



HAL
open science

Diversity, dynamics and evolution of bacterial epigenomes

Pedro H. Oliveira

► **To cite this version:**

Pedro H. Oliveira. Diversity, dynamics and evolution of bacterial epigenomes. Life Sciences [q-bio]. University of Paris Saclay, 2022. tel-04562706

HAL Id: tel-04562706

<https://hal.science/tel-04562706>

Submitted on 29 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITÉ PARIS-SACLAY
École Doctorale Structure et Dynamique des Systèmes Vivants

Mémoire de soutenance présenté en vue de l'obtention du
DIPLÔME D'HABILITATION A DIRIGER DES RECHERCHES

DIVERSITÉ, DYNAMIQUE ET ÉVOLUTION DES
ÉPIGÉNOMES BACTÉRIENS

DIVERSITY, DYNAMICS AND EVOLUTION OF
BACTERIAL EPIGENOMES

Pedro H. OLIVEIRA

Ingénieur de Recherche au Commissariat à l'Énergie Atomique et aux
Énergies Alternatives - Genoscope

Mars 2, 2022

Membres du Jury :

Examineur : Olivier LESPINET - Université Paris-Saclay
Examineur : Sylvain BRISSE - Institut Pasteur
Rapporteuse : Justine COLLIER - Université de Lausanne
Rapporteuse : Olga SOUTOURINA - Université Paris-Saclay
Rapporteur : Julien BRILLARD - Université de Montpellier

Acknowledgments

There are no proper words to convey my deep gratitude and respect to Genoscope's Director Patrick Wincker, for his enduring support and guidance of my activities at SeqLab. I further extend a sincere acknowledgement to all members of SeqLab and students who made this work possible, as well as to Eduardo Rocha (Institut Pasteur) and Gang Fang (Mount Sinai School of Medicine) for nurturing my curiosity on microbial epigenomics and supporting my activities as independent researcher. Finally, I would like to thank my parents Manuel and Beatriz for giving me all the opportunities that have made me who I am today, my wife Karina for her unfailing love and understanding, and my daughter Lia for making me believe that a brighter future is possible.

Contents

List of Acronyms	ix
1 Curriculum vitae	1
2 Scientific career	10
2.1 Career summary	10
2.2 Managerial and administrative activities	12
2.2.1 CEA, Genoscope (Lab Head)	12
2.3 Mentoring activities	13
2.3.1 CEA, Genoscope (Lab Head)	13
2.3.2 Mount Sinai School of Medicine, Department of Genetics and Genomic Sciences (Senior Scientist)	13
2.3.3 University of Lisbon, Department of Bioengineering (1 st Postdoc)	13
2.3.4 University of Lisbon, Department of Bioengineering (PhD)	14
2.4 Scientific collaborations	14
2.4.1 CEA, Genoscope (Lab Head)	14
2.4.2 Mount Sinai School of Medicine, Department of Genetics and Genomic Sciences (Senior Scientist)	14
2.4.3 Institut Pasteur, Department of Genomes and Genetics (2 nd Postdoc)	14
2.4.4 University of Lisbon, Department of Bioengineering (1 st Postdoc)	14
2.4.5 University of Lisbon, Department of Bioengineering (PhD)	15
2.5 Editorial activities	15
2.6 Grant and peer-reviewing activities	15
2.7 Teaching activities	16
2.8 Grants, fellowships, and awards	16
3 Past research	17
3.1 Ph.D. research (2006-2009)	18
3.1.1 Introduction	18
3.1.2 Deletion-formation events in DNA biopharmaceuticals	20
3.1.3 A predictive tool for estimating recombination frequency in plasmids	22
3.1.4 IS-mediated genetic instability in plasmid biopharmaceuticals	23
3.1.5 Transition to the first postdoctoral project	26
3.2 Postdoctoral project (2010-2013)	26
3.2.1 Introduction	26

3.2.2	Effect of hypoxia and prolonged passaging on the genomic stability of <i>in vitro</i> expanded human stem/stromal cells . . .	28
3.2.3	Effect of hypoxia and prolonged passaging on mitochondrial performance	29
3.2.4	Role of non-canonical DNA structures and sequence motifs on human mitochondrial DNA instability	32
3.2.5	Transition to the second postdoctoral project	35
3.3	Postdoctoral project (2013-2016)	35
3.3.1	Introduction	35
3.3.2	The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts	37
3.3.3	Regulation of genetic flux between bacteria by restriction-modification systems	41
3.3.4	The chromosomal organization of horizontal gene transfer in bacteria	45
3.3.5	Transition to a Senior Scientist position	49
3.4	Senior Scientist (2016-2020)	49
3.4.1	Introduction	49
3.4.2	Methylome analysis reveals great epigenomic diversity in <i>C. difficile</i>	51
3.4.3	Comparative analysis of methylation sites across <i>C. difficile</i> genomes	52
3.4.4	Non-methylated sites are enriched in regulatory elements . . .	52
3.4.5	Loss of methylation impacts transcription of sporulation genes	56
3.5	Conclusions	56
4	Current and Future research	60
4.1	Context	60
4.2	Role of persistent MTases in Bacteria	61
4.2.1	Introduction	61
4.2.2	Research program	63
4.3	Interplay between diversification of methylation systems, their target specificity and genetic mobility	66
4.3.1	Introduction	66
4.3.2	Research program	68
4.4	The anti-phage defensome of complex microbial populations	69
4.4.1	Introduction	69
4.4.2	Research program	70
4.4.3	Preliminary results	72
4.5	Final remarks	76
	Bibliography	80

Contents **v**

Appendices **100**

A 5 most important publications **101**

List of Figures

3.1	Structural instability in plasmid biopharmaceuticals	19
3.2	Structural instability in commercial plasmid vectors	21
3.3	Repeat-mediated recombination frequency in plasmids	23
3.4	IS-mediated plasmid instability	25
3.5	Gene expression analysis and microsatellite instability	30
3.6	Evaluation of mitochondrial properties in stem cells	31
3.7	Human mitochondrial DNA instability	34
3.8	Quantification and distribution of R-M systems in prokaryotic genomes	38
3.9	Quantification and distribution of R-M systems in MGEs	39
3.10	Analysis of complete R-M systems and solitary components	40
3.11	Analysis of HR and HGT events in bacteria	42
3.12	Gene flux in bacteria encoding R-M systems	44
3.13	Key concepts used in the analysis of core gene families	45
3.14	Abundance and distribution of HTgenes in hotspots	47
3.15	Genetic diversity in hotspots and coldspots	48
3.16	Methylomes of <i>C. difficile</i> strains	51
3.17	Abundance, distribution, and conservation of the motif <u>CAAAAA</u> . .	53
3.18	Overlap of transcription factor binding sites and transcription start sites with <u>CAAAAA</u>	55
3.19	Gene expression analysis	57
3.20	Overview of my past work	59
4.1	Summary of MTase conservation in bacterial genomes from GenBank	62
4.2	Persistent MTases: preliminary observations	64
4.3	Interplay between MGEs host range and MTase target specificity . .	67
4.4	Defensome analysis	71
4.5	The defensome of TARA MAGs	73
4.6	Defense islands in MAGs	75
4.7	Study of bacterial methylomes from clonal isolates, microbiomes, and holobionts	77

List of Acronyms

ABI Abortive Infection.....	69
ASCs Adipose-derived Stem Cells.....	27
BMSCs Bone-marrow MSCs.....	27
EMA European Medicines Agency.....	27
FDA US Food and Drug Administration.....	27
GOI Gene of Interest.....	18
ICEs/IMEs Integrative Conjugative Elements / Integrative Mobile Elements .	37
IS Insertion Sequence.....	11
HGT Horizontal Gene Transfer.....	10
HMM Hidden Markov Model.....	37
HR Homologous Recombination.....	23
MAGs Metagenome-Assembled Genomes.....	13
MTases MethylTransferases.....	12
MSCs Mesenchymal Stem Cells.....	13
MSI MicroSatellite Instability.....	27
MMR MisMatch Repair.....	27
MGEs Mobile Genetic Elements.....	10
mtDNA mitochondrial DNA.....	11
NHEJ Non-Homologous End-Joining.....	27
pDNA plasmid DNA.....	18
polyA polyAdenylation.....	18
PT PhosphoroThioation.....	69
REase Restriction Endonuclease.....	36
R-M Restriction-Modification.....	11
SMRT-seq Single-Molecule Real-Time sequencing.....	49

T-A Toxin-Antitoxin	63
TFs Transcription Factors	54
TFBSs Transcription Factor Binding Sites	54
TSSs Transcription Start Sites	54

CHAPTER 1

Curriculum vitae

Pedro H. OLIVEIRA

Group Leader



[in](#) pedroholiveira [twitter](#) pholive81 [github](#) oliveira-lab
[pholiveira.net](#) [Google Scholar](#)
[+33 769 668 874](#) [@pcphco@gmail.com](#)
Paris, France
Born on March 2nd 1981 in Lisbon, Portugal

Engineering and biotechnology background. Over the years, I prioritized efforts in acquiring a unique expertise in the broad fields of genomics and genome dynamics. My current interests are on identifying emerging challenges in the fast-growing research field of epigenomics, and on developing novel methods that can fundamentally address these challenges. My long term goal is to obtain biological insights that can be translated into more accurate disease diagnosis and more effective treatment for certain bacterial pathogens.

PROFESSIONAL EXPERIENCE

- | | |
|--------------------------|--|
| Currently
2020 | Lab Head, ATOMIC ENERGY AND ALTERNATIVE ENERGIES COMMISSION - GENOSCOPE, Paris, France
► Sequencing technologies and multi-scale biology
Genomics Epigenomics Bioinformatics |
| 2020
2016 | Senior Scientist, MOUNT SINAI SCHOOL OF MEDICINE, New York, USA
► Microbial epigenomics
Comparative Epigenomics / transcriptomics SMRT-seq R UNIX shell Microbial Genetics |
| 2016
2013 | Postdoctoral Fellow, INSTITUT PASTEUR, Paris, France
► Bacterial defense systems and genome organization
Unix Shell R Comparative Genomics Phylogenomics |
| 2012
2010 | Postdoctoral Fellow, UNIVERSITY OF LISBON INSTITUTO SUPERIOR TÉCNICO, Lisbon, Portugal
► Genetic instability in human stem cells for clinical applications
Flow cytometry Gene expression analysis Image analysis Multilineage differentiation Genome sequencing
Immunophenotyping Haplotyping Bioreactors |
| 2009
2008 | Visiting Researcher, MASSACHUSETTS INSTITUTE OF TECHNOLOGY HARVARD MEDICAL SCHOOL, Boston, USA
► Bacterial genome editing
Multiplex genome engineering Gene expression analysis |
| 2005
2005 | Research Assistant, UNIVERSITY COLLEGE LONDON, London, UK
Project financed by an ERASMUS fellowship
► Metabolite screening in bacteria
HPLC Mass spectrometry Antimicrobial susceptibility assays Bioreactors |

EDUCATION

- | | |
|-----------|---|
| 2022 | HDR - Habilitation for Research Direction Habilitation à Diriger des Recherches, University of Paris-Saclay, France |
| 2006-2010 | PhD in Biotechnology, University of Lisbon Instituto Superior Técnico |
| 1999-2005 | MSc in Biological Engineering, University of Lisbon Instituto Superior Técnico |

GRANTS | FELLOWSHIPS | AWARDS

- | | |
|------|--|
| 2017 | Grant PTDC/BTM-SAL/28624/2017 attributed by the Portuguese Ministry of Science (FCT) - Lactic Acid Bacteria as Cell Factories : A Synthetic Biology Approach for Plasmid DNA And Recombinant Protein Production. Role as consultant. |
| 2010 | Post-Doctoral grant BPD/64652/2009 attributed by the Portuguese Ministry of Science (FCT) |
| 2010 | Award for top-cited article during 2008-2010 |
| 2010 | Ph.D. grant BD/22320/2005 attributed by the Portuguese Ministry of Science (FCT) |
| 2007 | Travel grant - 32 nd FEBS Congress |
| 2005 | ERASMUS Fellowship |

Editorial Board Memberships

- 2021-Present Member of the Editorial Board of **Microbiology** [Link](#)
- 2020-Present Member of the Advisory Board of **Heliyon** (Cell group) [Link](#)
- 2020-2021 Member of the Editorial Board of **Genomics** [Link](#)
- 2019-Present Member of the Editorial Board of **mSystems** [Link](#)
- 2019-Present Member of the Editorial Board of **Scientific Reports** (Nature group) [Link](#)
- 2019-Present Member of the Editorial Board of **Scientific Data** (Nature group) [Link](#)
- 2019-Present Member of the Editorial Board of **BMC Microbiology** [Link](#)
- 2019-Present Member of the Editorial Board of **BMC Biotechnology** [Link](#)
- 2017-Present Member of the Editorial Board of **PLoS ONE** [Link](#)
- 2017-Present Member of the Editorial Board of **Frontiers in Genetics** [Link](#)
- 2017-Present Member of the Editorial Board of **Frontiers in Microbiology** [Link](#)

Society Memberships

- 2021-Present Member of the Executive Committee of **France Génomique** [Link](#)

Reviewing Activities

- 2021-2022 Grant reviewer for the **US National Science Foundation (NSF) Graduate Research Fellowship Program (GRFP) - Microbial Biology Panel** [Link](#)
- 2021 Grant reviewer for the **Swiss National Science Foundation (SNSF)** [Link](#)
- 2020 Grant reviewer for **Graduate Women in Science (GWIS)** [Link](#)
- 2012-Current External Reviewer for multiple journals (e.g. : *PLoS Computational Biology, Nucleic Acids Research, Scientific Reports, Bioinformatics, BMC Genomics, Cell Death and Differentiation, Stem Cells*). [Link](#)

Guest Editor Activities

- 2022 Editor of the book *Computational Epigenomics and Epitranscriptomics - Methods in Molecular Biology, Springer Nature*. [Link](#)
- 2015 Lead Guest Editor of a *Special issue on Biotherapeutics*. [Link](#)

PUBLICATIONS

- ▶ **Oliveira PH[†]**. Bacterial epigenomics : coming of age. *mSystems* (2021). 6(4):e0074721 [Link](#)
- ▶ **Oliveira PH[†]**, Fang G. Conserved DNA methyltransferases : a window into fundamental aspects of epigenetic regulation in Bacteria. *Trends in Microbiology* (2021). 29(1), 28-40 [Link](#)
- ▶ **Oliveira PH**, Ribis JW, Garrett EM, Trzilova D, Kim A, Sekulovic O, Mead EA, Pak T, Zhu S, Deikus G, Touchon M, Lewis-Sendari M, Beckford C, Zeitouni NE, Altman DR, Webster E, Oussenko I, Bunyavanich S, Aggarwal AK, Bashir A, Patel G, Wallach F, Hamula C, Huprikar S, Schadt EE, Sebra R, van Bakel H, Kasarskis A, Tamayo R, Shen A, Fang G. Epigenomic characterization of *Clostridioides difficile* finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. *Nature Microbiology* (2020). 5(1), 166-180 [Link](#)
- ▶ Cury J, **Oliveira PH**, de la Cruz F, Rocha EPC. Host range and genetic plasticity explain the coexistence of integrative and extrachromosomal mobile genetic elements. *Molecular Biology and Evolution*. (2018). 35(9), 2230-2239. [Link](#)
- ▶ **Oliveira PH^{††}**, Touchon M[†], Cury J, Rocha EPC. The chromosomal organization of gene transfer in bacteria. *Nature Communications*. (2017). 8(841), 1-11. [Link](#)
- ▶ **Oliveira PH[†]**, Touchon M, Rocha EPC. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. USA*. (2016). 113(20), 5658-5663. [Link](#)
- ▶ **Oliveira PH[†]**, Touchon M, Rocha EPC. The interplay of restriction modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Research*. (2014). 42(16), 10618-10631. [Link](#)
- ▶ **Oliveira PH[†]**, Lobato da Silva C, Cabral JMS. Genomic instability in human stem cells : current status and future challenges. *Stem Cells*. (2014). 32(11), 2824-2832. [Link](#)

- Gonçalves GAL[‡], **Oliveira PH[‡]**, Gomes AG, Lewis LA, Prather KLJ, Prazeres DMF, Monteiro GA. Evidence that the insertion events of IS2 transposition are biased towards abrupt compositional shifts in target DNA and modulated by a diverse set of culture conditions. *Applied Microbiology and Biotechnology*. (2014). 98(15), 6609-6619. [Link](#)
- **Oliveira PH[‡]**, Prazeres DMF, Monteiro GA. DNA instability in bacterial genomes : causes and consequences. In *Genome Analysis : Current Procedures and Applications* (Poptsova M. Eds.). Caister Academic Press, UK, 2014. [Link](#)
- **Oliveira PH**, Mairhofer J. Marker-free plasmids for biotechnological applications – implications and perspectives. *Trends in Biotechnology*. (2013). 31, 539-547. [Link](#)
- **Oliveira PH[‡]**, Lobato da Silva C, Cabral JMS. An appraisal of human mitochondrial DNA instability : new insights into the role of non-canonical DNA structures and sequence motifs. (2013). *PLoS ONE* 8(3) : e59907. [Link](#)
- **Oliveira PH**, Boura JS, Abecasis MM, Gimble JM, Lobato da Silva C, Cabral JMS. Impact of hypoxia and long-term cultivation on the genomic stability and mitochondrial performance of *ex-vivo* expanded human stem/stromal cells. *Stem Cell Research*. (2012). 9(3), 225-236. [Link](#)
- **Oliveira PH[‡]**, Prazeres DMF, Monteiro GA. Structural and segregational instability in plasmid biology. In *Plasmids : Genetics, Applications and Health* (Gonzalez FER; Lopez MI Eds.). Nova Science Publishers, NY, USA, 2012. ISBN : 978-1-62081-370-6. pp. 79-100. [Link](#)
- Borlido L, Azevedo AM, Sousa AG, **Oliveira PH**, Roque AC, Aires-Barros MR. Fishing human monoclonal antibodies from a CHO cell supernatant with boronic acid magnetic particles. *Journal of Chromatography B*. (2012). 903, 163-170. [Link](#)
- Lewis LA., Astatke M, Umekubo PT, Alvi S, Saby R, Afrose J, **Oliveira PH**, Monteiro GA, Prazeres DMF. Protein-DNA interactions define the mechanistic aspects of circle formation and insertion reactions in IS2 transposition. *Mobile DNA*. (2012). 3(1):1. [Link](#)
- **Oliveira PH[‡]**, Prather KLJ, Prazeres DMF, Monteiro GA. Mutation detection in plasmid-based biopharmaceuticals. *Biotechnology Journal*. (2011). 6(4), 378-391. [Link](#)
- **Oliveira PH[‡]**, Prather KLJ, Prazeres DMF, Monteiro GA. Analysis of DNA repeats in bacterial plasmids reveals the potential for recurrent instability events. *Applied Microbiology and Biotechnology*. (2010). 87(6), 2157-2167. [Link](#)
- **Oliveira PH**, Prather KLJ, Prazeres, DMF, Monteiro GA. Structural instability of plasmid biopharmaceuticals : challenges and implications. *Trends in Biotechnology*. (2009). 27, 503-511. [Link](#)
- **Oliveira PH**, Prazeres DMF, Monteiro GA. Deletion formation mutations in plasmid expression vectors are unfavored by run-away amplification conditions and differentially selected under kanamycin stress. *Journal of Biotechnology*. (2009). 143(4), 231-238. [Link](#)
- **Oliveira PH**, Lemos F, Monteiro GA, Prazeres DMF. Recombination frequency in plasmid DNA containing direct repeats – predictive correlation with repeat and intervening sequence length. (2008). *Plasmid*. 60(2), 159-165. [Link](#)
- Ribeiro SC[‡], **Oliveira PH[‡]**, Prazeres DMF, Monteiro GA. High frequency plasmid recombination mediated by 28 bp direct repeats. *Molecular Biotechnology*. (2008). 40(3), 252-260. [Link](#)
- **Oliveira PH**, Batagov A., Ward J, Baganz F, Krabben P. Identification of Erythrobactin, a hydroxamate-type siderophore produced by *Saccharopolyspora erythraea*. *Letters in Applied Microbiology*. (2006). 42(4), 375-380. [Link](#)
- Aleixo AI, **Oliveira PH**, Diogo HP, Minas da Piedade ME. Enthalpies of formation and lattice enthalpies of alkaline metal acetates. *Thermochimica Acta*. (2005). 428(1-2), 131-136. [Link](#)

[‡] Equal contribution

[†] Corresponding author

Publications as **first (or co-first) author** : 19 out of 23

Publications as **corresponding author** : 10 out of 23

EDITORIALS ABOUT MY PUBLICATIONS

- Jungbauer A. Improved products and processes through biochemical engineering science. *Biotechnology Journal*. (2011). 6(4), 362-363. [Link](#)

Oral communications (IS – invited speaker; SS – selected speaker)

- ▶ Guichard H, Delmont TO, Beraud M, Labadie K, Cruaud C, Poulain J, Wincker P, and **Oliveira PH**. The anti-phage defenses of TARA Oceans microbial metagenomes. *International Symposium on Environmental and Agronomical Genomics*, October 27-29, 2021. (IS)
- ▶ Guichard H, Delmont TO, Beraud M, Labadie K, Cruaud C, Poulain J, Wincker P, and **Oliveira PH**. Novel Mechanisms of Defence Against Foreign DNA. *World Microbe Forum*, June 20-24, 2021. (SS)
- ▶ Guichard H, Delmont TO, Beraud M, Labadie K, Cruaud C, Poulain J, Wincker P, and **Oliveira PH**. The landscape of cellular immune systems in marine microbial communities. *Microbiology Society Annual Conference*, April 26-30, 2021. (SS)
- ▶ Vacherie B, Mairey B, Payen E, Guérin T, Engelen S, Cruaud C, Labadie K, **Oliveira PH**, Lemainque A, Wincker P. Detailed comparison of DNA extraction methods for Nanopore sequencing of complex plant and animal genomes. *Nanopore Community Meeting Online*, December 1-3, 2020. (SS)
- ▶ **Oliveira PH**. Conserved DNA methyltransferases as keystones of epigenetic regulation in Bacteria. *26th Boston Bacterial Meeting Online*. July 16-17, 2020. (SS) [Link](#)
- ▶ **Oliveira PH**. Epigenomics in the era of third-generation sequencing: a large-scale study of the human pathogen *Clostridioides difficile*. *SMRT Leiden Online*. May, 2020. (SS) [Link](#)
- ▶ **Oliveira PH**. Large-scale detection and integrative functional characterization of DNA modifications in Bacteria. *Ohio State University - Department of Microbiology*, Ohio, USA. December 17, 2019. (IS)
- ▶ **Oliveira PH**. Large-scale detection and integrative functional characterization of DNA modifications in Bacteria. *Université Claude Bernard Lyon 1 - Molecular Microbiology and Structural Biochemistry*, Lyon, France. November 12, 2019. (IS)
- ▶ **Oliveira PH**. Large-scale integrative characterization of bacterial methylomes. *Center for Integrative Biology*, Toulouse, France. September 17, 2019. (IS)
- ▶ **Oliveira PH**. Epigenomic landscape of the human pathogen *Clostridium difficile*. Institut Pasteur - Microbiology Department Seminar, Paris. March 21, 2019. (IS)
- ▶ **Oliveira PH**, Kim A, Sekulovic O, Pak T, Zhu S, Mead EA, Deikus G, Touchon M, Lewis M, Beckford C, Zeitouni NE, Altman D, Webster E, Oussenko I, Aggarwal AK, Bashir A, Patel G, Hamula C, Huprikar S, Roberts RJ, Tamayo R, Schadt EE, Sebra R, van Bakel H, Kasarskis A, Shen A, Fang G. Epigenomic landscape of the human pathogen *Clostridium difficile*. *6th International Clostridium difficile Symposium*. Bled, Slovenia, September 12-14, 2018. (SS)
- ▶ **Oliveira PH**. The chromosomal organization of horizontal gene transfer. *Genome Science*, Nottingham, UK, September 4-6, 2018. (IS)
- ▶ **Oliveira PH**, Touchon M, Rocha EPC. Genetic mobility and the distribution of restriction modification systems in prokaryotes. *Department Days of Institut Pasteur*, Saint-Aignan-sur-Cher, France, September 22-24, 2014. (SS)
- ▶ **Oliveira PH**, Rocha EPC. A comparative genomics approach provides new insights into the distribution and evolutionary history of restriction modification systems in bacteria. *Annual Meeting of the Society for Molecular Biology and Evolution*, Chicago, USA, July 7-11, 2013. (SS)
- ▶ **Oliveira PH**. Mechanistic definition of IS2 transposition and the basis for its non-random insertional specificity in relevant DNA-based biopharmaceuticals. *3rd Annual World Congress of Molecular and Cell Biology*. Suzhou, China, June 14-16, 2013. (IS)
- ▶ **Oliveira PH**, da Silva CL, Cabral JMS. An appraisal of human mtDNA instability: role of DNA topology and sequence motifs. *15th European Congress on Biotechnology-Biocrossroads*. Istanbul, Turkey, September 23-26, 2012. (SS)
- ▶ **Oliveira PH**, Lobato da Silva C, Cabral JMS. An appraisal of human mtDNA instability: role of DNA topology and composition. *3rd Scientific Meeting of the Institute for Biotechnology and Bioengineering*. Lisbon, Portugal, March 16-17, 2012. (SS)
- ▶ **Oliveira PH**, Boura JS, Abecasis MM, Gimble J, da Silva CL, Cabral JMS. Genetic stability during the *ex-vivo* expansion of human mesenchymal stem cells for clinical applications. *6th International Meeting of the Portuguese Society for Stem Cells and*

Cellular Therapy (SPCE-TC). Cantanhede, Portugal, April 28-29, 2011. (SS)

- **Oliveira PH**. Biofármacos : Desafios e Limitações. *Tertúlias FNACiência*, FNAC Guimarães, Portugal, June 2, 2011. (IS)
- **Oliveira PH**. Da sala de aula ao laboratório – Uma experiência na primeira pessoa. *Jornadas de Engenharia Química e Biológica (JEQB)*. Instituto Superior Técnico, Lisbon, Portugal, March 21-25, 2011. (IS)
- **Oliveira PH**, Boura JS, Abecasis MM, da Silva CL, Cabral JMS. An appraisal of genetic stability in human mesenchymal stem cells. *1st Portuguese Meeting in Bioengineering*. Lisbon, Portugal, March 1-4, 2011. (SS)
- **Oliveira PH**, Lobato da Silva C, Cabral JMS. Bias in human mitochondrial DNA repeat distribution correlates with common deletion/amplification breakpoints involved in genetic disorders and cancer. *5th International Meeting of the Portuguese Society for Stem Cells and Cellular Therapy (SPCE-TC)*. Guimarães, Portugal, November 20-21, 2010. (SS)
- **Oliveira PH**, Lopes G, Prather KJ, Prazeres DMF, Monteiro GA. Structural instability in plasmid biopharmaceuticals for DNA vaccination. *2nd Scientific Meeting of the Institute for Biotechnology and Bioengineering*. Braga, Portugal, October 23-24, 2010. (SS)
- **Oliveira PH**, Prather KJ, Prazeres DMF, Monteiro GA. Structural instability in plasmid vectors for DNA vaccination. *III International Conference on Environmental, Industrial and Microbial Biotechnology (BioMicroWorld)*. Lisbon, Portugal, December 2-4, 2009. (SS)
- Santos R, Marques MPC, **Oliveira PH**, Carvalho F, Carvalho C, Monteiro GA, Cabral JMS, Frade R, Silva M, Fernandes P. The sunny side of mycobacteria. *30th Annual Congress of the European Society of Mycobacteriology*, Porto, Portugal, 5-8 July 2009. (SS)
- **Oliveira PH**, Prather KJ, Prazeres DMF, Monteiro GA. Structural instability in plasmid vectors for DNA vaccination. *MicroBiotec 2009*. Vilamoura, Portugal, November 28-30, 2009. (SS)
- Ribeiro SC, **Oliveira PH**, Prazeres DMF, Monteiro GA. Kanamycin-induced mutation in *E. coli* cells harbouring plasmids with direct repeats – The adaptive response. *31st FEBS Congress – Molecules in health and disease*. Istanbul, Turkey, June 24-29, 2006. (SS)
- Krabben P, **Oliveira PH**, Baganz F, Ward J. Exploring the substrate spectrum of the antibiotic producing bacteria *Saccharopolyspora erythraea*. *12th European Congress on Biotechnology*. Copenhagen, Denmark, August 21-24, 2005. (SS)

Poster communications

- Audouy T, Beluche O, Bertrand L, Bordelais I, Brun E, Dubois M, Dumont C, Estrada B, Ettetdgui E, Guerin T, El Hajji Z, Hamon C, Lebled S, Lenoble P, Louesse C, Magdelenat G, Mahieu E, Mangenot S, Martins N, Milani C, Muanga J, Orvain C, Paillard M, Payen E, Perroud P, Petit E, Robert D, Ronsin M, Sanchez S, Vacherie B, Barbance JM, Beraud M, Cruaud C, Labadie K, Perdereau A, Poulain J, Wincker P, and **Oliveira PH**. Genoscope's SeqLab at 25 : extending the genomic revolution to decode life. *International Symposium on Environmental and Agronomical Genomics*, October 27-29, 2021.
- Vacherie B, Mairey B, Payen E, Guérin T, Engelen S, Cruaud C, Labadie K, **Oliveira PH**, Lemainque A, Wincker P. Detailed comparison of DNA extraction methods for Nanopore sequencing of complex plant and animal genomes. *Nanopore Community Meeting Online*, December 1-3, 2020.
- Ribis J, **Oliveira PH**, Mead E, Sekulovic O, Tamayo R, Fang G, Shen A. Epigenetic regulation of *Clostridioides difficile* spore formation by a conserved orphan DNA methyltransferase. *Clostridial 11*, Leiden, Netherlands, August 19-22, 2019.
- **Oliveira PH**, Touchon M, Rocha EPC. Genetic mobility and the distribution of restriction modification systems in prokaryotes. *13th European Conference on Computational Biology (ECCB 2014)*, Strasbourg, France, September 7-10, 2014. (Selected for F1000 Posters : [Link](#))
- **Oliveira PH**, Touchon M, Rocha EPC. Genetic mobility and the distribution of restriction modification systems in prokaryotes. *EMBO Conference on Microbiology after the Genomics Revolution : Genomes 2014*, Paris, France, June 24-27, 2014.
- Gonçalves GAL, **Oliveira PH**, Lewis LA, Prazeres DMF, Monteiro GA. Identification of IS2 transposition in pVAX1-based plasmid, a common vector for DNA vaccine development. *15th European Congress on Biotechnology-Biocrossroads*. Istanbul, Turkey, September 23-26, 2012.

- **Oliveira PH**, Gonçalves GAL, Lewis LA, Prazeres DMF, Monteiro GA. Preferential transposition of the mobile element IS2 into AT- and GC-skew polarity switches. *15th European Congress on Biotechnology-Biocrossroads*. Istanbul, Turkey, September 23-26, 2012.
- Gomes AG and **Oliveira PH**. Comparative genomic analysis indicates strand compositional asymmetries as regional landmarks in plasmid evolution. *15th European Congress on Biotechnology-Biocrossroads*. Istanbul, Turkey, September 23-26, 2012.
- **Oliveira PH**, da Silva CL, Cabral JMS. Unusual DNA structures and instability motifs correlate with human mitochondrial deletion breakpoints involved in genetic disorders and cancer. *11th International Symposium on Mutations in the Genome*. Santorini, Greece, June 6-10, 2011.
- **Oliveira PH**, Boura JS, Abecasis MM, Gimble J, da Silva CL, Cabral JMS. An appraisal of genetic stability during the *ex-vivo* expansion of human mesenchymal stem cells. *11th International Symposium on Mutations in the Genome*. Santorini, Greece, June 6-10, 2011.
- **Oliveira PH**, Prather KJ, Prazeres DMF, Monteiro GA. Genome-wide analysis of DNA repeats in bacterial plasmids identifies genetically unstable hotspots and candidate regions for minimization. *8th European Symposium on Biochemical Engineering Science (ESBES)*. Bologna, Italy, September 5-8, 2010.
- Gonçalves GAL, **Oliveira PH**, Prather KJ, Prazeres DMF, Monteiro GA. Rational engineering of *E. coli* strains and vectors for improved manufacturing of plasmid biopharmaceuticals. *2nd MIT-Portugal Annual Conference : Creating Value through Systems Thinking*. Porto, Portugal, 28th September, 2010.
- Gonçalves GAL, **Oliveira PH**, Bower DM, Prazeres DMF, Monteiro GA, Prather KLJ. Rational engineering of *E. coli* strains for plasmid biopharmaceutical manufacturing. *Science 2010 Meeting*, Lisbon, Portugal, 4-7th July, 2010.
- **Oliveira PH**, Prazeres DMF, Monteiro GA. Hotspots for recombination are widespread among plasmid vectors. *3rd FEMScongress*. Gothenburg, Sweden, June/July 28-02, 2009.
- **Oliveira PH**, Prather KJ, Prazeres DMF, Monteiro GA. Rational engineering of *E. coli* strains and vectors for improved manufacturing of plasmid biopharmaceuticals. *1st MIT-Portugal Annual Conference : Engineering for Better Jobs*. Lisbon, Portugal, 7th July, 2009.
- Monteiro GA, Prazeres DMF, Azzoni A, Ribeiro S, Carvalho J, **Oliveira PH**, Freitas S, Magalhães S, Lima J, Beira J. Design and stability of plasmid vectors. *1st Scientific Meeting of the Institute for Biotechnology and Bioengineering (IBB)*. Faro, Portugal, 16th May, 2009.
- Prazeres DMF, Monteiro GA, Fonseca LP, Martins S, **Oliveira PH**, Azzoni A, Lima, J. Quality control and monitoring of plasmid biopharmaceuticals. *1st Scientific Meeting of the Institute for Biotechnology and Bioengineering (IBB)*. Faro, Portugal, 16th May, 2009.
- **Oliveira PH**, Prazeres DMF, Monteiro GA. Plasmid DNA instability mediated by direct repeats and type-2 insertion sequences. *EMBO conference Recombination Mechanisms*. Il Ciocco, Italy, May 19-23, 2008.
- **Oliveira PH**, Nunes SC, Ribeiro SC, Prazeres DMF, Monteiro GA. Kanamycin-induced recombination in *E. coli* cells harbouring plasmids with direct repeats. *32nd FEBS Congress – Molecular Machines*. Vienna, Austria, July 7-12, 2007.
- **Oliveira PH**, Prazeres DMF, Monteiro GA. Plasmid recombination by direct repeats and type 2 insertion sequences. *MicroBio-tec'07*. Lisbon, Portugal, 30 November – 2nd December, 2007.
- **Oliveira PH**, Marques MPC, Santos R, Claudino M, Monteiro GA, Cabral JMS, Fernandes P. Searching for a multipurpose *Mycobacterium*. *MicroBio-tec'07*. Lisbon, Portugal, 30 November – 2nd December, 2007.
- **Oliveira PH**, Linde E, Baganz F, Ward JM, Krabben P. From genome to improved antibiotic production by *Saccharopolyspora erythraea*. *ExGen - Exploiting Genomics Grant Holders*. Cambridge, UK, 19-20 July, 2006.
- **Oliveira PH**, Ribeiro SC, Prazeres DMF, Monteiro GA. Deletions in *Escherichia coli* plasmids harbouring direct repeats – Environmental factors and the adaptive response. *XVth National Congress of Biochemistry*. Aveiro, Portugal, December 8-10, 2006.
- Aleixo AI, **Oliveira PH**, Diogo HP, Minas da Piedade ME. Enthalpies of formation of alkaline metal acetates. *6th Mediterranean Conference on Calorimetry and Thermal Analysis*. Porto, Portugal, July 27-30, 2003.

MEDIA AND PUBLIC OUTREACH

- ▶ Participation in the Podcast series *Convidado Extra* (in Portuguese) from Jornal Observador (September 2020) [Link](#)
- ▶ Participation in the video series *Teach Me In 10* from Technology Networks (July 2020) [Link](#)
- ▶ Oliveira *et al.* (2020). Nature Microbiology highlighted by [GenomeWeb](#) [EurekAlert](#) [Phys](#) [TechnologyNetworks](#) [News Medical](#) [Mount Sinai](#) [Público](#) [Expresso](#) [DN](#) [JN](#) [IST](#) [RTP1](#) [MedPage Today](#) [JN](#) [Observador](#) [PacBio](#) [Visão](#)
- ▶ Profile highlighted in the journal Scientific Reports [Link](#)
- ▶ Oliveira *et al.* (2017). Nature Communications highlighted by [GEN](#) [Science et Vie](#) [GenomeWeb](#) [IST](#)
- ▶ Oliveira *et al.* (2016). PNAS highlighted by the Portuguese media [DN](#) [SIC](#)
- ▶ Interview given to Radio Antena 1 (in Portuguese) [Link](#)
- ▶ Interview given to Fundação Francisco Manuel dos Santos (in Portuguese) [Link](#)
- ▶ Research profile highlighted in the book *Conversas com Ciência*. ISBN : 978-989-89559-2-0 (in Portuguese) [Link](#)
- ▶ Interview given to project Ciência Com Vida (in Portuguese) [Link](#)
- ▶ Interview given to Radio channel Antena 1 (in Portuguese) [Link](#)
- ▶ Research profile highlighted in the Portuguese newspaper Público [Link](#)
- ▶ Book chapter on *Biopharmaceuticals* for a general audience (in Portuguese).
Marcos JC and **Oliveira PH.** (2012). Biofármacos : Desafios e Limitações. In *Conversas com Ciência*. Escola de Ciências da Universidade do Minho, Braga, Portugal, 2012. ISBN : 978-989-98077-0-9, pp. 63-65. [Link](#)


STUDENT SUPERVISION / TEACHING

- 2022 Auriane Lacroix (Master student (M2) at Genoscope, Évry, France) : *The anti-phage defensible in MAGs reconstructed from human complex microbial communities* - Role as PI (100%)
- . 2022 Angelina Beavogui (Master student (M2) at Genoscope, Évry, France) : *The anti-phage defensible in MAGs reconstructed from environmental complex microbial communities* - Role as PI (100%)
- . 2021-2022 Invited Academic Seminars - Master in Bioinformatics. *Advances in Sequencing Technologies*. University of Evry, France.
- 2021 Hadrien Guichard (Master student (M2) at Genoscope, Évry, France) : *The anti-phage defensible of complex microbial communities* - Role as PI (100%)
- . 2021 Angelina Beavogui (Master student (M1) at Genoscope, Évry, France) : *Solitary DNA methyltransferases and control of gene flux in bacteria*. - Role as PI (100%)
- . 2016-2017 Alex Kim (Research Fellow at Mount Sinai School of Medicine, NY, USA) : *Epigenomic Landscape in the Human Pathogen Clostridioides difficile* - Role as co-PI (50%)
- . 2014 Invited Academic Seminars. *Therapeutic Potential of Human Stem Cells*. Université Paris 3 – Sorbonne Nouvelle, Paris, France.
- 2012 Invited Academic Seminars. *Genomic Instability*. MIT-Portugal course, Instituto Superior Técnico, Lisbon, Portugal.
- 2011-2012 Invited Academic Seminars. *Biological Reactors*. Instituto Politécnico de Setúbal (IPS). Portugal.
- 2010-2012 Márcia Mata (Research Fellow at Instituto Superior Técnico, Lisbon, Portugal) : *Scalable Reactor Systems for the Production of Human Mesenchymal Stem Cells* - Role as co-PI (80%)
- . 2008 Principal investigator of the MIT-Portugal program i-Teams with the project entitled : *Fermentative Siderophore Production*.
- 2007 Fanny Trouvat (Research Fellow at Instituto Superior Técnico, Lisbon, Portugal) : *Plasmid Instability Mediated by Direct-Repeats* - Role as co-PI (90%)
- .

LANGUAGES

Portuguese	●	●	●	●	●
English	●	●	●	●	●
French	●	●	●	○	○
Mandarin	●	○	○	○	○

OTHER INTERESTS

› Photography  [Link](#)

Scientific career

Contents

2.1	Career summary	10
2.2	Managerial and administrative activities	12
2.2.1	CEA, Genoscope (Lab Head)	12
2.3	Mentoring activities	13
2.3.1	CEA, Genoscope (Lab Head)	13
2.3.2	Mount Sinai School of Medicine, Department of Genetics and Genomic Sciences (Senior Scientist)	13
2.3.3	University of Lisbon, Department of Bioengineering (1 st Postdoc)	13
2.3.4	University of Lisbon, Department of Bioengineering (PhD)	14
2.4	Scientific collaborations	14
2.4.1	CEA, Genoscope (Lab Head)	14
2.4.2	Mount Sinai School of Medicine, Department of Genetics and Genomic Sciences (Senior Scientist)	14
2.4.3	Institut Pasteur, Department of Genomes and Genetics (2 nd Postdoc)	14
2.4.4	University of Lisbon, Department of Bioengineering (1 st Postdoc)	14
2.4.5	University of Lisbon, Department of Bioengineering (PhD)	15
2.5	Editorial activities	15
2.6	Grant and peer-reviewing activities	15
2.7	Teaching activities	16
2.8	Grants, fellowships, and awards	16

2.1 Career summary

Bacterial genomes are remarkably stable from one generation to the next but are plastic on an evolutionary time scale, being substantially shaped by Horizontal Gene Transfer (HGT), genome rearrangements/mutations, and the activity of Mobile Genetic Elements (MGEs). This implies the existence of a delicate balance between the maintenance of genetic stability and generation of variability. My past research avenues versed on different but complementary aspects of this balance: from the

detection of novel instability events to the in-depth analysis of gene flux-controlling systems; from the organization of HGT in bacterial chromosomes to the organization and diversification of epigenomes. I also have moved between 'wet-' and dry-lab' environments, in order to fully harness the interdisciplinary nature of these topics.

During my Ph.D. (2006-2009), I investigated the occurrence of novel structural instability events in bacterial plasmids, with an emphasis on those used as DNA biopharmaceuticals in clinical applications (gene therapy and DNA vaccination). I computationally predicted, and experimentally validated, the occurrence of multiple repeat-mediated deletions/amplifications and spontaneous Insertion Sequence (IS) transpositions in these vectors [Oliveira *et al.* 2009b, Gonçalves *et al.* 2014, Ribeiro *et al.* 2008, Oliveira *et al.* 2008, Oliveira *et al.* 2010, Oliveira *et al.* 2011, Oliveira *et al.* 2009a, Oliveira & Mairhofer 2013, Lewis *et al.* 2012]. These aberrations not only affect the clinical consistency of the end-products, but represent a safety concern upon administration to humans. I also developed a meta-analysis-driven mathematical model capable of predicting the frequency of these instability events [Oliveira *et al.* 2008]. During the second half of my Ph.D. (2008-2009), I worked as visiting student at the Massachusetts Institute of Technology and Harvard Medical School, where I used gene editing techniques to build safer bacterial chassis devoid of active ISs [Oliveira 2010]. Towards the end of my Ph.D. I actively collaborated with researchers from the City University of New York in order to unravel the insertion mechanisms and the genetic factors behind the insertion specificity of certain bacterial ISs [Lewis *et al.* 2012, Gonçalves *et al.* 2014].

For my first postdoctoral project (2010-2012), I considered gaining some expertise in eukaryotic cell biology, while still working on aspects linked to genome dynamics. In particular, I focused on studying the genetic (in)stability of human adult stem cells for clinical applications. I was able to provide a deeper understanding of the effects of *ex-vivo* expansion and hypoxia on the genomic stability of these cells [Oliveira *et al.* 2012, Oliveira *et al.* 2014a]. I also found for the first time a link between the presence of non-B DNA and rearrangements in the human mitochondrial DNA (mtDNA) with potential clinical relevance [Oliveira *et al.* 2013].

Given my expertise in wet-lab environments, I considered fundamental to strengthen my skills in bioinformatics/computational biology, while still keeping a focus on genome dynamics. This brought me to Eduardo Rocha's Lab at Institut Pasteur in Paris (2013-2016) for a second post-doctoral project. During this period, I made the most comprehensive and detailed analysis of Restriction-Modification (R-M) systems in prokaryotes, their relations with the vectors of HGT, and with other cell's defense systems [Oliveira *et al.* 2014b]. I also provided a more precise view on how R-M systems regulate gene flux and how their repertoire and/or specificity may shape population structure [Oliveira *et al.* 2016]. Towards the end of my stay, I performed a first of its kind genome-wide study for characterizing the landscape of HGT in bacteria [Oliveira *et al.* 2017].

My first independent position was as senior scientist at Mount Sinai School of Medicine in New York (2016-2020) where I led a project at the crossroads of bacterial genomics, epigenomics, metatranscriptomics, and clinical microbiology. In

particular I leveraged third generation long-read sequencing data to evaluate the role of highly pervasive DNA MethylTransferases (MTases) (*e.g.*: at the species level) as key players in the virulence and clinical success of human bacterial pathogens. I integrated multiple -omics data, performed the first comprehensive characterization of the epigenomic landscape of *Clostridioides difficile*, and discovered novel biological insights with important biomedical implications. Namely, I found a core MTase that regulates sporulation and additional relevant phenotypes, opening a new epigenomic dimension to study this critical pathogen, and a possibility to more effectively battle *C. difficile* infections [Oliveira *et al.* 2020].

From 2020 onward I assumed the position of head of the sequencing laboratory (SeqLab) at the French Alternative Energies and Atomic Energy Commission - National Sequencing Center - Genoscope. I am currently managing a team of 30 people, including technicians and senior scientists, and overseeing multiple small- to large-scale sequencing projects at both the national and international level. Simultaneously, I am setting up my own line of research on bacterial epigenomics, that builds both on SeqLab's rich portfolio in sequencing equipment, and its unique access to a large diversity of microbial samples. I am combining high-throughput (epi)genomic technologies and bioinformatic approaches to address outstanding questions put forward in the bacterial epigenomics field. In Chapter 4 of this document and in a recent Opinion paper [Oliveira 2021], I describe how these research questions are unfolding and call out the challenges ahead.

In a nutshell, I am a **Ph.D. + 11 years** senior researcher with experience in **both wet- and dry-lab** environments. I have a track record of **24 publications** and **52 communications**. The impact of my work has given me the opportunity to make several **invited review and commentary articles** as first/corresponding author. I have attained a certain recognition by the scientific community, resulting in being invited for the **editorial board** of **11 journals**, such as **mSystems**, **Microbiology**, and **Scientific Data**. Some of my work has been **highlighted by the media** (*e.g.*: **GEN**, **Science&Vie**, **GenomeWeb**). I was **awarded** an ERASMUS fellowship and Ph.D. / Post-Doctoral grants by the Portuguese Ministry of Science (FCT) to support my research. I have acted as **external reviewer** for **20 different journals**, **lead guest-edited** one special issue, and I am currently editing a Springer Nature **Methods in Molecular Biology** book entitled *Computational Epigenomics and Epitranscriptomics* to be published in 2022. **I led work packages** of several European H2020 and national-level projects, and **supervised / co-supervised** multiple undergraduates, master students, and research fellows. The following subsections, describe in more detail some aspects of my scientific career.

2.2 Managerial and administrative activities

2.2.1 CEA, Genoscope (Lab Head)

- September 2020 - Present. Lab Head at Genoscope. Management of a team of 30 people, including senior scientists and technicians.

- March 2021 - Present. Member of the executive committee of the France Génomique network.
- June 2021 - Present. Permanent invited member of the council of the UMR 8030.

2.3 Mentoring activities

2.3.1 CEA, Genoscope (Lab Head)

- January 2022 - June 2022. MSc (master II) advisor of Auriane Lacroix. Auriane's project deals with the abundance and diversity of anti-phage defense systems (defensome) in Metagenome-Assembled Genomes (MAGs) reconstructed from complex microbial communities obtained from the human gut.
- January 2022 - June 2022. MSc (master II) advisor of Angelina Beavogui. Angelina's project deals with the abundance and diversity of anti-phage defense systems (defensome) in MAGs reconstructed from complex microbial communities obtained from the TARA Oceans project.
- April 2021 - June 2021. MSc (master I) advisor of Angelina Beavogui. Angelina's project dealt with the control of gene flux in bacteria by orphan MTases. In particular, she studied the interplay between MTase target degeneracy and genetic mobility.
- February 2021 - July 2021. MSc (master II) advisor of Hadrien Guichard. Hadrien's project dealt with a large-scale analysis of the defensome in MAGs reconstructed from complex marine communities.

2.3.2 Mount Sinai School of Medicine, Department of Genetics and Genomic Sciences (Senior Scientist)

- September 2016 - January 2017. Co-PI of Alex Kim (research fellow). Alex's project involved the development of a bioinformatic pipeline for refinement of methylation motifs predicted by Pacbio sequencing.

2.3.3 University of Lisbon, Department of Bioengineering (1st Postdoc)

- January 2010 - December 2012. Co-PI of Márcia Mata (research fellow). Márcia's project involved studying the genetic stability of Mesenchymal Stem Cells (MSCs) during their *ex-vivo* expansion in bioreactors.

2.3.4 University of Lisbon, Department of Bioengineering (PhD)

- January 2007 - June 2007. Co-PI of Fanny Trouvat (research fellow). Fanny's project involved studying the genetic stability of non-viral DNA biopharmaceuticals under different environmental conditions.

2.4 Scientific collaborations

2.4.1 CEA, Genoscope (Lab Head)

- Thierry Naas (Hôpital Bicêtre AP-HP): Mutant construction and phenotype testing of *Klebsiella pneumoniae*.
- Suzana Salcedo (CNRS, Université de Lyon): Mutant construction and phenotype testing of *Acinetobacter baumannii*.
- Ryan Rego (Institute of Parasitology, AVCR): Epigenome of *Borrelia burgdorferi*.
- Multiple other individual and institutional collaborations within the frame of incoming sequencing projects.

2.4.2 Mount Sinai School of Medicine, Department of Genetics and Genomic Sciences (Senior Scientist)

- Richard Roberts (New England Biolabs): Diversity of methylation systems in *Clostridioides difficile*.
- Aimee Shen (Tufts University School of Medicine): Sporulation phenotype in *Clostridioides difficile*.
- Rita Tamayo (University of North Carolina at Chapel Hill): *In-vivo* testing of sporulation in *Clostridioides difficile*.

2.4.3 Institut Pasteur, Department of Genomes and Genetics (2nd Postdoc)

- Fernando de la Cruz (Universidad de Cantabria): Transmissibility of MGEs.

2.4.4 University of Lisbon, Department of Bioengineering (1st Postdoc)

- Jeffery M. Gimble (Tulane University): Genomic stability of adipose-derived human stem cells.

2.4.5 University of Lisbon, Department of Bioengineering (PhD)

- Kristala J. Prather (Massachusetts Institute of Technology): Development of genetically safer bacterial chassis for propagation of DNA biopharmaceuticals.
- George Church (Harvard Medical School): Multiplex accelerated genome engineering of Bacteria.
- Leslie A. Lewis (City University of New York): Insertion specificity of IS₂.

2.5 Editorial activities

- 2021 - Present: Member of the Editorial Board of [Microbiology](#)
- 2020 - 2021: Member of the Editorial Board of [Genomics](#)
- 2020 - Present: Member of the Advisory Board of [Heliyon](#)
- 2019 - Present: Member of the Editorial Board of [mSystems](#)
- 2019 - Present: Member of the Editorial Board of [Scientific Reports](#)
- 2019 - Present: Member of the Editorial Board of [Scientific Data](#)
- 2019 - Present: Member of the Editorial Board of [BMC Microbiology](#)
- 2019 - Present: Member of the Editorial Board of [BMC Biotechnology](#)
- 2017 - Present: Member of the Editorial Board of [PLoS ONE](#)
- 2017 - Present: Member of the Editorial Board of [Frontiers in Genetics](#)
- 2017 - Present: Member of the Editorial Board of [Frontiers in Microbiology](#)

2.6 Grant and peer-reviewing activities

- 2021 - 2022: Reviewer of grant applications for the US National Science Foundation (NSF) [Graduate Research Fellowship Program \(GRFP\)](#)-Microbial Biology Panel
- 2021: Reviewer of grant applications for the [Swiss National Science Foundation \(SNSF\)](#)
- 2020: Reviewer of grant applications for [Graduate Women in Science \(GWIS\)](#)
- [External Reviewer for multiple journals](#) (e.g.: [PLoS Computational Biology](#), [Nucleic Acids Research](#), [Scientific Reports](#), [Bioinformatics](#), [BMC Genomics](#), [Cell Death and Differentiation](#), [Stem Cells](#)).

2.7 Teaching activities

- 2021 - 2022: University of Évry (Master II in Bioinformatics). Invited academic seminars on *Advances in Sequencing Technologies*.
- 2014: University Paris 3 - Sorbonne Nouvelle. Invited academic seminar on *Therapeutic Potential of Human Stem Cells*.
- 2012: University of Lisbon - MIT Portugal PhD Program. Invited academic seminar on *Genomic Instability*.

2.8 Grants, fellowships, and awards

- 2021: *MYCOXPLORE* - Horizon H2020 MSCA. Role as partner (submitted).
- 2021: *Biodiversity Genomics Europe* - Horizon H2020. Role as partner (submitted).
- 2021: *Aplicaciones ómicas en sistemas biológicos para usos específicos en biodiversidad, salud y ambiente* - CONCYTEC. Role as partner.
- 2021: *Climate Genomics of Antarctic Toothfish (ClimGenAT)* - Institut Pierre-Simon Laplace. Role as partner.
- 2017: *Lactic Acid Bacteria as Cell Factories: A Synthetic Biology Approach for Plasmid DNA and Recombinant Protein Production* - Grant PTDC/BTM-SAL/28624/2017 attributed by the Portuguese Ministry of Science (FCT). Role as consultant.
- 2010: Post-Doctoral grant BPD/64652/2009 attributed by the Portuguese Ministry of Science (FCT).
- 2010: Award for top-cited article during 2008-2010.
- 2010: Ph.D. grant BD/22320/2005 attributed by the Portuguese Ministry of Science (FCT).
- 2007: Travel grant - 32nd FEBS congress.
- 2005: ERASMUS fellowship.

Past research

Contents

3.1 Ph.D. research (2006-2009)	18
3.1.1 Introduction	18
3.1.2 Deletion-formation events in DNA biopharmaceuticals	20
3.1.3 A predictive tool for estimating recombination frequency in plasmids	22
3.1.4 IS-mediated genetic instability in plasmid biopharmaceuticals	23
3.1.5 Transition to the first postdoctoral project	26
3.2 Postdoctoral project (2010-2013)	26
3.2.1 Introduction	26
3.2.2 Effect of hypoxia and prolonged passaging on the genomic stability of <i>in vitro</i> expanded human stem/stromal cells	28
3.2.3 Effect of hypoxia and prolonged passaging on mitochondrial performance	29
3.2.4 Role of non-canonical DNA structures and sequence motifs on human mitochondrial DNA instability	32
3.2.5 Transition to the second postdoctoral project	35
3.3 Postdoctoral project (2013-2016)	35
3.3.1 Introduction	35
3.3.2 The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts	37
3.3.3 Regulation of genetic flux between bacteria by restriction-modification systems	41
3.3.4 The chromosomal organization of horizontal gene transfer in bacteria	45
3.3.5 Transition to a Senior Scientist position	49
3.4 Senior Scientist (2016-2020)	49
3.4.1 Introduction	49
3.4.2 Methylome analysis reveals great epigenomic diversity in <i>C. difficile</i>	51
3.4.3 Comparative analysis of methylation sites across <i>C. difficile</i> genomes	52
3.4.4 Non-methylated sites are enriched in regulatory elements	52
3.4.5 Loss of methylation impacts transcription of sporulation genes	56
3.5 Conclusions	56

3.1 Ph.D. research (2006-2009)

3.1.1 Introduction

In the last three decades plasmid-based gene delivery has gained widespread and growing attention, particularly due to its promising potential to correct genetic defects and prevent or treat infectious diseases and cancer [Bunnell & Morgan 1998, Kay 2011]. When compared to their viral counterparts, plasmid DNA (pDNA) biopharmaceuticals offer appealing advantages, especially in terms of safety, stability, lower toxicity, and easier scalability of production. The levels of transgene expression attained with nonviral alternatives are, however, typically weaker and poorly sustained, and call for further improvements in current pDNA vector systems and their mode of delivery. A therapeutic pDNA molecule is typically built around a modular structure, comprising a unit essential for propagation in a bacterial host (usually *Escherichia coli*) and an eukaryotic transcription unit harboring, among other features, a Gene of Interest (GOI) and a polyAdenylation (polyA) signal (**Fig. 3.1a**). Remarkable improvements in the efficacy of these molecules have in many cases arisen by tinkering with different elements of its structure. Some examples, include the modification of polyA sequences to further increase transgene expression and resistance to nucleases [Azzoni *et al.* 2007], alterations in the type and load of CpG motifs [Coban *et al.* 2005], changes in codon bias [Uchijima *et al.* 1998], and the elimination of extraneous sequences [Mairhofer *et al.* 2008]. Such changes can have profound effects on the cellular uptake of pDNA molecules, intracellular stability, nuclear transport, and consequently duration of transgene expression.

Although DNA-based biopharmaceuticals are typically more stable than cell- or protein-based ones, one of the major problems encountered during their design is the assurance of structural integrity. The latter can be defined as a series of spontaneous events that culminate in an unforeseen rearrangement, loss, or gain of genetic material. Such events are frequently triggered by the transposition of MGEs or by the presence of instability-prone elements such as non-canonical (non-B) structures. Over the years, several groups have described the occurrence of multiple types of spontaneous mutations in therapeutic pDNA (reviewed in [Oliveira *et al.* 2009a, Oliveira & Mairhofer 2013]), the majority involving deletion-formation events mediated by direct/inverted repeats or transposition of IS elements (**Fig. 3.1b**). Despite the usual low mutation frequency (typically 10^{-9} - 10^{-3} total cells) [Oliveira *et al.* 2008], the detrimental impact of these contaminant populations in terms of productivity loss should not be disregarded, as their propagation may be favored under particular environmental conditions.

During my Ph.D. I focused on accessory regions of the plasmid backbone with the potential to engage in a wide range of structural instability phenomena. And addressed key points that remained unclear by the time: Where are these multiple unstable regions located? How many are they and which conditions favor their instability? What is their impact on plasmid safety? I started the analyses with a candidate plasmid model against rabies, and later extended it to several widely

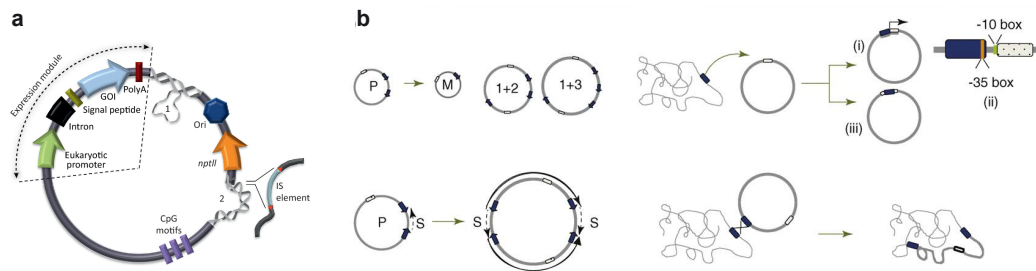


Figure 3.1: Structural instability in plasmid biopharmaceuticals. (a) Structural elements and possible detrimental sequence motifs on a conventional therapeutic plasmid. Structural elements can be organized into (i) a bacterial propagation unit and (ii) a eukaryotic expression unit. The bacterial expression unit typically contains an ori for the amplification of the plasmid in *E. coli* and an antibiotic resistance gene (here *nptII*) used for the selection of plasmid-bearing cells. The eukaryotic expression unit contains a promoter sequence, a 5' untranslated region (5' UTR), the therapeutic GOI, and a polyA. Also shown are examples of elements prone to undergo structural rearrangements. Situation 1 illustrates a slipped-misalignment between direct repeats. Situation 2 depicts the spontaneous transposition of an IS upstream of the *nptII* gene. Adapted from [Oliveira & Mairhofer 2013]. (b) Possible recombination events involving pDNA. (*upper left*) Recombination between direct repeats (blue arrows) in a parental (P) pDNA can give rise to monomeric (M) and heterodimeric (1+2, 1+3) deletion-formation products. (*bottom left*) Recombination between inverted repeats (blue arrows) results in head-to-head dimers bearing two pairs of interposed inverted repeats, inverted spacer sequences (S) and two giant inverted repeats (large arrows). (*upper right*) Transposition of ISs (blue box) from the chromosome (shown as coiled molecule) to pDNA can be detected as up-regulation of gene (white box) expression (i). This occurs via the generation of a hybrid promoter composed of a -35 region from the IS extremity (orange box) and a resident -10 hexamer (green box) present in pDNA (ii). Alternatively, ISs can promote gene inactivation or suppression by interrupting the coding region (iii). (*bottom right*) Recombination between similar regions (blue boxes) present in pDNA and in gDNA can lead to integration of the plasmid into the host genome. Adapted from [Oliveira *et al.* 2009a].

used bacterial and mammalian plasmids.

3.1.2 Deletion-formation events in DNA biopharmaceuticals

Plasmid pCIneo (pMB1-derived) is used as backbone to construct candidate DNA vaccines against rabies [Bahloul *et al.* 1998, Jallet *et al.* 1999]. It contains an ampicillin resistance (*amp^R*) gene for selection in bacteria, and a neomycin phosphotransferase (*neo^R*) gene under the regulation of an SV40 enhancer and early promoter region, which is used as a selectable marker in mammalian cells. Interestingly, I found that several *recA⁻* *E. coli* strains harbouring pCIneo were able to grow in the presence of the aminoglycoside antibiotic kanamycin after approximately 20–40 h [Ribeiro *et al.* 2008]. No growth was observed in control experiments carried out with the same strains devoid of plasmids. A subsequent restriction analysis of plasmids extracted from cells which had been subjected to kanamycin stress revealed the presence of unexpected low and high molecular-weight restriction fragments which were later found to be resulting from deletion-duplication events triggered by the presence of directly repeated sequences [Ribeiro *et al.* 2008]. Such recombination events were found to bring the *neo^R* gene closer to a promoter-like sequence located in the origin of replication [Ribeiro *et al.* 2010], thus allowing transcription of the gene under the presence of this selector. When *E. coli* cells were grown in liquid media in the absence of kanamycin, the number of recombined plasmid molecules remained approximately constant and at low levels (around 1.5×10^4 per 2×10^5 cells) (Fig. 3.2a). This means that 1 out of 13 cells would have a molecule of recombined plasmid (in *ca.* 200–300 molecules of parental plasmid), representing a contamination of about 0.02% in a purified plasmid batch. Because recombinant plasmids provided a competitive advantage to cells when growth was conducted under selective pressure, the number of recombined molecules reached a maximum of 3.0×10^7 per 2×10^5 cells in the same period of time (Fig. 3.2a).

In order to examine the influence of increasing kanamycin concentrations on the recombination products generated from pCIneo, I analyzed pDNA restriction patterns from several individual colonies of *E. coli*. (Fig. 3.2b) shows the variation in the relative percentage of each of the recombinant plasmid forms with 30, 50 or 70 $\mu\text{g ml}^{-1}$ of kanamycin. Twenty-four hours after plating, I observed a decrease in the percentage of monomeric (M) form with increasing kanamycin concentrations. This reduction was accompanied by a gradual increase in the percentage of the dimeric (1 + 2) form. Also, at 24 h, the trimeric (1 + 3) form only came out as a minor recombination product. This low abundance of the 1 + 3 form among recombination products was also reported by other authors [Bi & Liu 1996]. The higher selective pressure for the 1 + 2 form with increasing kanamycin concentrations likely reflects a compromise between faster replication rates of dimeric plasmids [Summers *et al.* 1993] and size-related low competitive advantage.

Despite the accumulating evidence of plasmid deletion-formation events, no consistent mapping of repeated regions had been performed at the time. This prompted me to analyze the density of direct, inverted and tandem repeats in 33 commonly

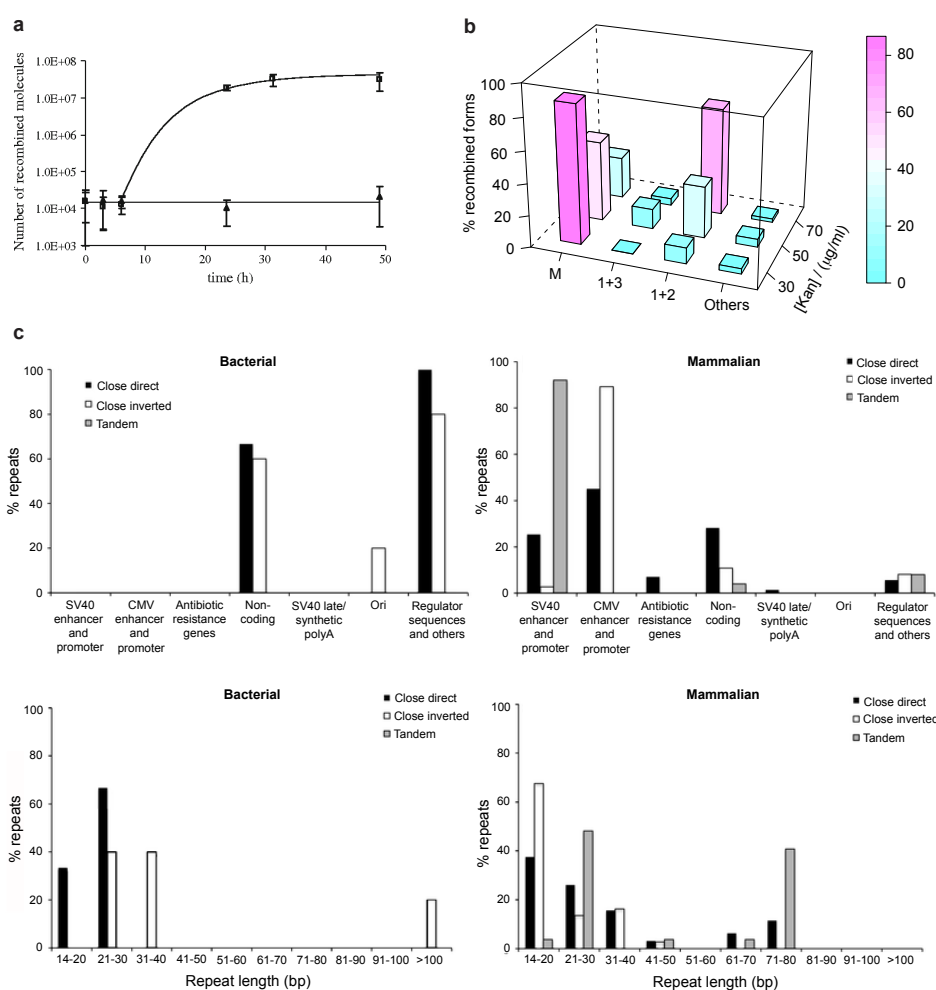


Figure 3.2: Repeat-mediated structural instability in commercial plasmid vectors. (a) Time-course evolution of pCIneo plasmid in DH5 α cells grown in LB liquid medium supplemented with 100 $\mu\text{g}/\text{mL}$ ampicillin (open triangles) or 30 $\mu\text{g}/\text{mL}$ kanamycin (open squares). The number of recombined molecules was determined by quantitative real-time PCR. Error bars represent standard deviations of at least three independent measures. Taken from [Ribeiro *et al.* 2008]. (b) Variation in the relative percentage of recombinants with varying kanamycin concentrations (between 30 and 70 $\mu\text{g ml}^{-1}$) (adapted from [Oliveira *et al.* 2009b]). (c) Repeat mapping in expression and cloning vectors shows that close direct and close inverted repeats are preferentially located within non-coding and regulator sequences (*upper left*), while in mammalian expression vectors, they essentially map within eukaryotic promoters and non-coding regions (*upper right*). The majority of close direct and inverted repeats found were 14–40 bp in length while tandem repeats had typical periods of 21 and 72 bp (*bottom*). Taken from [Oliveira *et al.* 2010].

used bacterial and mammalian expression vectors available from multiple commercial vendors [Oliveira *et al.* 2010]. I found an unexpectedly high density of repeats, particularly in mammalian expression vectors, with values as high as 3.8 pairs per kb for close direct repeats. In bacterial vectors, repeats were predominantly located within non-coding regions and regulatory sequences (such as transcriptional terminators or binding sites), whereas in mammalian vectors they mainly mapped to eukaryotic promoters and SV40 enhancer (**Fig. 3.2c**). Hence, multiple commercial vectors show an elevated recombination potential due to the presence of multiple repetitive elements. If used as backbones for therapeutic applications, there is a real risk of injecting numerous undetectable mutated plasmids with unknown biological properties. In this regard, I proposed a set of measures to curb these instability events, namely: *i*) the generation of minimal plasmids devoid of any non-essential and potentially detrimental backbone sequences; *ii*) the use of more stable and efficient promoters; and *iii*) avoiding the use of selective markers. It is thus foreseeable that the next few years witness exciting progress in the development of structurally fine-tuned plasmid molecules, which will unfold in tandem with an increased perception of the extent to which these systems may be applied across different research fields.

3.1.3 A predictive tool for estimating recombination frequency in plasmids

I next sought to develop a tool able to predict the frequency at which deletion-formation takes place in plasmids. Therefore, I proposed a non-linear mathematical function that accurately predicts recombination frequencies (F_R) in bacterial pDNA harbouring directly repeated sequences [Oliveira *et al.* 2008]. The mathematical function (**Eq. 1**), which was developed on the basis of published data on deletion-formation in multicopy plasmids containing direct-repeats (L_R , 14–856 bp) and intervening sequences (L_S , 0–3,872 bp), also accounts for the strain genotype in terms of its *recA* function.

$$F_R(L_R, L_S) = (A + L_S)^{-\frac{\alpha}{L_R}} \cdot \frac{L_R}{1 + B \cdot L_R + C \cdot L_S} \quad (\text{Eq. 1})$$

where A , B , C , and α are constants.

Development of **Eq. 1** was made using prior knowledge available on the trend of the dependence of recombination frequency on repeat length and on intervening sequence distance [Oliveira *et al.* 2008]. (**Fig. 3.3a**) summarizes the parameters determined by least squares regression, together with 95% confidence intervals estimated by bootstrap analysis for this equation. The differences observed in terms of presence/absence of the C parameter in both genotypes are likely to be related with the fact that *recA*-independent recombination is much more sensitive to the distance between the repeats than *recA*-dependent recombination is. Globally, the trend of the experimental data points is the same as that predicted using the respective equations (**Fig. 3.3b,c**).

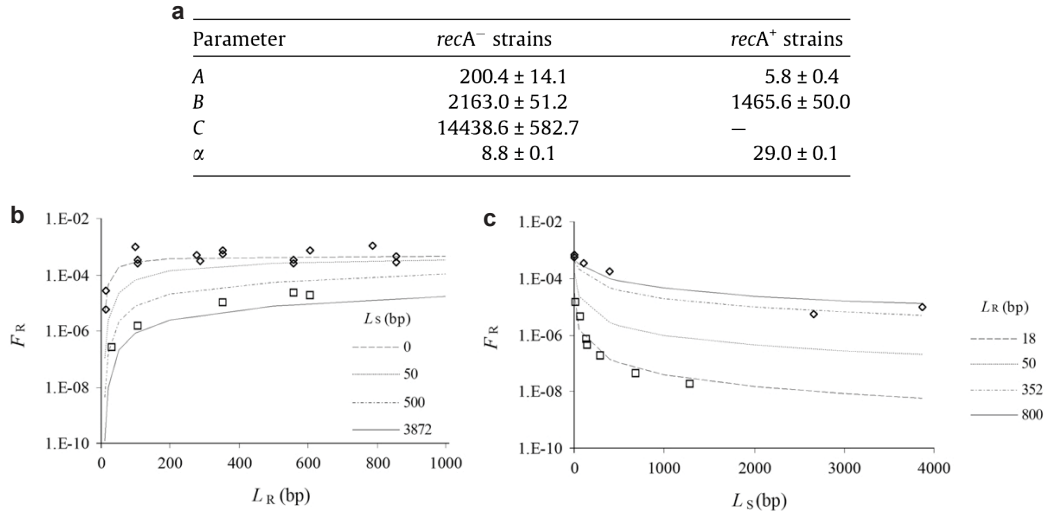


Figure 3.3: (a) **Eq. 1** parameters and confidence intervals (95%) calculated by bootstrap analysis for *recA*⁻ and *recA*⁺ strains. (b, c) F_R as a function of L_R (L_S) for fix values of L_S (L_R) and for *recA*⁻ strains [Oliveira *et al.* 2008]. Lines correspond to theoretical predictions (**Eq. 1**) and symbols correspond to experimental data for $L_S = 0$ or $L_R = 352$ bp (\diamond) and $L_S = 3,872$ or $L_R = 18$ bp (\square). Taken from [Oliveira *et al.* 2008].

All the data used for the development of **Eq. 1** refers to plasmids harbouring fully identical directly repeated sequences. This is particularly important because sequence composition and similarity are known to influence Homologous Recombination (HR) [Eckert & Yan 2000] and thus, F_R . Although I focused my attention on deletion frequencies in *recA*⁻ and *recA*⁺ strains, considerable deviations to the proposed equation should be expected when other recombinant-deficient strains are used. Interestingly, I found that predictions performed by **Eq. 1** closely matched recombination frequency values calculated for *Bacillus subtilis* chromosome harbouring direct repeats [Chedin *et al.* 1994], suggesting that, its applicability might be useful for other systems beside multicopy plasmids. A similar dependence of F_R on L_R and L_S has also been recently reported in the genome of *Acinetobacter baylyi* [Gore *et al.* 2006]. In conclusion, this study provides valuable and fairly accurate predictions that might be useful in similar systems.

3.1.4 IS-mediated genetic instability in plasmid biopharmaceuticals

ISs are small cryptic mobile elements, ubiquitous in most Eubacteria and Archaea, and tremendously relevant to the field of plasmid bioprocessing. They can severely impact plasmid function and yield, by leading to deletions and rearrangements, activation, downregulation or inactivation of neighboring gene expression

[Mahillon & Chandler 1998]. Such IS-containing molecules can become predominant throughout the culture stage, either as a result of high antibiotic selective pressure, or the presence of certain genes and regulatory elements capable of enhancing transposition frequency. The *kan^R* gene for example, in particular its 5' half and approximately 100 bp upstream, constitutes a hotspot for the transposition of several ISs, including IS1 [Prather *et al.* 2006], IS3 [Szeverenyi *et al.* 1996], IS150 [Szeverenyi *et al.* 1996], and IS186 [Szeverenyi *et al.* 1996]. In addition, I found this region to attract spontaneous IS2 transpositions as well [Oliveira *et al.* 2009b]. These mutants were able to subsist under kanamycin stress due to the generation of a hybrid promoter between the inserted 5' end of IS2 and an existing *kan^R* -10 box (**Fig. 3.4a**).

To evaluate the kinetics of IS2 accumulation on a smaller scale, typical of many laboratory settings, cells were grown under standard conditions (shake flask culture, LB medium, 37°C) for an initial period of 48 h after which they were used to reinoculate a new shake flask every 48 h [Gonçalves *et al.* 2014]. Such extended growth until late-stationary phase is frequently used during the overproduction of certain proteins or chemical compounds of biological interest, and it was used here as a worst-case scenario in terms of IS accumulation. The procedure was repeated four times until reaching a maximum accumulated growth time of 192 h. Plasmid samples taken at the end of each growth period were analyzed by real-time PCR for quantification of IS2 insertions (**Fig. 3.4b**).

These results show that detectable levels of transposition (approximately 1 in 2×10^7 plasmid molecules) can be attained at very early stages of cell culture (soon after transformation and during inoculum growth, time 0). Assuming a purely stochastic distribution of IS2-contaminated plasmids in the cell population, we can roughly estimate that at this stage, approximately 1 out of 6.7×10^4 cells will contain one plasmid with IS2 insertions (assuming an average number of 300 copies per cell). The results also point out that growing *E. coli* for one single round of 48 h is enough to quickly increase the number of IS2 insertions as much as 1,000-fold (**Fig. 3.4b**).

Asymmetries in DNA sequence composition have been previously linked to conformational changes capable of favorably influencing the integration of MGEs. For example, retroviral integration occurs more efficiently at the outer face of curved DNA segments composed of alternate AT- and GC-rich tracts [Muller & Varmus 1994]. Also, integration of T-DNA from *Agrobacterium tumefaciens* was described to take place in regions of the *Arabidopsis* genome coincident with inversely correlated and sharply asymmetric GC- and AT-skews [Schneeberger *et al.* 2005, Zhang *et al.* 2007]. Taking the above into consideration, I decided to perform an extensive literature search for IS2 target sites, in order to investigate whether these sites significantly correlate with the presence of nearby compositional asymmetries. For this purpose, I performed a literature search and gathered 41 IS2 target sites, each having 1.4 kb in length (± 0.7 kb flanking the target site) [Gonçalves *et al.* 2014]. GC and AT skew profiles were subsequently computed using 200 bp windows and a 100 bp step. I observed a negative to positive transition in GC skew and a rather symmetric and less-prominent inversion in

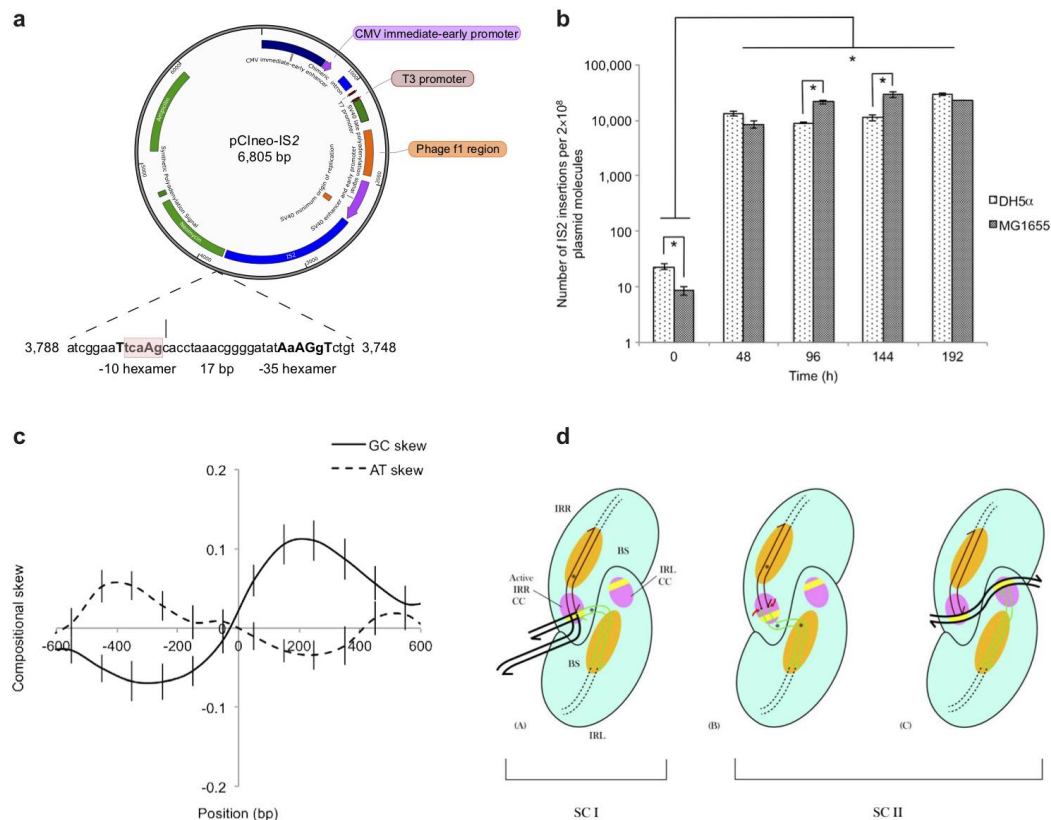


Figure 3.4: IS-mediated plasmid instability. (a) Plasmid pCIneo::IS2. Shown is the hybrid promoter generated by the IS2 -35 and pCIneo -10 hexamer sequences. Capital letters match the *E. coli* promoter consensus sequence TTGACAN_{16–18}TATAAT. The boxed region represents the 5 bp duplicated sequence generated upon IS2 insertion. The arrow indicates the junction site between the right inverted repeat of IS2 and pCIneo. Numeration (in bp) is relative to the beginning of the CMV immediate early enhancer and promoter region in pCIneo::IS2. (b) Effect of cultivation time under standard growth conditions (shake flask, LB medium, 37 °C) on the accumulation of IS2 in pCIneo. Four sequential growths of 48 h each were performed with two different *E. coli* host strains: DH5α and MG1655. (c) Average profiles of GC- and AT-skew 550 bp upstream and downstream of the insertion site of IS2. (d) Schematic model for the IS2 transposition pathway. Each synaptic complex is shown as a dimer with a DNA binding site (orange) to which protein binding domains are bound, and a catalytic center (pink). SC - Synaptic Complex. Error bars represent the SEM of at least three independent experiments. * $P < 0.05$. Adapted from [Gonçalves *et al.* 2014, Lewis *et al.* 2012].

AT skew, both taking place at the position corresponding to the insertion site (**Fig. 3.4c**). Concomitantly, IS2 target sites show elevated intrinsic curvature/bending [Lewis *et al.* 2012], that seem to play a crucial role in the transposition pathway (**Fig. 3.4d**). Bending results in a compaction of the DNA structure that might contribute to an increased level of sequence discrimination and a proper alignment of the target DNA within the catalytic domains of the transpososome. Altogether, our results extend the current knowledge on the mechanisms driving transposition of MGEs and open up exciting perspectives for further research. A deeper understanding of the detailed mechanism of IS transposition may allow us to conceive of redirecting insertion specificity in a predictable way [Guynet *et al.* 2009], or instead, pave the way for the design of safer and more evolutionary robust genetic circuits.

3.1.5 Transition to the first postdoctoral project

During my Ph.D. I was able to acquire multiple competences in the broad fields of molecular biology, microbial genetics, and biotechnology. Since many of the questions addressed during this period versed on events of genetic instability during biopharmaceutical development, I considered as valuable asset if I could acquire a broader view on this topic, by expanding my lines of research to eukaryotic cell-based biopharmaceuticals. More specifically, I investigated the onset of certain genomic aberrations during the *in vitro* expansion of human MSCs for clinical applications.

3.2 Postdoctoral project (2010-2013)

3.2.1 Introduction

The broad field of Regenerative Medicine brings the exciting promise of using stem cells and/or their progeny to replace injured tissues damaged by disease, either through the cell's integration (engraftment) into the target tissue and/or the cell's ability to deliver soluble signaling factors. Stem cells can be derived from multiple tissues, namely from embryonic and adult sources. Over the last three decades, a wealth of exciting research findings on the immunomodulatory, antiapoptotic, proangiogenic, and anti-inflammatory properties of MSCs appears to lend further support to their use as a promising cell source in the treatment of immune disorders and in tissue repair [Gao *et al.* 2016, Dimarino *et al.* 2013]. With the increasing demand of human adult stem cells for both research and clinical purposes (typically 1–5 million cells per kg of body weight are required per treatment, it becomes of utmost importance to bridge the gap between the need to expand the cells *in vitro* and the capability of harnessing the factors underlying replicative senescence. Adult stem cells are known to have a limited lifespan *in vitro* and to enter replicative senescence almost undetectably upon the start of *in vitro* culturing [Bonab *et al.* 2006]. This process is typically characterized by a gradual decrease in proliferative and differentiation capacity, morphological changes (cells typically display enlarged, flattened, and more irregular shapes), high levels of tumor suppressors, and accelerated telom-

ere erosion [Baxter *et al.* 2004]. For safety reasons, the number of population doublings (PDs) should be kept at reasonably low levels (typically below 20) if clinical applications are envisaged [Sensebe *et al.* 2011]. And despite the fact that adult stem cells typically divide less frequently than pluripotent stem cells, they are also prone to acquire chromosomal aberrations during expansion in culture.

Multiple studies have been carried out to address the effect of hypoxic preconditioning on the genomic stability of human adult stem cells, often with contradictory results. Some authors point to enhanced structural instability and aneuploidy events at early passages (P1–P7) under low oxygen (5% O₂) [Ueyama *et al.* 2012]. Others describe physiological O₂ concentrations (1–7%) to significantly reduce or prevent chromosomal aberrations [Holzwarth *et al.* 2010, Estrada *et al.* 2012, Li & Marban 2010, Tsai *et al.* 2011]. Rodríguez-Jiménez *et al.* [Rodríguez-Jimenez *et al.* 2008] have shown that low oxygen environments (1% O₂) repress the MisMatch Repair (MMR) system through epigenetic chromatin inactivation and diminished SP1 binding, resulting in increased MicroSatellite Instability (MSI) in mouse neural stem cells and human Bone-marrow MSCs (BMSCs) as soon as 6 h.

Hence, quality control during *in vitro* expansion becomes critical for a safer clinical implementation of stem cell therapies. In their Reflection Paper on Stem Cell-Based Medicinal Products, the European Medicines Agency (EMA) highlighted the tumorigenic potential associated with manipulation steps and culture of pluripotent and somatic cells and made recommendations on performing cytogenetic analysis and evaluating parameters such as telomerase activity, proliferative capacity, and senescence status [EMA 2011]. Similar concerns have been addressed by the International Stem Cell Banking Initiative (ISCBI) [Initiative 2009]. In this matter, the US Food and Drug Administration (FDA) has stepped up its oversight of the increasing number of clinics usually operating under poorly regulated jurisdictions and offering unproven treatments against a myriad of pathologies (reviewed in [Lysaght & Campbell 2011]). The lack of a sound and reliable scientific follow-up has in some cases led to fatal outcomes [Cyranoski 2010].

Hence, I developed in this postdoctoral work a side by side comparison of human BMSCs and Adipose-derived Stem Cells (ASCs) concerning the impact of prolonged passaging (>P10, PP) and use of a low oxygen tension (2%) on the expression of the proto-oncogene *c-MYC* and tumor suppressor gene *p53*, as well as on critical genes that mediate MMR, HR, and Non-Homologous End-Joining (NHEJ). I further examined the influence of the aforementioned conditions in the onset of MSI, changes in telomere length, and mitochondrial performance. In addition, to gain insight into so far unnoticed discrepancies between BMSCs and ASCs responses to hypoxia, these results provide data potentially useful in the specification of quality-control requirements for stem-cell-based products for use in Cellular Therapy applications. Moreover, they extend the current knowledge on the continuous and far reaching changes occurring at the cellular level, that need to be taken into account when considering the *in vitro* expansion of stem cells for therapeutic applications.

3.2.2 Effect of hypoxia and prolonged passaging on the genomic stability of *in vitro* expanded human stem/stromal cells

Stem/stromal cells consist of a population of non-hematopoietic multipotent progenitors, able to differentiate into different mesodermal cell lineages, and which can be isolated from adult tissues (e.g. bone marrow, the most studied source) or postnatal tissues (e.g. umbilical cord matrix). An increasing body of evidence has demonstrated the immense clinical potential of these types of stem cells, mainly resulting from their immunomodulatory properties and multilineage differentiation ability [Strauer *et al.* 2002, Chen *et al.* 2004, Horwitz *et al.* 2001, Lazarus *et al.* 2005]. Since BMSC titers are typically very low and decline with age (roughly 0.01% of bone marrow mononuclear cells in a newborn in opposition to 0.00005% in elderly patients) [Caplan 2007], it becomes necessary to expand these cells *in vitro* in order to achieve clinically relevant cell numbers. Concurrently, other sources such as the adipose tissue are emerging as an alternative source of stem cells, which can thus be obtained in a less invasive way, and in larger titers than those found in the bone marrow [Gimble & Guilak 2003].

The process of expanding adult stem cells *in vitro*, has proven to be challenging not only from the bioengineering standpoint [Santos *et al.* 2011], but also from the perspective of maintaining stem cell characteristics while avoiding or delaying senescence and genetic instability [Tarte *et al.* 2010, Sensebe *et al.* 2011, Ueyama *et al.* 2012]. Moreover, the use of low oxygen tensions (1–5%, referred to as hypoxia) has been exploited as a strategy to mimic the physiologic microenvironment of these cells in order to increase cell proliferation [Dos Santos *et al.* 2010, Tsai *et al.* 2011]. Nevertheless, the mechanisms involved are poorly understood and some conflicting results have been obtained.

Since some studies have implicated hypoxia as a promoter of “stemness” and cell proliferation, I decided to investigate its effect as well as that of PP, on the expression of genes involved in HR (*RAD51*, *BRCA1*), NHEJ (*Ku80*), and MMR (*MLH1*) pathways. Low passage (P1–P3) human BMSCs and ASCs were exposed to normoxic (20% O₂) or hypoxic (2% O₂) conditions for a period of time ranging from 6 h to 21 days. After this 3-week period, normoxic cells were allowed to expand for a total of 10–20 passages (PP), while hypoxic cells were submitted to a reoxygenation period of 48 h (AR, after reoxygenation).

In the first 21 days under normoxia, the expression levels of the four genes did not differ significantly from that of the control for both cell sources. (Fig. 3.5a, left). However, a surprising and rather pronounced (3- to 5-fold) increase in gene expression was observed mainly within the first 24 h upon plating for the two HR genes and for both cell sources, regardless of the O₂ tension studied (Fig. 3.5a), with exception of *BRCA1* in hypoxic BMSCs). Since *BRCA1* was implicated in the adhesion, spreading, and motility of breast cancer cells by interaction with F-actin and the ezrin/radixin/moesin complex [Coene *et al.* 2011], it thus seems plausible to assume that this gene together with *RAD51*, may also play an important role in the adhesion of stem cells to culture-treated plastic. Prolonged passaging in normoxia

led to a significant down-regulation of *Ku80* and *BRCA1* in BMSCs, and *RAD51* and *BRCA1* in ASCs.

Under hypoxia, I observed a down-regulation of *MLH1* and *BRCA1* as soon as 6 h, but a more generalized repression of the four genes was only visible after 21 days. After a 48 h period of reoxygenation, cells previously exposed to hypoxia had expression levels close to those of time 0, with the exception of *RAD51* whose over-expression was still around 4–5-fold (**Fig. 3.5a**). The latter is probably an effect reminiscent of cell passaging, which points to a delayed recovery of *RAD51* basal levels of expression. I also evaluated the impact of hypoxia and PP on the expression of the tumor suppressor gene *p53* and the *c-MYC* proto-oncogene. Despite their well-known involvement in a plethora of biological processes that include cell-cycle progression, DNA replication, apoptosis, and angiogenesis, studies focusing on the expression levels of *p53* and *c-MYC* during *in vitro* expansion of stem cells are rather limited and have generated conflicting results. I found that for both types of cells the responses of *p53* and *c-MYC* to hypoxia and PP are similar. Upon long-term cultivation involving consecutive passages (> P10), for both BMSCs and ASCs, these genes were down-regulated (**Fig. 3.5b**), as previously reported by other authors [[Kim et al. 2009](#), [Efimenko et al. 2011](#)].

Since down-regulation of DNA repair genes is known to induce genomic instability [[Vaish 2007](#), [Rodriguez-Jimenez et al. 2008](#)], I further examined the status of several microsatellite sequences present in the genomic DNA of both cell sources. From day 14 onwards, I observed fragments absent in the dinucleotide markers *D17S250* and *D2S123* of both BMSCs and ASCs cultured under hypoxia, whereas the same instabilities and changes in *D5S346* were only seen after PP for “normoxic” cells (**Fig. 3.5c**). Since MSI was found to be present in only one or two markers, it should be respectively classified as stable (MSI-S) or low frequency (MSI-L).

3.2.3 Effect of hypoxia and prolonged passaging on mitochondrial performance

The evaluation of mitochondrial function as a quality control procedure during expansion of stem cells, has gained increasing interest in the recent years. Nevertheless, some aspects remain poorly explored, such as a thorough search for genetic aberrations in mtDNA. The results obtained in this work support the concept of a relatively stable mitochondrial genome, both in terms of small indels as well as larger aberrations. The majority of mutations found were haplogroup-specific or occurred at highly polymorphic repeat tracts, and their load did not change under hypoxia or after PP (**Fig. 3.6a**). One such example is the mononucleotide C tract ($C_{6-10}TC_6$) commonly termed D310, which is located within the displacement loop (D-loop) region. Mutations found within the D310 region were observed in all samples with exception of those from donor 2, and consisted in 1 or 2-bp C insertions (thus lying within the polymorphic length range of 6 C–10 C [[Legras et al. 2008](#)]) (**Fig. 3.6a**).

The mutation m523-524dAC found in BMSCs from donor 2 was relatively fre-

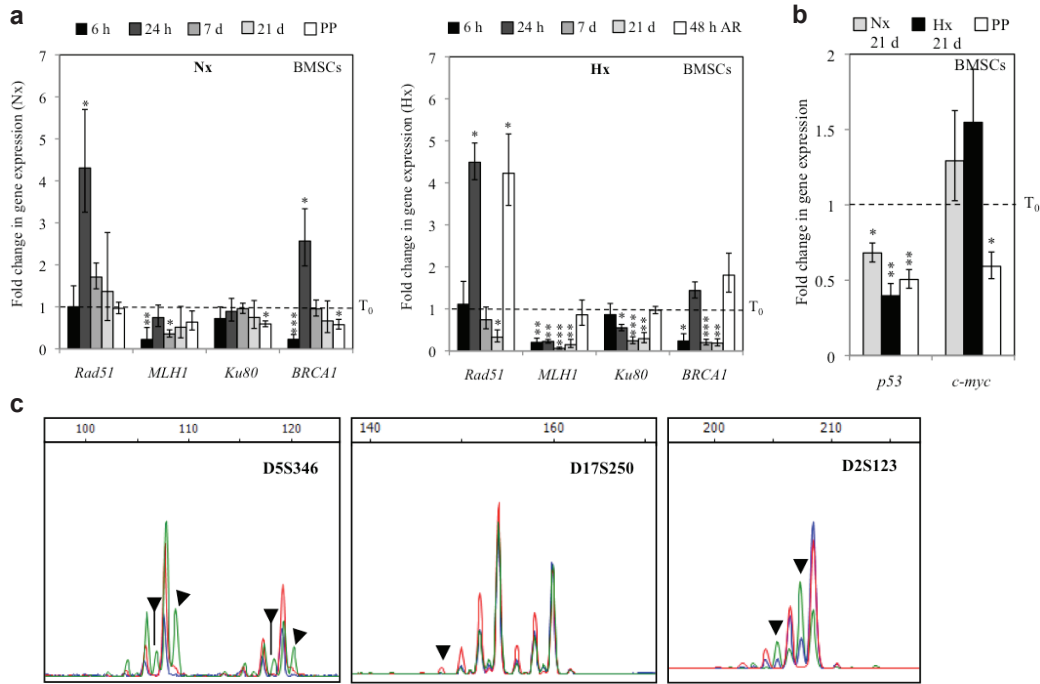


Figure 3.5: Gene expression analysis and microsatellite instability. (a) Changes in the expression level of DNA repair genes, tumor suppressor *p53*, and *c-MYC* proto-oncogene in BMSCs expanded for different incubation times under normoxia (Nx), hypoxia (Hx) and 48 h after reoxygenation (AR). Fold change in the expression of HR (*RAD51*, *BRCA1*), NHEJ (*Ku80*), and MMR (*MLH1*) genes, normalized to T₀. (b) Fold change in the expression of *p53* and *c-MYC*, normalized to T₀. *GAPDH* was used as housekeeping gene. (c) Examples of microsatellite instability (MSI) in the Bethesda panel of markers (*BAT25*, *BAT26*, *D5S346*, *D17S250*, and *D2S123*) as well as in *TP53Alu* and *RB*. Representative electropherograms depicting unstable loci are shown (T₀ — blue; Hx 21 days — red; PP — green), with additional or missing peaks indicated by arrows. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Taken from [Oliveira *et al.* 2012].

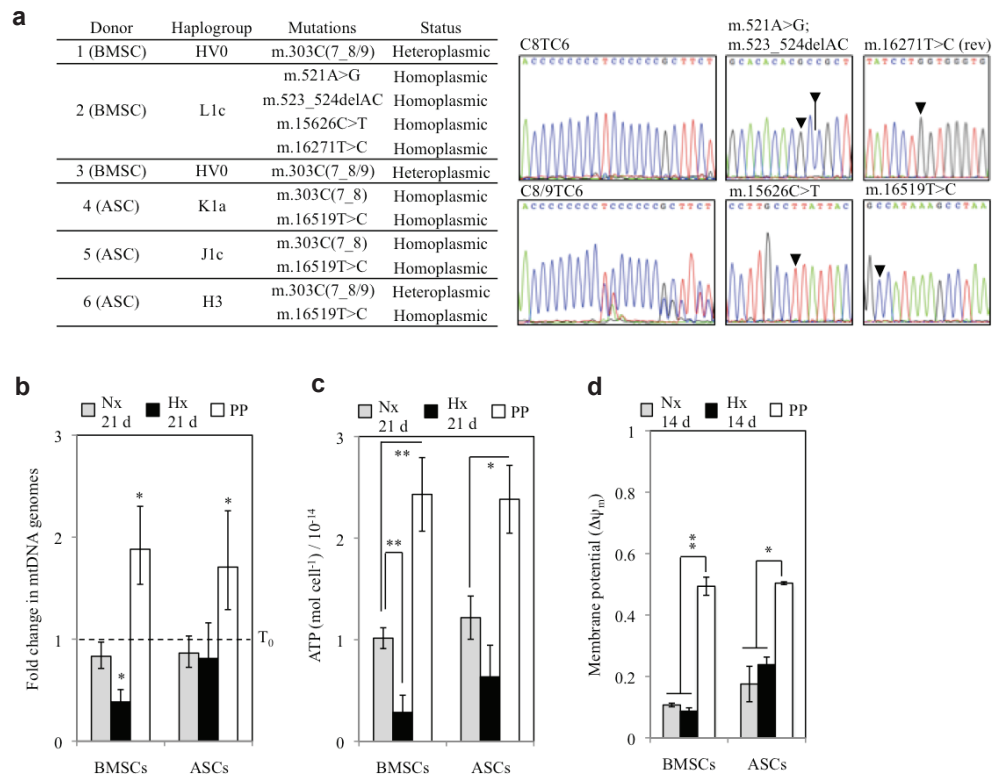


Figure 3.6: Evaluation of mitochondrial properties in stem cells. (a) Sequencing of the complete mitochondrial genomes of three BM and three ASC donors allowed me to determine the corresponding haplogroup, as well as non-haplogroup variations. The latter were essentially common variants located in the D-loop region of mtDNA and their number or status was not observed to change during PP or expansion under hypoxia. One exception was the m.521A > G variation, which was not found in specialized databases, thus representing, to the best of our knowledge, a novel homoplasmic polymorphism. PP led to an increase of mtDNA content (b), intracellular ATP (iATP) (c), and membrane potential ($\Delta\Psi_m$) (d), whereas expansion under hypoxia led to a decrease of more than twofold in the mtDNA content and iATP of BMSCs. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Taken from [Oliveira *et al.* 2012].

quent arising from slippage events occurring within a short microsatellite of 5 tandem AC repeats. Both the m.15626C > T synonymous transition located in the cytochrome b gene and m.16271T > C located in the D-loop have been previously identified [Ingman & Gyllensten 2006]. Yet, a thorough search in the literature as well as in specialized databases (www.mitomap.org) led me to conclude that m.521A > G is most probably, a novel homoplasmic polymorphism. Another variant found in some of our samples was 16519T>C (**Fig. 3.6a**), which has been shown to be related to an increased risk of development of several types of carcinomas, or to be in linkage disequilibrium with functional variants that increase that risk [Bai *et al.* 2007, Peng *et al.* 2011]. Since some of the latter mutations map within the non-coding D-loop region involved in the regulation of replication and transcription, I hypothesized that changes in copy number could occur. I did not find evidence of functional impairment of mitochondrial activity as a result of the presence of these mutations. In particular, I did not find significant differences in the mitochondrial content of the several donors (data not shown), suggesting that none of these mutations is *per se* the causative agent of any detrimental effect.

I further evaluated the effect of hypoxia and PP on other performance parameters such as the number of mitochondrial genomes, intracellular ATP (iATP), and membrane potential ($\Delta\Psi_m$). I observed that all of the latter parameters roughly increased 1.8-fold after prolonged passaging for both cell sources (**Fig. 3.6b-d**). These changes are characteristic of a more mature state associated with differentiation or senescence [Rehman 2010, Prigione *et al.* 2010]. After 21 days in hypoxia, the number of mtDNA genomes decreased more than half in BMSCs, whereas no significant difference was seen in ASCs (**Fig. 3.6b**). These results were found to correlate with variations in iATP levels for both types of cells (**Fig. 3.6c**). Taking into consideration the role of *p53* in maintaining mtDNA oxidative phosphorylation and copy number [Lebedeva *et al.* 2009], its hypoxia-mediated down-regulation likely contributed to the observed decrease in iATP content and number of mitochondrial genomes. Moreover, the decrease in mtDNA content and the previously observed down-regulation of *RAD51* under hypoxia are two events that apparently cannot be dissociated. The mtDNA was recently shown to be a substrate for *RAD51*, in a pathway that possibly facilitates the completion of mtDNA replication in the presence of DNA lesions [Sage *et al.* 2010]. Older passage cells also have an increased amount of ROS (as shown by changes in $\Delta\Psi_m$) (**Fig. 3.6d**) and are more exposed to oxidative damage, a fact that may jeopardize their potential clinical use.

3.2.4 Role of non-canonical DNA structures and sequence motifs on human mitochondrial DNA instability

Rearrangement of mtDNA and consequent loss of mitochondrial function has been implicated in the aging process and in a broad range of clinical phenotypes (reviewed in [Tuppen *et al.* 2010]). Short direct and inverted repeats are found to flank the majority of such rearrangements ($\sim 85\%$), a fact that has led to the assumption that deletion formation arises from slipped mispairing during replication or repair

of damaged DNA [Krishnan *et al.* 2008]. It appears, however, that no significant correlation exists between the density of repeat pairs and distribution of deletion breakpoints [Samuels *et al.* 2004], suggesting that additional factors are likely to contribute to the mutational spectra.

Here I evaluated the importance of alternative non-canonical structures (*e.g.*: intrinsically curved DNA, G-quadruplexes, triplex DNA and Z-DNA), whose impact on mtDNA instability is unknown or has so far been poorly explored in the literature. I observed a clear preference for 5' breakpoints to map in the vicinity of position 7.7 kb and 3' breakpoints to map in the vicinity of positions 14.5 and 16.1 kb (**Fig. 3.7a**). Well-known examples of deletion hotspots located in the close vicinity of these positions are the "common deletion" (nucleotide positions 5' 8,470–8,482; 3' 13,447–13,459) or the displacement loop (D-loop) 16,070 regions. The large majority of human mtDNA deletions (86%) affect solely the major arc (nucleotide positions 5,799–16,569 and 1–109), 2% affect the minor arc (nucleotide positions 442–5,720) and 12% affect the origins of replication (nucleotide positions 110–441 and 5,721–5,798 respectively for O^H and O^L) (**Fig. 3.7b**).

It was recently shown that intra-strand DNA hairpins and cloverleaf-like elements are enriched in common breakpoint sites of the human and mouse mitochondrial genomes [Damas *et al.* 2012]. These observations prompted me to investigate if breakpoints were preferentially located within or in the close vicinity of other classes of non-B DNA elements. The intrinsic flexibility of a DNA molecule (bendability) and its tendency to form a bent structure in the absence of external forces (curvature propensity) are parameters commonly used to describe secondary structure. A highly bendable molecule is less rigid, and does not necessarily retain intrinsic curvature as it allows a mixture of many different conformational states [Perez-Martin & de Lorenzo 1997]. Thus, regions having high curvature/bendability ratios are more prone to adopt curved and rigid conformations with elevated topological stress. In this sense, I decided to evaluate if human mitochondrial deletion breakpoints were preferentially clustered in regions under high torsional stress, and if known hotspots such as the "common deletion" are located in regions with particularly high ratios. The highest breakpoint densities were found in those regions with the highest curvature/bendability ratios, and departed significantly from density values estimated to occur randomly (**Fig. 3.7c**). Despite the generalized decrease in breakpoint density observed at ratios below 3, 90.5% of all breakpoints locate in the close vicinity of regions with ratios above the average value for the human mitochondrial genome (0.85). Regarding other non-B conformations, I found five G-tetrads, three sequences prone to generate triplex DNA and one sequence prone to generate Z-DNA (**Fig. 3.7d**). The local average density of breakpoints found in these nine sequences was 14.0 per 0.1 kb, which corresponds to a fold increase of 1.5 when compared to the average for the mitochondrial genome. When I compared the real breakpoint densities within these structures with those predicted from the random and partially random models, I verified that only the G2 and G5 elements were significantly enriched in breakpoints (**Fig. 3.7d**). The average breakpoint density found for these two elements was 41.7 per 0.1 kb, which

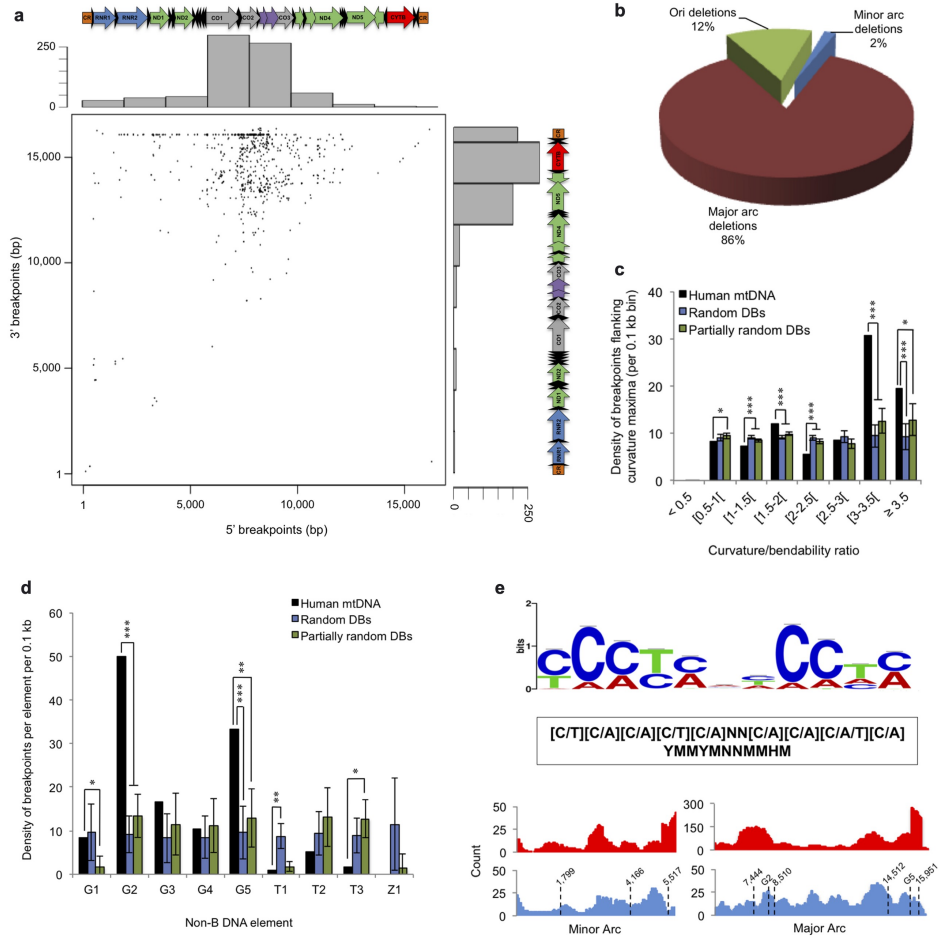


Figure 3.7: non-B DNA and human mitochondrial DNA instability. (a) Distribution of the 5' and 3' positions corresponding to 1,508 breakpoints, as well as corresponding histograms and positions along the mitochondrial genome. CR-control region; RNR-Ribosomal RNA; ND-NADH dehydrogenase; CO-cytochrome oxidase; CYTB-cytochrome B. Black arrows correspond to tRNA genes. (b) Pie chart indicating the proportion of deletions occurring exclusively in the major arc (positions 1–109 and 5,799–16,569), minor arc (442–5,720) or involving the origins of replication O^H (110–441) and O^L (5,721–5,798). (c) Density of deletion breakpoints (Σ number of breakpoints/ Σ fragment sizes) computed in 0.1 kb bins flanking each curvature maximum (black bars). Also shown are the density values computed after randomization of breakpoint positions (shown in blue and green). (d) Density of deletion breakpoints per 0.1 kb bins of each non-B element without (black) or with (blue, green) randomization. Error bars represent standard deviations. (e) Sequence logo of the degenerate 11-mer motif over-represented in the close vicinity (± 15 bp) of the non-repeated breakpoint dataset. Distribution profiles of breakpoints (red) and 11-mer motif (blue) along the minor arc (left) and major arc (right). Stippled lines indicate the positions of highly bent regions as well as G2 and G5 motifs. Adapted from [Oliveira *et al.* 2013].

corresponds to a fold increase of 4.6 and 3.3 when respectively compared to the genome average and major arc densities. These observations on the presence of compositional asymmetries near unstable regions are not only in line with my previous findings of non-B elements, but together with literature evidence, raise the possibility for the presence of other over-represented motifs. To evaluate this scenario, I carried out a search for conserved motifs in the close vicinity (± 15 bp) of our non-repeated breakpoint dataset ($n = 1,115$). For this purpose, as well as to attain more reliable conclusions on over-represented motifs, I used two different motif discovery tools, MEME and AlignAce, followed by motif edge trimming using STAMP. An 11-mer degenerate consensus YMMYMNNMMHM) was found to be over-represented in our dataset (**Fig. 3.7e**). This motif occurs 469 times in the human mtDNA, and was found to be over-represented when compared to shuffled mitochondrial genomes [Oliveira *et al.* 2013]. 50.3% of all mtDNA breakpoints were observed at a distance of less than 5 bp from one of such motifs [Oliveira *et al.* 2013]. Also, the distribution of this motif was positively correlated with the distribution of deletion breakpoints in both arcs (Spearman $\rho = 0.38$; $P < 0.001$) (**Fig. 3.7e**). Still, some regions depart from this tendency and show extremely high counts of breakpoints, despite a weak increase in the number of YMMYMNNMMHM motifs (e.g. nucleotide positions position 5.5 in the minor arc and 7.5 and 16 kb in the major arc) (**Fig. 3.7e**). Bearing in mind the findings on intrinsic curvature, these local discrepancies correlate with the nearby presence of highly bent regions at positions 5,517, 7,444 and 15,951, which as I mentioned previously, likely play a destabilizing role in these regions. Hence, the findings here described here may help to understand and redefine the multiple mechanisms by which deletion formation occurs in the human mitochondrial genome.

3.2.5 Transition to the second postdoctoral project

During my Ph.D. and first postdoctoral project I was able to gain multiple competences in wet lab environments (studying both prokaryotic and eukaryotic genomes). At this stage I considered as valuable asset if I could develop stronger bioinformatic skills. This would allow me to strengthen the ensemble of my competences, and acquire a more complete and competitive scientific profile. Nevertheless, I continued to work on aspects of genome dynamics, in particular, I investigated the interplay between prokaryotic defense systems and MGEs, and delved into the chromosomal organization of HGT.

3.3 Postdoctoral project (2013-2016)

3.3.1 Introduction

The flow of genetic information between bacterial cells by HGT drives bacterial evolution [Treangen & Rocha 2011, Gogarten *et al.* 2002], and R-M systems are key moderators of this process [Thomas & Nielsen 2005, Labrie *et al.* 2010]. They are

thought to be ubiquitous in Bacteria and Archaea [Makarova *et al.* 2013], and operate like many poison-antidote systems: they typically encode an MTase function that modifies a particular sequence and a Restriction Endonuclease (REase) function that cleaves a DNA when its recognition sequence is unmethylated [Mruk & Kobayashi 2014, Loenen *et al.* 2014, Ishikawa *et al.* 2009]. The three classical types of R-M systems differ in their molecular structure, sequence recognition, cleavage position and cofactor requirements [Roberts *et al.* 2003]. R-M systems are major players in the co-evolutionary interaction between MGEs and their hosts. Closely related strains have different systems and distantly related species sometimes have similar systems, suggesting frequent HGT. Incoming DNA is unlikely to be modified in a way compatible with the R-M systems of the new host and will be degraded. This has led to very early proposals that R-M systems are bacterial innate immune systems [Arber & Linn 1969], since they effectively allow self- from non-self discrimination. R-M systems might preferentially cluster with and stabilize other antiviral defense systems in so-called defense islands, *i.e.* discrete DNA segments that include a plethora of defense systems [Makarova *et al.* 2013, Makarova *et al.* 2011]. In some cases, different defense systems have been shown to operate synergistically in order to increase the overall resistance to phage infection [Dupuis *et al.* 2013]. Some R-M systems can also propagate horizontally in a selfish way. Incoming DNA carrying an R-M system induces ‘genetic addiction’ to the host by post-segregational killing [Naito *et al.* 1995]. This behavior leads to the stabilization of MGEs against challenge by competitor elements as long as the R-M system is present [Kulakauskas *et al.* 1995, Takahashi *et al.* 2011, Mochizuki *et al.* 2006]. Accordingly, genes encoding R-M systems have been reported to move between prokaryotic genomes within MGEs [Kulakauskas *et al.* 1995, Kita *et al.* 2003, Burrus *et al.* 2001, Kobayashi *et al.* 1999].

The study of R-M systems is at a key point in time. On the one hand, a number of studies have enlarged the known scope of activity of these systems in bacterial cells [Makarova *et al.* 2011, Dupuis *et al.* 2013, Fukuda *et al.* 2008], and a single resource, REBASE [Roberts *et al.* 2015], has re-grouped most of this information. On the other hand, the recent availability of tools to characterize bacterial methylomes is opening new perspectives on the effect of R-M systems in bacterial epigenetics [Furuta *et al.* 2014, O’Connell Motherway *et al.* 2014]. However, there is a lack of recent studies on some of the original questions put forward regarding R-M systems. How abundant are these systems? Which are more abundant? How rapidly do they evolve? How many systems are actually in MGEs? Which MGEs? Is there an association between R-M systems and different mechanisms of genetic mobility? In this first postdoctoral work, I used a comparative genomics approach to answer these questions.

3.3.2 The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts

After having developed a dedicated set of Hidden Markov Model (HMM) profiles for all MTase and REase types, I used them to query all complete prokaryotic genomes from Genbank [Oliveira *et al.* 2014b]. I identified a total of 4,743 R-M systems in 2,261 prokaryotic genomes (Fig. 3.8a). Type II systems are the most intensely studied and were also the most abundant (42.4%). Type IIC systems, in which the REase and MTase are part of the same polypeptide, accounted for more than a third (38.8%) of all Type II R-M systems. Type I were the second most abundant, corresponding to $\sim 29.5\%$ of all R-M systems. Type IV (methylation targeted) REases were found to be much more abundant (19.9%) than Type III (8.2%). I found similar trends in the relative amounts of R-M systems when the analysis was performed in chromosomes and plasmids separately. This large number of Type IV REases is somewhat surprising, given the very few studies devoted to them. Thus, the frequency of R-M systems was found to vary widely among bacterial large phyla (Fig. 3.8b), and to depend on genome size, taxonomy and lifestyle.

I observed a positive correlation between the total number of R-M systems and genome size (Spearman's $\rho = 0.2256$, $P < 10^{-4}$) (Fig. 3.8c). For small genomes (< 2 Mb), there is a quick increase in the number (Spearman's $\rho = 0.4758$, $P < 10^{-4}$) and density (Spearman's $\rho = 0.3810$, $P < 10^{-4}$) of R-M systems with genome size. For larger genomes (≥ 2 Mb), the average number of R-M systems is nearly independent of genome size (Spearman's $\rho = -0.0284$, $P > 0.2$) and kept around two per genome. Accordingly, the density of R-M systems decreased with increasing genome size for this group (Spearman's $\rho = -0.3434$, $P < 10^{-4}$).

Several works have shown that R-M systems stabilize plasmids in cells by their addictive behavior [Dupuis *et al.* 2013, Kulakauskas *et al.* 1995, Mochizuki *et al.* 2006]. Surprisingly, I found very few systems in plasmids when compared to chromosomes (219 versus 3802) (Fig. 3.9a). The rarity of R-M systems in plasmids might be the consequence of the presumably lower genetic mobility of these elements. To test this hypothesis, I divided the plasmids into two classes: plasmids encoding the conjugation machinery or at least the relaxase that allows them to be mobilized in trans by another conjugation machinery (MOB⁺, 44.6% from total), and plasmids lacking even the relaxase (MOB⁻, 55.4%). More MOB⁺ than MOB⁻ plasmids were found to contain R-M systems (respectively 113 versus 75, $P < 10^{-4}$; Chi-square test) (Fig. 3.9a). I then aimed at identifying if some types of R-M systems were over- or under-represented in certain types of MGEs. For this, I computed the observed/expected (O/E) ratios of the number of each type of R-M system present within plasmids, prophages and Integrative Conjugative Elements / Integrative Mobile Elements (ICEs/IMEs) (Fig. 3.9b). Type III systems appear as particularly over-represented in ICEs/IMEs ($P < 10^{-4}$; Chi-square test). Type IV REases are under-represented in all MGEs, which is consistent with their role in defense against invading epigenetic DNA methylation systems [Fukuda *et al.* 2008]. The most compact systems (Type IIC) are overabundant in all MGEs, and especially

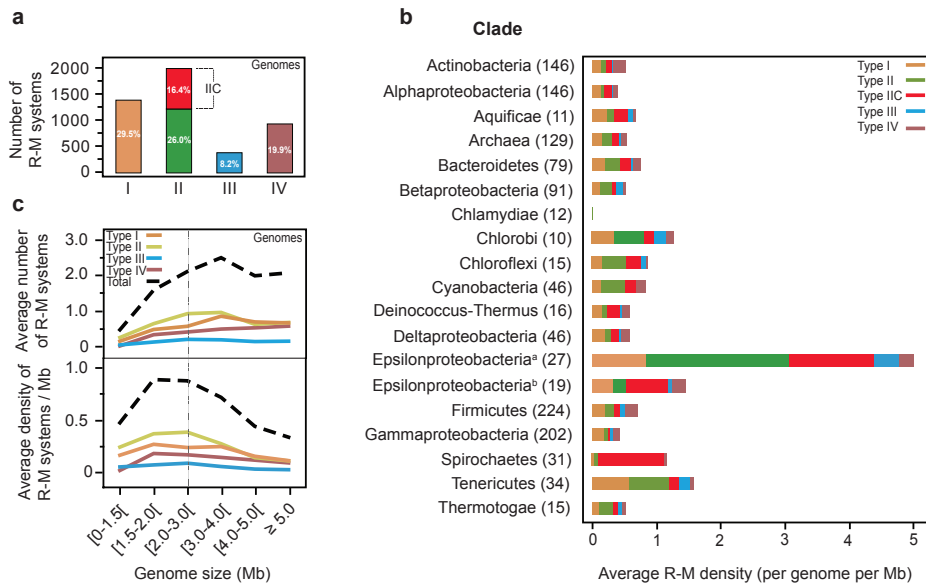


Figure 3.8: Quantification and distribution of R-M systems in 2,261 prokaryotic genomes. (a) Amount of Types I, II, IIC, III R-M systems and Type IV REases found in genomes. Corresponding percentages are indicated. (b) Average R-M density (per genome per Mb) according to clade. The largest peak on R-M density observed for *Epsilonproteobacteria*^(a) results from the presence of multiple systems particularly among *Helicobacter* species. For comparison, I also show the density for *Epsilonproteobacteria* without *Helicobacter*^(b). Only clades with at least 10 different species were considered for comparison. The number of species within each clade is indicated next to its name. (c) Distribution of the average number of R-M systems per genome (upper graph) and average density per genome per Mb (bottom graph) according to genome size (Mb). Stippled line separates the regions having small and large genomes. Genomes of *Helicobacter* were not included to avoid obtaining extremely inflated values in the [1.5–2.2] Mb genome size range. Taken from [Oliveira *et al.* 2014b].

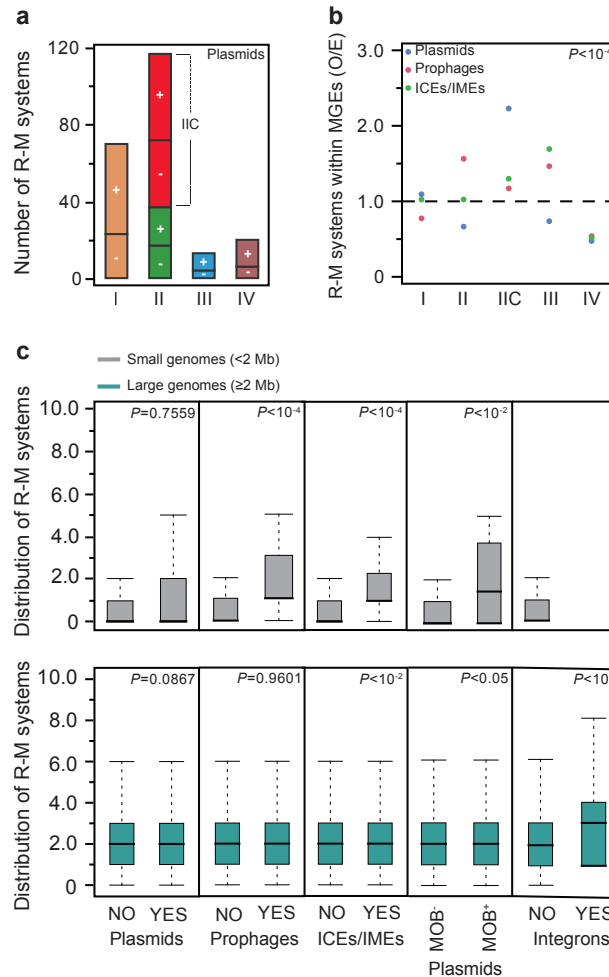


Figure 3.9: Quantification and distribution of R-M systems in MGEs. **(a)** Amount of Types I, II, IIC, III R-M systems and Type IV REases found in plasmids. The latter were classified according to their transmissibility: plasmids encoding the entire conjugation machinery or at least the relaxase (MOB⁺, shown as +), and plasmids lacking even the relaxase (MOB⁻, shown as -). **(b)** Observed/expected (O/E) ratios of R-M systems in plasmids, prophages and ICEs/IMEs. Expected values were obtained by multiplying the total number of each type of R-M system by the fraction of R-M systems assigned to each MGE. **(c)** Co-occurrence of R-M systems and MGEs. Box plots of the genomic co-occurrence of R-M systems with plasmids, prophages, ICEs/IMEs and integrans in small (<2 Mb) and large (≥2 Mb) genomes. Error bars represent standard deviations. Mann–Whitney–Wilcoxon test *P* values are indicated. Taken from [Oliveira *et al.* 2014b].

in plasmids.

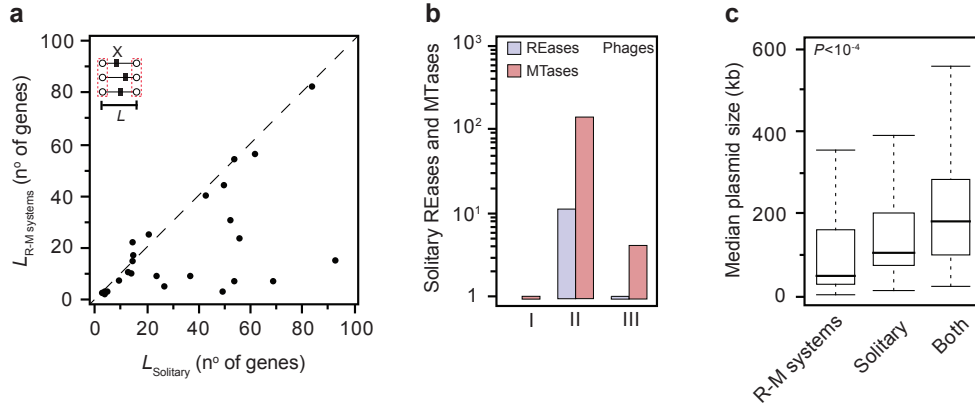


Figure 3.10: Comparative analysis of complete R-M systems and solitary components. (a) Median size of regions (L , expressed in the number of genes) harboring complete volatile R-M systems versus solitary volatile R-M elements (indicated as X). Stippled line corresponds to the identity. (b) The number of solitary REases and MTases in phages. Over 90% of the total hits were found to correspond to solitary MTases. (c) Median plasmid size (kb) for plasmids containing only complete R-M systems, solitary components or both. Mann–Whitney–Wilcoxon test P value is indicated next to the box plots. Taken from [Oliveira *et al.* 2014b].

Since MGEs appear to carry few R-M systems, I decided to quantify the association between the number of R-M systems and the presence/absence of the different MGEs in genomes. Large genomes (≥ 2 Mb) show no strong association between the number of R-M systems and the presence or absence of plasmids (independently of being MOB^- or MOB^+), prophages and ICEs/IMEs (Fig. 3.9c). Only integrons are more likely to be found in large genomes with more R-M systems. This latter observation is in agreement with previous works suggesting that R-M systems stabilize super-integrans [Rowe-Magnus *et al.* 2003]. Among small genomes (< 2 Mb), the number of R-M systems is positively correlated with the presence of prophages, ICEs/IMEs and MOB^+ plasmids (Fig. 3.9c). Overall, these data suggest that the abundance of R-M systems is indeed associated with genome size and the presence of MGEs because they are both associated with higher rates of horizontal transfer.

Previous works have found many solitary MTases and some solitary REases in genomes, suggesting they result from the genetic degradation of intact systems. Partial loss of R-M systems would lead to solitary MTases or REases that could eventually become domesticated by the host genome [Seshasayee *et al.* 2012, Ershova *et al.* 2012]. If solitary R-M genes were derived from genetic degradation of R-M systems arising by horizontal transfer then they should be encoded in MGEs ongoing genetic degradation. Hence, MGEs carrying solitary systems should be smaller than those encoding complete R-M systems. To test this hypothesis, I com-

puted the distance between the two flanking core genes surrounding solitary R-M genes and complete R-M systems. I found that for $\sim 80\%$ of the species containing non-persistent R-M proteins, this distance was smaller for complete systems than for solitary proteins ($P < 10^{-2}$; Binomial test) (**Fig. 3.10a**). Accordingly, solitary elements are very abundant in large MGEs such as phages and conjugative elements, whereas we showed above that complete systems are rare in these elements. In phages I identified 155 solitary genes (**Fig. 3.10b**), even though phages encode very few R-M systems. Plasmids containing only solitary R-M proteins are larger than those carrying complete R-M systems (median sizes of 106 and 50 kb, respectively, $P < 10^{-4}$, Mann–Whitney–Wilcoxon test) (**Fig. 3.10c**). These results strongly suggest that R-M systems and solitary proteins are transferred independently through distinct MGEs and/or transfer mechanisms. Additionally, this implies that solitary components of R-M systems are not systematically part of an ongoing genetic degradation process. Instead, it suggests solitary components are acquired as such by bacterial chromosomes.

3.3.3 Regulation of genetic flux between bacteria by restriction–modification systems

In bacterial population genetics, the events of gene transfer are usually termed HGT when they result in the acquisition of new genes, and HR when they result in allelic replacements. Several recent large-scale studies of population genomics have observed more frequent HR within than between lineages [Didelot *et al.* 2011, Doroghazi & Buckley 2010]. This suggests that HR might favor the generation of cohesive population structures within bacterial species [Fraser *et al.* 2007]. Specific lineages of important pathogens that have recently changed their R-M repertoires show higher sexual isolation, such as *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Burkholderia pseudomallei*, and *Staphylococcus aureus* [Budroni *et al.* 2011, Croucher *et al.* 2014, Nandi *et al.* 2015]. For example, a Type I R-M system decreased transfer to and from a major methicillin-resistant *S. aureus* lineage [Roberts *et al.* 2013]. Diversification of R-M target recognition sites could thus reduce transfer between lineages with different systems while establishing preferential gene fluxes between those with R-M systems recognizing the same target motifs (cognate R-M). However, these results can be confounded by evolutionary distance: closely related genomes are more likely to encode similar R-M systems, inhabit the same environments (facilitating transfer between cells), and have similar sequences (that recombine at higher rates). The advantages conferred by new genes might be higher when transfer takes place between more similar genetic backgrounds.

Here, I aimed at testing the effect of R-M systems on the genetic flux in bacterial populations. I concentrated on Type II R-M systems because they are the best studied, very frequent, and those for which I could predict sequence specificity. I inferred genome-wide counts of HR and HGT and tested their association with the frequency of R-M systems encoded in the genomes. I then made a more precise test

of the key hypothesis that bacteria carrying similar R-M systems establish highways of gene transfer, independently of phylogenetic proximity and clade-specific traits.

I analyzed a dataset of 79 core genomes and pangenomes corresponding to a total of 884 complete genomes. I used five different programs to detect HR in the core genome and Count [Csuros 2010] to infer the events of HGT from the patterns of presence and absence of gene families in the species' trees. I identified 236,894 events of gene transfer in the 79 pangenomes. These events were very unevenly distributed among clades, from close to none in the genomes of obligatory endosymbionts to 1,538 events per genome in *Rhodopseudomonas palustris* (Fig. 3.11a). Genome size had a strong direct effect on HR and HGT (Fig. 3.11b).

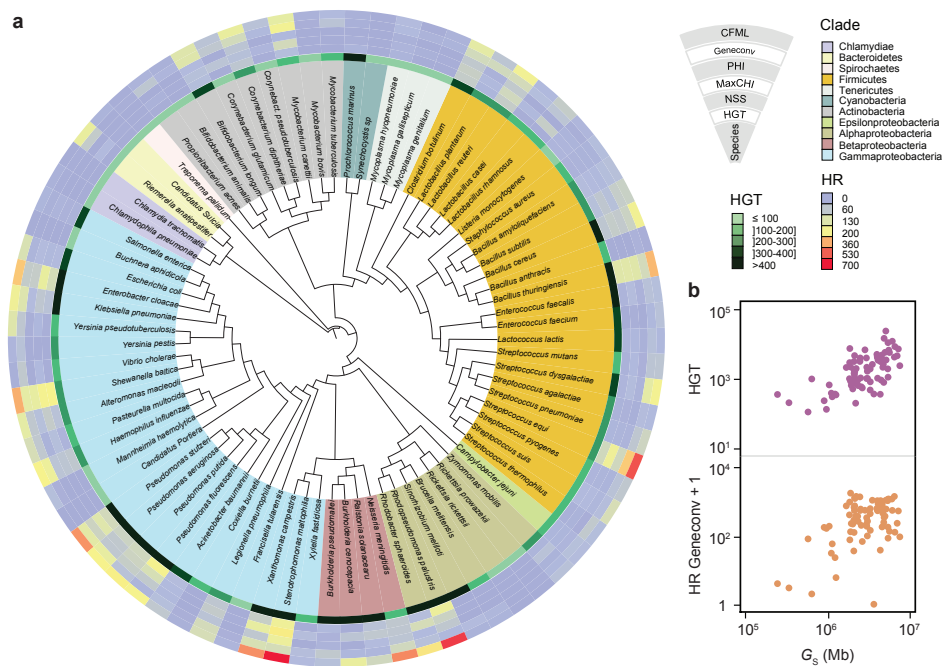


Figure 3.11: Analysis of HR and HGT events. in bacteria (a) 16S rRNA phylogenetic tree of the 79 bacterial species. The tree was drawn using the iTOL server (itol.embl.de/index.shtml). The innermost circle layer indicates the species and associated clade. The six subsequent layers correspond (in an outwardly direction) to the average number of HGT events per genome computed using Count; the number of recombined genes per genome given by NSS, MaxChi, and PHI; and the number of recombination events per genome given by Geneconv and CFML (outermost layer), respectively. (b) Distribution of the average number of HGT events and HR events (inferred by Geneconv) per clade according to genome size (G_s). Spearman's $\rho_{HGT} = 0.65$, $P_{HGT} < 10^{-4}$, Spearman's $\rho_{Geneconv} = 0.32$, $P_{Geneconv} < 10^{-2}$. Taken from [Oliveira *et al.* 2016].

I further found that the number of HGT and HR events was higher in genomes with more R-M systems, and especially in those with Type II systems

[Oliveira *et al.* 2016]. The observation of higher genetic fluxes in the presence of R-M systems might seem unexpected in the light of the role of the latter in degrading exogenous DNA. To explain these results, I put forward two hypotheses:

Hypothesis 1: The relative abundance of R-M systems in a clade results from the selective pressure imposed by the abundance of MGEs in that clade. Selection for multiple R-M systems is expected to be stronger for clades enduring infections by many MGEs. R-M systems have limited efficiency and might not completely prevent MGE infection and transfer [Korona *et al.* 1993]. This results in a weak positive association between transfer of genetic information and the abundance of R-M systems.

Hypothesis 2: R-M systems favor transfer of genetic material between cells by generating restriction breaks that stimulate recombination between homologous sequences.

To distinguish between the first two hypotheses, I analyzed the genetic flux between pairs of genomes with cognate Type II R-M systems. If R-M systems predominantly prevent genetic transfer (hypothesis 1), then the flux of genetic material between genomes encoding cognate R-M systems should be higher. If R-M systems predominantly stimulate genetic transfer (hypothesis 2), then pairs of genomes encoding cognate R-M systems should show lower than average genetic flux. I found that lineages associated with genomes encoding cognate R-M systems co-exchanged more genetic information (**Fig. 3.12a-c**). To verify that the presence of cognate R-M systems is associated with increased genetic exchange independently of evolutionary distance, I binned the comparisons between events occurring in terminal branches in terms of the phylogenetic distance between pairs of genomes. I then ran the same analysis in each bin separately. These analyses showed more co-transfer between genomes encoding cognate R-M systems in nearly all bins, even if this analysis had lower statistical power (fewer comparisons per bin) (**Fig. 3.12b,c**). Importantly, this difference was always significant for the most distant pairs of genomes. Hence, pairs of genomes encoding cognate R-M systems were associated with more frequent HR and HGT, independently of the evolutionary distances between them.

This work shows that noncognate genomes have reduced DNA exchanges. This decreases the power of natural selection and increases the effect of drift, potentially leading to the accumulation of deleterious mutations. Importantly, R-M systems' diversification at the origin of a lineage may increase its genetic cohesion by disfavoring exchanges with the closest related ones, as previously suggested for some pathogens [Budroni *et al.* 2011, Croucher *et al.* 2014, Nandi *et al.* 2015]. Interestingly, diversification can also increase the genetic flux between distant bacteria encoding cognate R-M systems with which there were previously few genetic exchanges. Hence, R-M systems might shape population structure in complex ways depending on the repertoire of R-M systems in the other lineages.

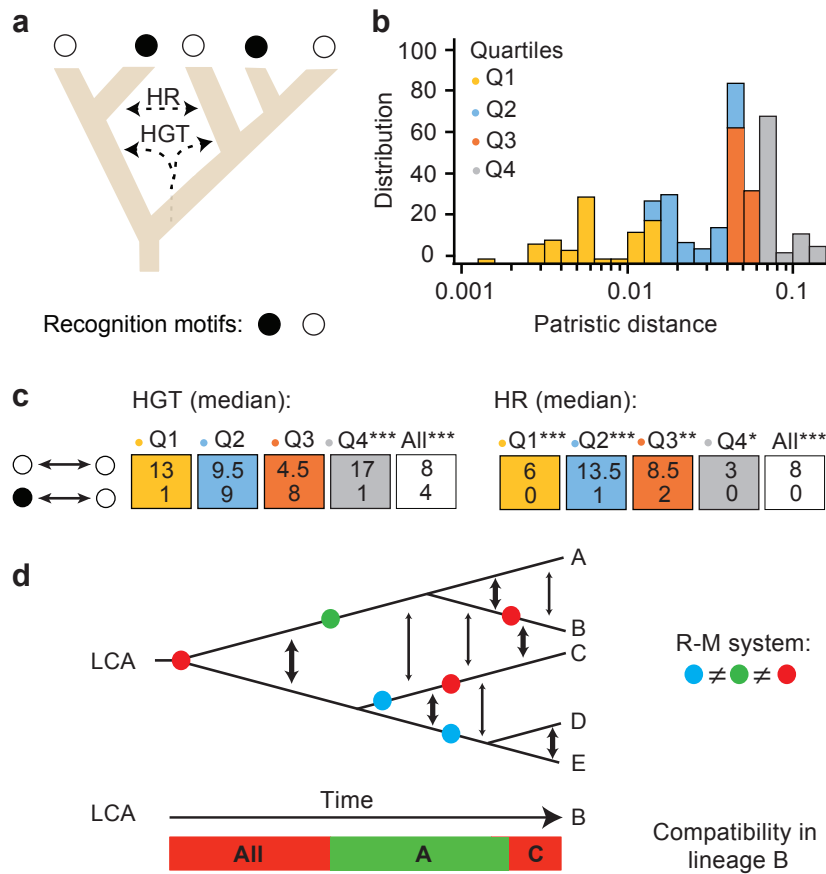


Figure 3.12: Gene flux in bacteria encoding R-M systems. (a) I analyzed the patterns of HR and HGT in the tree of each clade, comparing the flux between tips ending in cognate (similar recognition motifs) or noncognate (different motifs) extant taxa. (b) Histogram of patristic distances (colored by quartiles) between bacteria with Type II R-M systems. (c) Median values of HGT and recombination events for each quartile (Q) and for the full dataset (All) between terminal branches of bacteria with Type II R-M systems recognizing (or not) the same target motif. We analyzed *Bacillus amyloliquefaciens*, *Bifidobacterium longum*, *E. coli*, *Haemophilus influenza*, *Listeria monocytogenes*, *Neisseria meningitidis*, *Salmonella enterica*, and *Streptococcus pneumoniae*. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$. (d) Genetic flux in function of time and the presence of R-M systems. As lineages diverge and R-M systems change (circles indicate such changes), the lineages with cognate R-M systems (same color) share more genetic material than the other lineages. For example, the lineage B changes R-M systems twice since the last common ancestor (LCA). Initially transfer is favored with all lineages, then with the sister lineage A, and finally with the distantly related lineage C. Taken from [Oliveira *et al.* 2016]

3.3.4 The chromosomal organization of horizontal gene transfer in bacteria

Gene repertoires of bacterial species are often very diverse, which is central to bacterial adaptation to changing environments, new ecological niches, and co-evolving eukaryotic hosts [Wilmes *et al.* 2009]. In this regard, genomic diversification is shaped by the balancing processes of gene acquisition and loss [Mira *et al.* 2001], moderated by positive selection on some genes, and purifying selection on many others [Koskiniemi *et al.* 2012]. Chromosomes are organized to favor the interactions of DNA with the cellular machinery [Rocha 2008]. For example, most bacterial genes are co-transcribed in operons, leading to strong and highly conserved genetic linkage between neighboring genes [Overbeek *et al.* 1999]. At a more global level, early-replicating regions are enriched in highly expressed genes in fast-growing bacteria to enjoy replication-associated gene dosage, creating a negative gradient of expression along the axis from the origin (*ori*) to the terminus (*ter*) of replication (*ori*->*ter*) [Vieira-Silva & Rocha 2010, Sharp *et al.* 1989]. These organizational traits can be disrupted by the integration of novel genetic information. So there is an equilibrium between selection for organization, and selection for diversification.

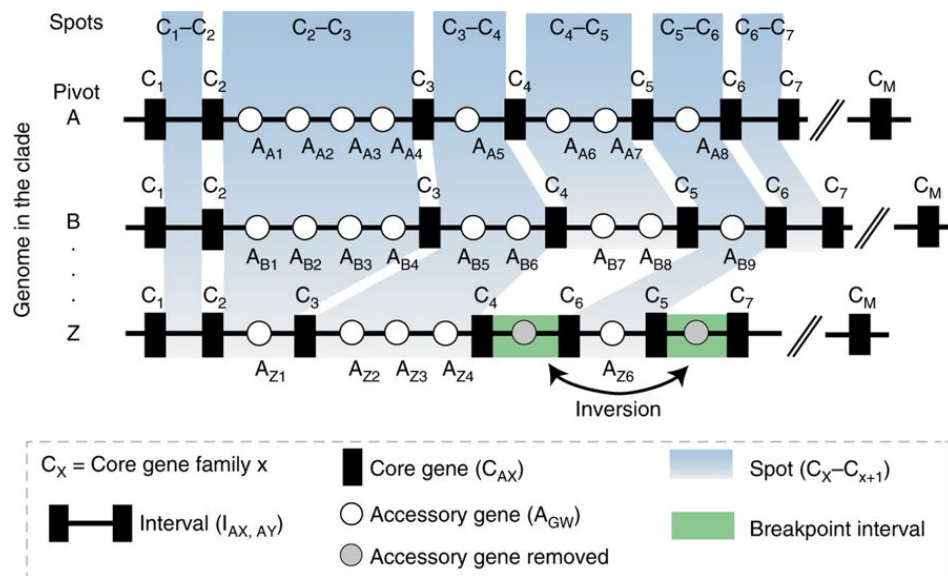


Figure 3.13: Scheme depicting key concepts used in this study. Intervals flanked by the same core gene families (C_X , C_Y) as those from pivot genome A were defined as syntenic intervals (i.e., the members of the core gene families X and Y were also contiguous in the pivot). The intervals that do not satisfy this constraint were classed as breakpoint intervals (green-shaded regions) and excluded from our analysis. For every interval in the pivot genome, we defined spot as the set of intervals flanked by members of the same core gene families (blue-shaded regions). Taken from [Oliveira *et al.* 2017].

MGEs carrying similar integrases tend to integrate at the same sites in the chromosome, leading to regions with unexpectedly high frequency of MGEs at homologous regions. This concentration of MGEs in few sites has been frequently described [Balbontin *et al.* 2008, Boyd *et al.* 2009], especially in relation to the presence of neighboring tRNA and tmRNA genes [Williams 2002]. Yet, a previous work described the existence of regions with high rates of diversification in *E. coli* (hotspots), some of which lacked recognizable integrases [Touchon *et al.* 2009]. In particular, the genes flanking two hotspots were associated with high rates of homologous recombination (*rfb* and *leuX*). In *Streptococcus pneumoniae*, the chromosomal genes flanking MGEs also showed higher rates of HR [Croucher *et al.* 2011, Chancey *et al.* 2015]. In this species, it was suggested that integration of MGEs close to core genes under selection for diversification could be adaptive by facilitating the transfer and subsequent recombination of the latter [Everitt *et al.* 2014]. Here, I defined and identify hotspots in a large and diverse panel of bacterial species and show how they reflect the mechanisms driving genome diversification by HGT.

To study the distribution of gene families in bacterial chromosomes, I analyzed 932 complete genomes of 80 bacterial species. I inferred the core genome, the pan-genome, the accessory genome, and the phylogeny of each species. I partitioned the genomes into an array of core genes and intervals (**Fig. 3.13**). The latter were defined as the positions between consecutive core genes. I defined a spot as the set of intervals delimited by members of the same two families of core genes in the genomes of the clade. Spots contained 170,041 HTgenes (Horizontally Transferred Genes). I next quantified the clustering of these genes by counting the minimal number of spots required to accumulate at least 50% of the HTgenes (HTg50). I found that < 2% of the largest hotspots accumulate >50 % of all HTgenes [Oliveira *et al.* 2017]. Conversely, 72.6% of the spots were on average empty, *i.e.*, had no accessory gene in any genome. Hence, most HTgenes were found integrated in a very small number of sites in the genome.

I next used simulations to infer the statistical thresholds for the degree of clustering of HTgenes in each clade. I identified the spot with the highest number of HTgenes in each simulation (MaxHTg,i), and computed the 95th percentile of the distribution of these maximal values (T_{95%}). Spots with more than T_{95%} HTgenes were called hotspots, spots lacking accessory genes were called empty, and the others were called coldspots. I found a total of 1841 hotspots in the 80 clades (**Fig. 3.14a**). They represent only 1.2% of the spots, but they concentrate 47% of the accessory gene families and 60% of the HTgenes. The number of hotspots differed widely among clades, from none or very few in *Acetobacter pasteurianus*, *Bacillus anthracis*, and the obligatory endosymbionts, to more than 60 in *Bacillus thuringiensis*, *E. coli*, and *Pseudomonas putida* (**Fig. 3.14a**). This variance was partly a function of chromosome size (**Fig. 3.14b**), but was especially associated with the number of HTgenes (**Fig. 3.14c**). Hence, a few hotspots aggregate most of the genes acquired by horizontal transfer and this trend is more pronounced when the rates of transfer are high.

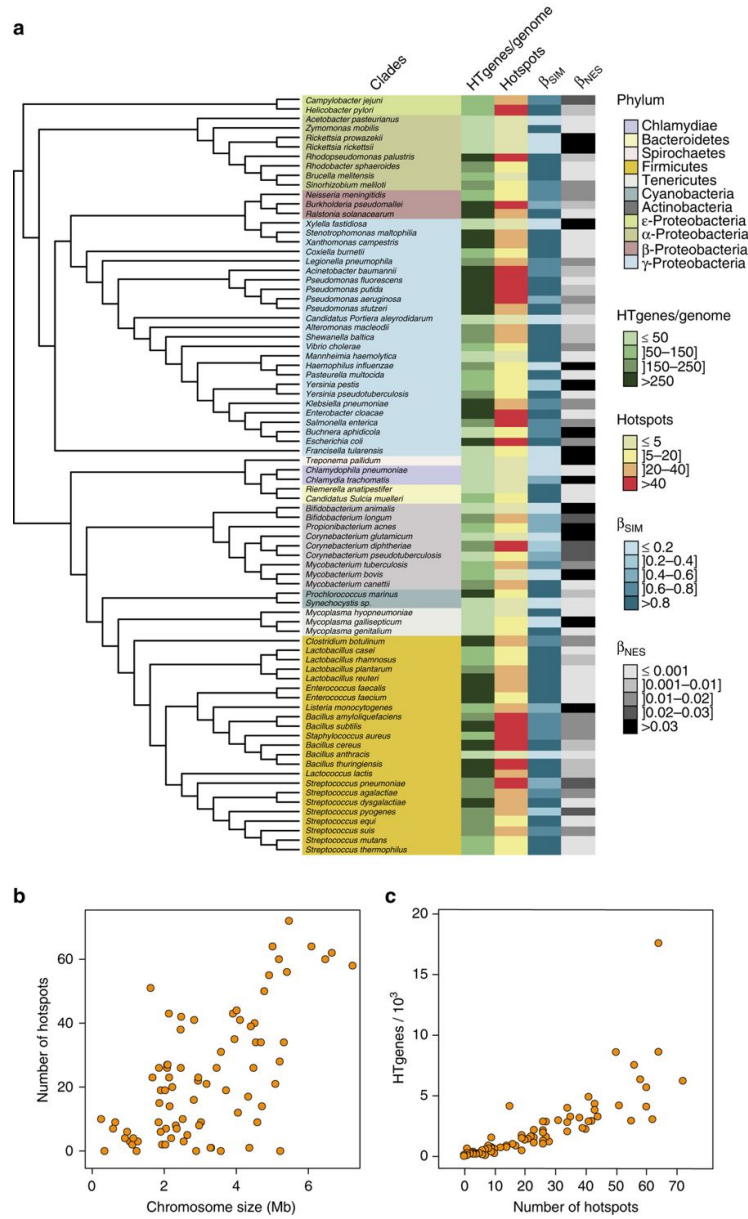


Figure 3.14: Analysis of HTgenes and the abundance and distribution of hotspots. (a) 16S rRNA phylogenetic tree of the 80 bacterial clades. The tree was drawn using the iTOL server (itol.embl.de/index.shtml). The first column indicates the clade and is colored by phylum. The four subsequent columns correspond respectively to: the average number of HTgenes per genome computed using Count, the number of hotspots, the average Simpson dissimilarity index (β_{SIM} , accounting for turnover), and the average multiple-site dissimilarity index accounting only for nestedness (β_{NES}). (b) Distribution of the average number of hotspots per clade according to the average genome size (G_S). (c) Association between the number of hotspots and the number of HTgenes in the clade. Taken from [Oliveira *et al.* 2017].

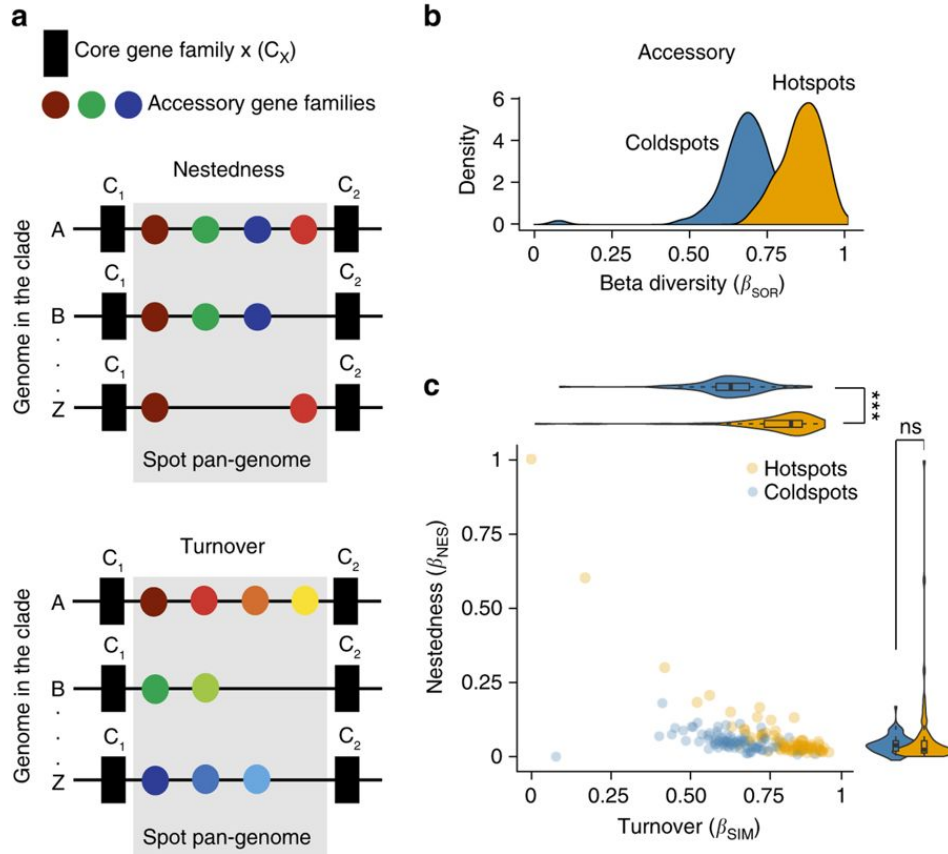


Figure 3.15: Genetic diversity of the accessory genes present in hotspots and coldspots. **(a)** Examples of gene nestedness and turnover in a spot. Turnover measures the segregation between intervals in terms of gene families, *i.e.*, it accounts for the replacement of some genes by others. Nestedness accounts for differential gene loss and measures how the gene repertoires of some intervals are a subset of the repertoires of the others. It typically reflects a non-random process of gene loss. **(b)** Distributions of β diversity (β_{SOR}) in hotspots and coldspots. **(c)** Partition of β_{SOR} in its components of nestedness (β_{NES}) and turnover (β_{SIM}) for hotspots and coldspots ($\beta_{SOR} = \beta_{NES} + \beta_{SIM}$). *** $P < 10^{-3}$; Mann-Whitney-Wilcoxon test; ns: not significant. Taken from [Oliveira *et al.* 2017].

To assess whether genetic diversity in hotspots was compatible with one single ancient integration event, I introduced measures derived from the analysis of beta diversity in Ecology, where it is used to measure the differences in species composition between different locations. Here I used it to measure the difference in gene repertoires among a set of intervals from the same spot. I measured the Sørensen index (β_{SOR}) for hotspots and coldspots of each species using the binary matrix of gene presence/absence. Diversity results from a mixture of independent gene acquisitions and replacements (turnover) and differential gene loss (nestedness), and β_{SOR} can be partitioned into the two related additive terms: turnover (β_{SIM}) and nestedness (β_{NES}) ($\beta_{\text{SOR}} = \beta_{\text{NES}} + \beta_{\text{SIM}}$) (**Fig. 3.15a**). Beta diversity of accessory genes was higher in hotspots than in coldspots (**Fig. 3.15b**). This difference was caused by turnover, since only β_{SIM} was significantly higher in hotspots than in coldspots (**Fig. 3.15c**). The values of β_{NES} were very low in both cases; confirming that most hotspots are not caused by singular events of integration of MGEs. Hence, while genetic diversity is high in hotspots and coldspots, these results show faster diversification in hotspots because they endure higher genetic turnover.

3.3.5 Transition to a Senior Scientist position

During my previous postdoctoral project I was able to acquire multiple programming skills as well as competences in comparative genomics and biology of DNA methylation systems. At this point I sought to become more autonomous and implement a line of research in which I could explore multiple but complementary angles of bacterial methylomics. In particular, I moved to Mount Sinai School of Medicine in New York, to leverage Pacbio Single-Molecule Real-Time sequencing (SMRT-seq) data, and perform the first comparative epigenomics study across a diverse collection of isolates of the human pathogen *C. difficile*.

3.4 Senior Scientist (2016-2020)

3.4.1 Introduction

C. difficile is a spore-forming Gram-positive obligate anaerobe and the leading cause of nosocomial antibiotic-associated disease in the developed world [Smits *et al.* 2016]. Clinical symptoms of *C. difficile* infection (CDI) in humans range in severity from mild self-limiting diarrhea to severe, life-threatening inflammatory conditions, such as pseudomembranous colitis or toxic megacolon. Since the vegetative form of *C. difficile* cannot survive in the presence of oxygen, CDIs are transmitted via the fecal/oral route through *C. difficile*'s metabolically dormant spore form; these spores subsequently germinate into actively growing, toxin-producing vegetative cells that are responsible for disease pathology [Paredes-Sabja *et al.* 2014]. CDI progresses in an environment of host microbiota dysbiosis, which disrupts the colonization resistance typically provided by a diverse microbiota and may enhance spore germination [Seekatz & Young 2014]. In the

last two decades, there has been a dramatic rise in outbreaks with increased mortality and morbidity due in part to the emergence of epidemic-associated strains with enhanced growth [Zidaric & Rupnik 2016, Collins *et al.* 2018], toxin production [Lanis *et al.* 2010], and antibiotic resistance [Valiente *et al.* 2014]. *C. difficile* was responsible for half a million infections in the United States in 2011, with 29,000 individuals dying within 30 days of the initial diagnosis [Lessa *et al.* 2015]. Those most at risk are older adults, particularly those who take antibiotics that perturb the normally protective intestinal microbiota.

Despite the significant progress achieved in the understanding of *C. difficile* physiology, genetics, and genomic evolution [Sebaihia *et al.* 2006, He *et al.* 2013], the roles played by epigenetic factors, namely DNA methylation, have not been studied. In the bacterial kingdom, there are three major forms of DNA methylation: N6-methyladenine (6mA, the most prevalent form representing roughly 80%), N4-methylcytosine (4mC), and 5-methylcytosine (5mC). Although bacterial DNA methylation is most commonly associated with R-M systems that defend hosts against invading foreign DNA [Oliveira *et al.* 2014b, Oliveira *et al.* 2016], increasing evidence suggests that DNA methylation also regulates a number of biological processes such as DNA replication and repair, cell cycle, chromosome segregation and gene expression, among others [Casadesus & Low 2006, Low *et al.* 2001, Cohen *et al.* 2016, Manso *et al.* 2014, Atack *et al.* 2015, Wion & Casadesus 2006]. Efficient high-resolution mapping of bacterial DNA methylation events has only recently become possible with the advent of SMRT-seq [Flusberg *et al.* 2010], which can detect all three types of DNA methylation, albeit at different signal-to-noise ratios: high for 6mA, medium for 4mC, and low for 5mC. This technique enabled to characterize the first bacterial methylomes [Fang *et al.* 2012, Murray *et al.* 2012], and since then, more than 4,500 (as of 12/2021) have been mapped, heralding a new era of "bacterial epigenomics" [Davis *et al.* 2013].

Herein, I mapped and characterized the DNA methylomes of 36 human *C. difficile* isolates using SMRT-seq and comparative epigenomics. I observed great epigenomic diversity across *C. difficile* isolates, as well as the presence of a highly conserved MTase. Inactivation of this MTase resulted in a functional impact on sporulation, a key step in *C. difficile* transmission, consistently supported by multi-omics data and genetic experiments. Further integrative transcriptomic analysis suggested that epigenetic regulation by DNA methylation is also associated with *C. difficile* host colonization and biofilm formation. Finally, the epigenomic landscape of *C. difficile* also allowed me to perform a data-driven joint analysis of multiple defense systems and their contribution to gene flux. These discoveries are expected to stimulate future investigations along a new epigenetic dimension to characterize, and potentially repress medically relevant biological processes in this critical pathogen.

3.4.2 Methylome analysis reveals great epigenomic diversity in *C. difficile*

From a Pathogen Surveillance Project at Mount Sinai Medical Center, 36 *C. difficile* isolates were collected from fecal samples of infected patients. A total of 15 different MLST sequence types (STs) belonging to clades 1 (human and animal, HA1) and 2 (so-called hypervirulent or epidemic) are represented in our dataset (**Fig. 3.16a**).

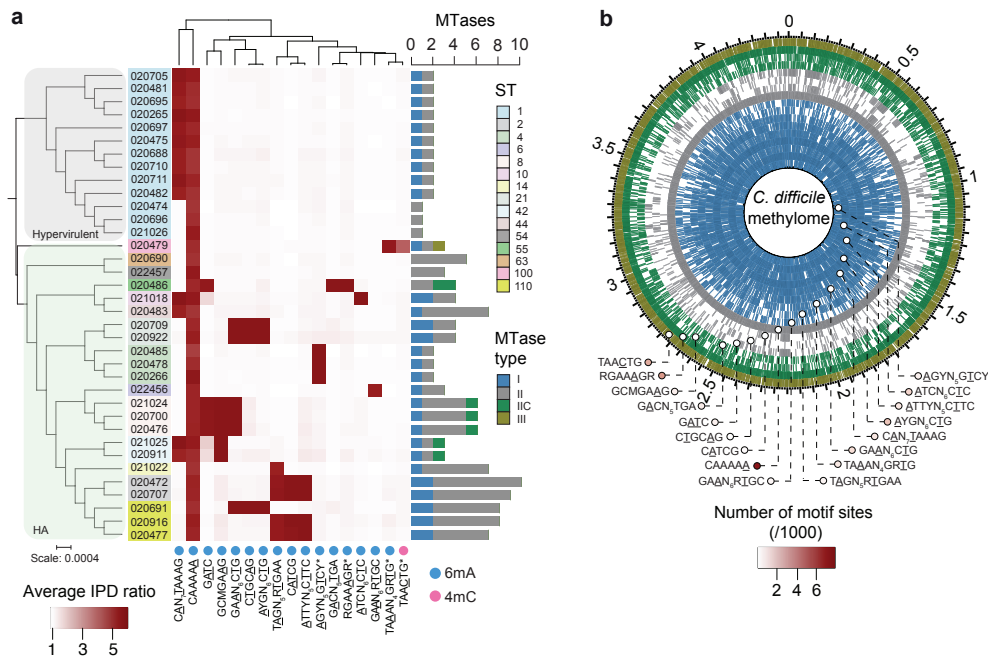


Figure 3.16: Methyomes of 36 *C. difficile* strains. (a) Phylogenetic tree of 36 *C. difficile* strains colored by clade (hypervirulent, human and animal (HA) associated) and MLST sequence type (ST). Heatmap depicting the landscape of methylated motifs per genome, and their average interpulse duration (IPD) ratio. Asterisks refer to new motifs not previously listed in the reference database REBASE. Methylated bases are underlined. The CAAAAA motif was consistently methylated across isolates. Barplot indicates the number and types of active MTases detected per genome. In Type IIC systems, MTase and REase are encoded in the same polypeptide. (b) Representation of the *C. difficile* methylome. Shown are the positions of all methylation motif sites in the reference genome of *C. difficile* 630, colored according to MTase type. Also shown are the average motif occurrences per genome (across the 36 isolates). Taken from [Oliveira *et al.* 2020].

Methylation motifs were found using the SMRT-portal protocol. I found a total of 17 unique high-quality methylation motifs in the 36 genomes (average of 2.6 motifs per genome) (**Fig. 3.16a**). The large majority of target motifs were of 6mA type, one motif (TAACTG) belonged to the 4mC type, and no confident 5mC motifs were

detected. Like most bacterial methylomes, >95% of the 6mA and 4mC motif sites were methylated (**Fig. 3.16b**). One 6mA motif, CAAAAA, was intriguingly present across all genomes, which led us to raise the hypothesis that 6mA methylation events at this motif, and its corresponding MTase, may play an important and conserved functional role in *C. difficile*. Consistent with the presence of a highly conserved CAAAAA motif, I identified a Type II 6mA solitary DNA MTase (577 aa) present across isolates. This MTase is encoded by *CD2758* in *C. difficile* 630, a reference strain that was isolated from a *C. difficile* outbreak in Switzerland. The ubiquity of this MTase was not restricted to the 36 isolates, as I was able to retrieve orthologs in a list of approximately 300 global *C. difficile* isolates from GenBank.

3.4.3 Comparative analysis of methylation sites across *C. difficile* genomes

The *C. difficile* genome has an average of 7,721 (± 197 , sd) CAAAAA motif sites. Adjusted by the k-mer frequency of the AT-rich *C. difficile* genome (70.9%) using Markov models, CAAAAA motif sites are significantly under-represented in the chromosome, particularly in intragenic regions. To evaluate if specific chromosomal regions are enriched or depleted for this motif, I used a multi-scale signal representation (MSR) approach (**Fig. 3.17a**).

Briefly, MSR uses wavelet transformation to examine the chromosome at a succession of increasing length scales by testing for enrichment or depletion of a given genomic signal. While scale values <10 are typically associated with regions <100 bp, genomics regions enriched for CAAAAA sites at scale values >20 correspond to segments larger than 1 kb (i.e. gene and operon scale), and include genes related to sporulation and colonization (**Fig. 3.17a**): stage 0 (*spo0A*), stage III (*spoIIIAA-AH*), and stage IV (*spoIVB*, *sigK*) of sporulation, membrane transport (PTS and ABC-type transport systems), transcriptional regulation (e.g. *iscR*, *fur*), and multiple cell wall proteins (CWPs).

To further characterize CAAAAA motif sites, I built on the large collection of methylomes of the same species and sought to categorize these motif sites on the basis of their positional conservation across genomes. I performed whole genome alignment of 37 *C. difficile* genomes (36 isolates + *C. difficile* 630 as reference), and classified each motif position in the alignment as either: (i) conserved orthologous (devoid of SNPs or indels); (ii) variable orthologous (in which at least one genome contains a SNP or indel); and (iii) non-orthologous (**Fig. 3.17b**). I found a total of 5,828 conserved orthologous motif positions, 1,050 variable orthologous positions (885 with SNPs and 165 with indels), and an average of 843 non-orthologous positions per genome. The latter were, as expected, largely mapped to MGEs.

3.4.4 Non-methylated sites are enriched in regulatory elements

DNA methylation is highly motif driven in bacteria; i.e. in most cases, >95% of the occurrence of a methylation motif is methylated. However, a small fraction of

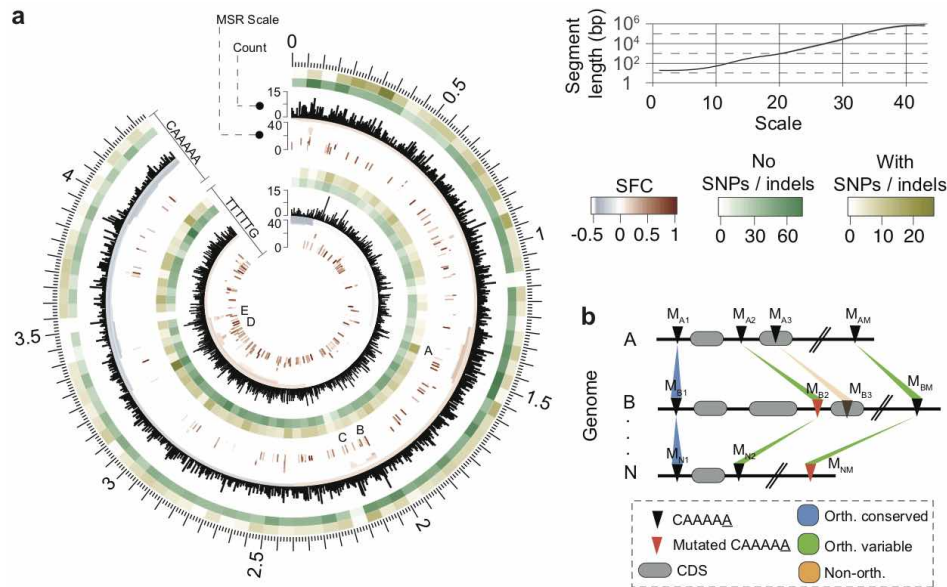


Figure 3.17: Abundance, distribution, and conservation of CAAAAA motifs. (a) Distribution of CAAAAA sites in both strands of the reference *C. difficile* 630 genome, and corresponding genomic signal obtained by MSR. Letters (A-E) represent regions with particularly high abundance of CAAAAA at scales above 20, i.e., typically above the single gene level. Relation between MSR scale and segment length is also shown. The significant fold change (SFC) corresponds to the fold change (\log_2 ratio) between observed and randomly expected overlap statistically significant at $P=10^{-6}$. Heatmap layers correspond to the number of orthologous conserved (no SNPs/indels, green) and orthologous variable (with SNPs/indels) CAAAAA positions. (b) Whole genome alignment (36 isolates + *C. difficile* 630 as reference) was performed using Mauve. I defined an orthologous occurrence of the CAAAAA motif (black triangles), if an exact match to the motif was present in each of the 37 genomes (conserved, blue regions), or if at least one motif (and a maximum of $n-1$, being n the number of genomes) contained positional polymorphisms (maximum of two SNPs or indels per motif) (variable, green regions). Non-orthologous occurrences of CAAAAA are indicated as orange-shaded regions. The results are shown in (Fig. 3.17a) in the form of heatmaps. Taken from [Oliveira *et al.* 2020].

methylation motif sites can be non-methylated. The on/off switch of DNA methylation in a bacterial cell can contribute to epigenetic regulation through competitive binding between DNA MTases and other DNA binding proteins (e.g. Transcription Factors (TFs)) as previously described for *E. coli* [Wion & Casadesus 2006, Lim & van Oudenaarden 2007, Ardissonne *et al.* 2016, Cota *et al.* 2016]. Previous bacterial methylome studies analyzing one or few genomes usually had insufficient statistical power to perform a systematic interrogation of non-methylated motifs sites. Building on the rich collection of 36 *C. difficile* methylomes, I performed a systematic detection and analysis of non-methylated CAAAAA sites. The latter were found dispersed throughout the full length of the *C. difficile* genome, yet were overrepresented in orthologous variable and non-orthologous CAAAAA positions (O/E=respectively 1.51 and 1.49) and underrepresented in orthologous conserved CAAAAA positions (O/E 0.84) ($P < 10^{-4}$; Chi-square test). This is consistent with the idea that variable positions are more likely to be non-methylated to provide breadth of expression variation. A minor percentage of positions (5.5%) remained non-methylated in at least one third of the isolates, suggesting that competitive protein binding is expected to be more active in certain genomic regions (**Fig. 3.18a**).

The non-methylated CAAAAA positions detected across the 36 *C. difficile* genomes allowed a systematic search for evidence of overlap with Transcription Factor Binding Sites (TFBSs) and Transcription Start Sites (TSSs). To test this, I queried the genomes for putative binding sites of 21 TFs pertaining to 14 distinct families and found overlaps between prominent peaks of non-methylated CAAAAA (**Fig. 3.18a**) and the TFBSs of CodY and XylR (**Fig. 3.18b**). Performing the analysis at the genome level, both CodY and XylR binding sites showed significant enrichment for non-methylated CAAAAA (**Fig. 3.18c**). In a similar enrichment analysis using 2,015 TSSs reconstructed from RNA-seq data coverage, we found a genome-level enrichment: of non-methylated CAAAAA sites preferentially overlapped with TSSs (**Fig. 3.18d,e**).

In addition to the on/off epigenetic switch driven by competitive binding between the MTase and other DNA binding proteins at CAAAAA sites, we hypothesized that epigenetic heterogeneity within a clonal population could also stem from DNA replication errors at CAAAAA sites, especially because homopolymer tracts are expected to be more error-prone [Moxon *et al.* 2006]. To test this hypothesis, we investigated if CAAAAA had a larger than expected frequency of mutated reads (compared to baseline sequencing errors). Specifically, we tested if CAAAAA sites in *C. difficile* are indeed particularly error prone during DNA replication, compared to the control motifs TAAAAA or GAAAAA (here called KAAAAA).

To account for the confounding effect of mutation rate at different sequence contexts, I further computed the % mutated reads (SNPs + indels) in four distinct species with a broad dispersion of mutation rates (*Mycobacterium tuberculosis* < *E. coli* \approx *C. difficile* < *Helicobacter pylori*). Interestingly, I indeed found a significantly higher % of mutated reads mapping to CAAAAA sites compared to KAAAAA sites in *C. difficile* (**Fig. 3.18f**), while this difference was not observed in any of the control genomes.

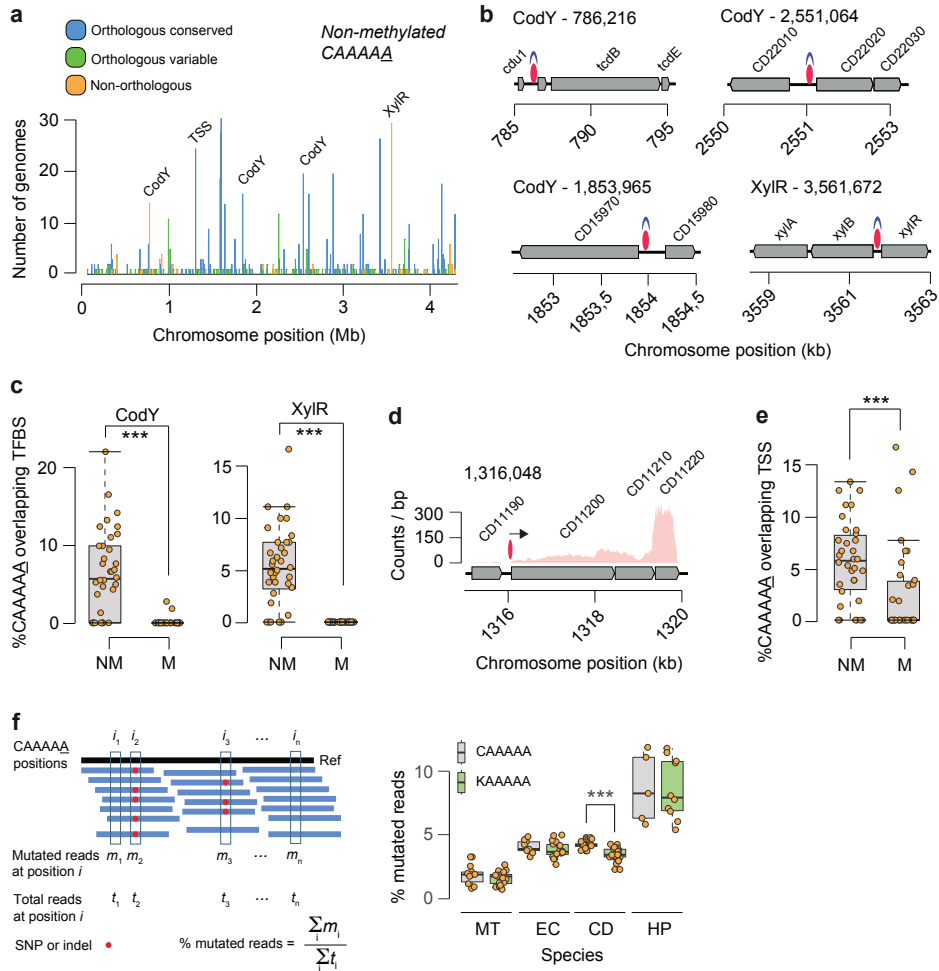


Figure 3.18: Distribution of non-methylated CAAAAA motif sites, and overlap with transcription factor binding sites (TFBS) and transcription start sites (TSS). (a) Number of *C. difficile* isolates for which non-methylated CAAAAA motif sites were detected at a given chromosome position (coordinates are relative to the reference genome of *C. difficile* 630). Peak colors correspond to orthologous (conserved and variant) and non-orthologous CAAAAA positions. Some of the major peaks of non-methylated CAAAAA positions were found to overlap with TFBS (e.g.: CodY, XylR) and TSS. (b) Genetic regions for which overlap was observed between highly conserved non-methylated CAAAAA motif sites (red ovals) and TFs (CodY and XylR, blue forms). (c) % CAAAAA motif sites (non-methylated and methylated) overlapping CodY and XylR for each *C. difficile* isolate. (d) Example of a chromosomal region in which non-methylated CAAAAA motifs overlap a TSS (shown as arrow). (e) % CAAAAA motifs (non-methylated (NM) and methylated (M)) overlapping TSSs for each *C. difficile* isolate. (f) % mutated reads (SNPs + indels) in CAAAAA and K(G or T)AAAAA motifs for *M. tuberculosis* (MT), *E. coli* (EC), *C. difficile* (CD) and *H. pylori* (HP). AAAAAA was not considered as control motif as it would theoretically be more error-prone. *** $P < 10^{-3}$, Mann-Whitney-Wilcoxon test. Taken from [Oliveira *et al.* 2020].

Collectively, these results highlight two types of variations that affect CAAAAA methylation status in *C. difficile*: on/off epigenetic switch of CAAAAA sites preferentially overlap with putative TFBSs and TSSs, and higher mutation rates at the CAAAAA motif sites that contribute to cell-to-cell heterogeneity.

3.4.5 Loss of methylation impacts transcription of sporulation genes

To study the functional significance of methylation at CAAAAA sites, I used RNA-seq to compare the transcriptomes of wild-type *C. difficile* 630 with that of the knockout mutant $\Delta CD2758$: during exponential (mid-log) growth phase, following sporulation induction (15 and 19 h), and during stationary phase. Of the 3,896 genes annotated in *C. difficile* 630, 405 – 715 (10.4 – 18.3%, depending on the growth stage) were differentially expressed (DE) at a 1% FDR, with effect sizes ranging from 13-fold for under-expressed and 8-fold for over-expressed genes. The set of DE genes was enriched for the sporulation Gene Ontology (GO) term, as well as some additional terms: flagellum, cell division, ribosome/translation, ATP-coupled transport, and de-novo UMP biosynthesis (all $P < 10^{-3}$ and FDR < 5%) (**Fig. 3.19a**).

Considering the high conservation of *CD2758* across *C. difficile* genomes, I attempted to explore the RNA-seq data further beyond the sporulation phenotype and the GO term analyses discussed above, with a special focus on biological processes critical to *C. difficile* infection. I took an integrative transcriptomic strategy to search for critical *C. difficile* biological processes that may involve epigenetic regulation. Specifically, I performed an overlap analysis between the list of DE genes from our RNA-seq data (wild-type vs. $\Delta CD2758$ mutant; four different time points), and those from published studies focusing on the colonization and infection by this pathogen. Using DE genes obtained from murine gut isolates at increasing time points after infection [Fletcher *et al.* 2018], I found significant overlaps with DE genes in the $\Delta CD2758$ mutant (O/E=1.8, $P < 10^{-4}$, Chi-Square test) (**Fig. 3.19b**).

Hence, these results consolidate the involvement of *CD2758* in the sporulation process (particularly visible at the 19 h sporulation time point) and provide molecular evidence supporting reduced sporulation initiation in $\Delta CD2758$ cells. Moreover, this integrative overlap analysis provide additional evidence that DNA methylation events by *CD2758* may directly and/or indirectly affect the expression of multiple genes involved in the *in vivo* colonization and biofilm formation of *C. difficile*, and inspire future work to elucidate the mechanisms underlying the functional roles of CAAAAA methylation in *C. difficile* pathogenicity.

3.5 Conclusions

Fig. 3.20 shows an overview of my past research, highlighting major questions addressed, main methods / approaches used, and corresponding publications. In a

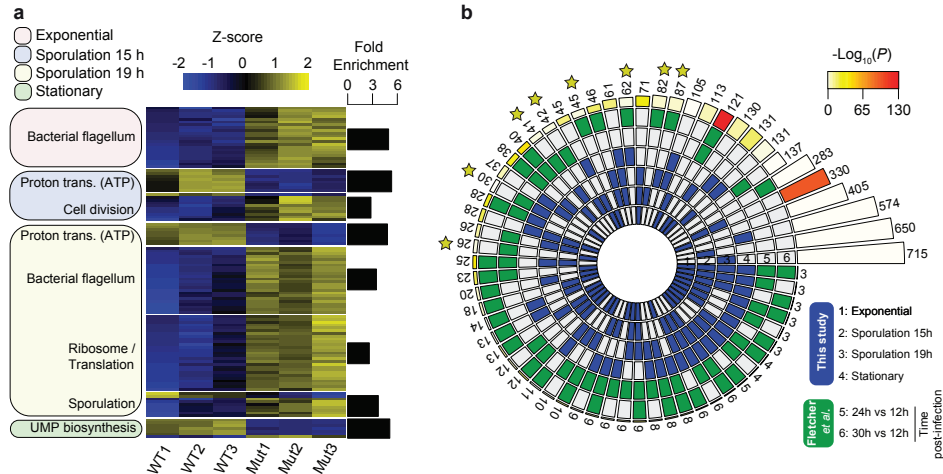


Figure 3.19: Gene expression analysis. (a) Heatmap of 92 DE genes in three replicates of *C. difficile* 630 Δ *erm* compared to equal number of replicates of *C. difficile* 630 Δ *erm* Δ *CD2758*. The Z score reflects the degree of down- (Z-score < 0) or up- (Z-score > 0) regulation, computed by subtracting the mean of the log-transformed expression values and dividing by the standard deviation for each gene over all samples scored. (b) Significance of overlap between multiple datasets of DE genes. Comparisons were performed between DE genes called in this study for each time point (blue-shaded) and those from Fletcher et al. (green-shaded). The latter corresponds to *C. difficile* DE genes called 24 and 36 h (versus 12 h) of post-infection time in a murine model. Color intensities of the outermost layer represent the *P*-value significance of the intersections (3,896 genes were used as background). The height of the corresponding bars is proportional to the number of common genes in the intersection (indicated at the top of the bars). Stars indicate pairwise comparisons between the different studies. Taken from [Oliveira *et al.* 2020].

nutshell, I showed during my Ph.D. that multiple unwanted DNA instability events take place in plasmid vectors commonly used in the development of DNA biopharmaceuticals. I identified these events, quantified them, built safer host genomes for vector propagation, and proposed safety guidelines to be followed by the scientific community. Altogether, my Ph.D. allowed me to gain expertise in microbial genetics, multiple wet-lab techniques, and state-of-the-art genome editing approaches (**Fig. 3.20**). It also allowed me to develop research and acquire unique know-how at world-renowned institutions, including the Massachusetts Institute of Technology (Kristala Prather Lab) and Harvard Medical School (George Church Lab).

For my first postdoc, I explored additional aspects of genetic instability, this time in human stem cells. This allowed me to leverage my knowledge of eukaryotic biology, and my skills in different techniques used during the handling, expansion, differentiation, and quality-control of these cell products for clinical applications.

For my second postdoc, I decided to move to Eduardo Rocha's lab at Institut Pasteur in Paris. This allowed me to deepen my skills in bioinformatics, while still working on aspects related with genome dynamics. During this period, I acquired multiple novel competences in comparative genomics, phylogenomics, population genomics, and UNIX / R programming. My work contributed to a more complete picture of bacterial defense systems, and their abundance and distribution across MGEs.

As a senior scientist at Mount Sinai School of Medicine, I explored a different but complementary angle of bacterial epigenetics, by extracting critical information from DNA polymerization kinetics in SMRT-seq, and by integrating it with meta-transcriptomics and functional genomics data. This allowed me to recently perform the first comprehensive characterization of the epigenomic landscape of *C. difficile*, and to discover a conserved MTase that regulates aspects of its biology that are critical to its transmission to humans.

My past work has proven to be influential in three ways. Firstly, it provided evidence for multiple novel events of genetic instability in DNA and cell-based biopharmaceuticals. Secondly, it advanced the current understanding of bacterial defense systems, and their interplay with HGT. Thirdly, it suggested a broader and more intricate role of DNA methylation in regulating key biological traits in bacteria. This raises numerous interesting and novel questions, which I would like to further pursue. Ultimately, I pretend to further explore the diversity, dynamics and evolution of bacterial epigenomes combining comparative epigenomics, meta-transcriptomics, and microbial genetics.

	QUESTION	METHODS	PUBLICATIONS
PhD Thesis 2006-2009	Detection of novel deletion-formation events in plasmid-based biopharmaceuticals	Molecular biology techniques Microbial genetics Gene expression analyses	Ribeiro et al. 2008 Oliveira et al. 2009a, 2009b, 2010, 2011, 2013
	Development of a predictive tool for estimating recombination frequencies in multicopy plasmids	Meta analysis of recombination frequencies Model development	Oliveira et al. 2008
	Insertion sequence-mediated genetic instability in plasmid biopharmaceuticals	Bioreactor technology Genome editing Basic bioinformatic analyses	Lewis et al. 2012 Gonçalves et al. 2014
1st Postdoc 2010-2012	Effect of hypoxia and prolonged passaging on the genomic stability of <i>ex-vivo</i> expanded human stem / stromal cells.	Stem cell culture / expansion Microscopy / Image analysis Immunophenotyping and stem cell differentiation Microsatellite, telomere analysis	Oliveira et al. 2012 Oliveira et al. 2014a
	Effect of hypoxia and prolonged passaging on mitochondrial performance	Whole mtDNA sequencing and analysis Haplotyping Flow cytometry	Oliveira et al. 2012 Oliveira et al. 2014a
	Role of non-canonical DNA structures and sequence motifs on human mitochondrial DNA instability	Meta-analysis of deletion breakpoints Bioinformatic analyses of non-B DNA De-novo motif finding, selection and validation	Oliveira et al. 2013
2nd Postdoc 2013-2016	Interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts	Programming skills (UNIX shell) Phylogenetic analysis Comparative genomics	Oliveira et al. 2014b
	Regulation of genetic flux between bacteria by restriction modification systems	Techniques to infer homologous recombination Reconstruction of the evolution of gene families	Oliveira et al. 2016
	Chromosomal organization of horizontal gene transfer in bacteria	Core / pan genome analyses Gene functional analysis Measures of gene repertoire diversity	Oliveira et al. 2017
Senior Scientist 2016-2020	Methylome analysis of <i>Clostridioides difficile</i>	Programming skills (R) SMRT-sequencing and genome assembly Wavelet analysis	Oliveira et al. 2020
	Transcriptional analysis of <i>Clostridioides difficile</i>	RNA-seq, read mapping Differential expression analysis Gene functional analysis	Oliveira et al. 2020
	Detection and analysis of multiple genomic elements	Multiple pipelines to detect TSSs, TFBSs, CRISPRs, TAs, prophages, conjugative/mobilizable elements, integrons, BREX, Abi, DISARM	Oliveira et al. 2020

Figure 3.20: Overview of my past work.

Current and Future research

Contents

4.1	Context	60
4.2	Role of persistent MTases in Bacteria	61
4.2.1	Introduction	61
4.2.2	Research program	63
4.3	Interplay between diversification of methylation systems, their target specificity and genetic mobility	66
4.3.1	Introduction	66
4.3.2	Research program	68
4.4	The anti-phage defensible of complex microbial populations	69
4.4.1	Introduction	69
4.4.2	Research program	70
4.4.3	Preliminary results	72
4.5	Final remarks	76

4.1 Context

The information content of the genetic alphabet is not limited to the primary nucleotide sequence but is also conveyed by chemical modifications of individual bases. This additional layer of 'epigenetic' information shows distinct degrees of chemical diversity and complexity, and is found to be pervasive across all kingdoms of life. Propelled by recent progresses in third-generation sequencing technologies — SMRT-seq by Pacific Biosciences and nanopore sequencing by Oxford Nanopore Technologies (ONT) — more than 4,500 methylomes have been mapped to date. As a consequence, the field of bacterial epigenomics is witnessing a remarkable expansion beyond single methylome analyses to the realm of multi-omic data integration. My recent findings on the epigenomics of *C. difficile* [Oliveira *et al.* 2020] add to the growing number of studies integrating multi-omics profiling to identify putative epigenetic regulation networks. Fueled by this exciting momentum, I propose to combine high-throughput (epi)genomic technologies and bioinformatic approaches to address outstanding questions linked to the diversity, dynamics and evolution of bacterial methylation systems. The next three subsections describe in detail ongoing research and how these lines of study will likely unfold in the future.

4.2 Role of persistent MTases in Bacteria

4.2.1 Introduction

I recently found that the number of solitary MTases exceeds that of complete R-M systems in bacteria [Oliveira *et al.* 2014b]. This can be explained by their involvement in biological roles beyond cell defense including initiation of DNA replication, DNA repair, and gene regulation. Some well-known examples include Dam and Dcm from *E. coli*, CcrM from *Caulobacter crescentus* [Mouammine & Collier 2018], and more recently, a GCGC-specific 5mC MTase from *H. pylori* [Estibariz *et al.* 2018], and a 6mA MTase recognizing the CAAAAA motif in *C. difficile* [Oliveira *et al.* 2020]. Overall, the above-mentioned solitary MTases appear as promising targets for better understanding the involvement of epigenetic regulation in clinically relevant phenotypes: they are persistent, i.e., endure strong selective pressure for retention in the majority (or eventually the totality) of genomes of a species, and deemed (quasi-)essential for cellular viability and survival in a particular environment. Interestingly, they seem to be the tip of a much larger iceberg. In a recent large-scale analysis in bacteria, I found that 45% of the species contained at least one persistent MTase [Oliveira & Fang 2021], which shows that the latter are abundant and suggests that their epigenetic modifications may be important and frequent (Fig. 4.1).

The growing number of available bacterial epigenomes has spurred a surge in comparative epigenomic studies. However, despite the large abundance of solitary and persistent MTases in bacteria, more comprehensive studies are needed to fully characterize the precise mechanisms by which DNA methylation modulates gene expression and alters phenotypes. Here I propose to use short- and long-read sequencing technologies combined with microbial genetics, and comparative genomics to advance our understanding of the diversity and functional relevance of DNA methylation in bacteria. This first research line addresses outstanding questions put forward in the bacterial epigenomics field, namely:

- What are the functional roles of persistent MTases in Bacteria?
- Is there a link between methylome and the clinical success of certain human pathogens?
- To what extent is methylation shaping (and being shaped) by DNA topology?

In particular, I will focus on the persistent MTases recognizing RAATTY and CCWGG, which are ubiquitous in *A. baumannii* and *K. pneumoniae*. The following considerations led to these choices: (i) both species have emerged as major causes of health care-associated infections due to the extent of its antimicrobial resistance and propensity to cause large nosocomial outbreaks; (ii) they are ranked by the World Health Organization as priority pathogens for purposes of research and development; (iii) there are no large-scale epigenomic studies performed in these bacteria so far; and (iv) these persistent MTases have homologues with relevant

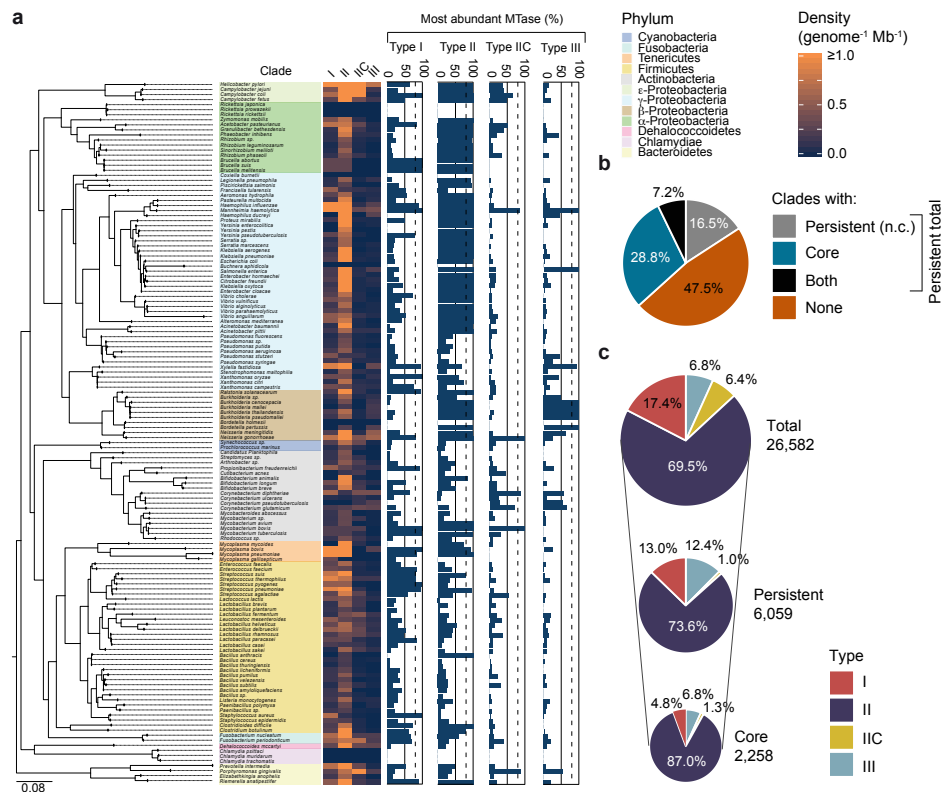


Figure 4.1: (a) Phylogenetic tree of 139 bacterial species (colored by Phylum), for which at least ten complete genomes were available at Genbank (corresponding to a total of 5,568 genomes). Heatmap corresponds to the density (per genome per Mb) of Types I, II, III MTases and type IIC R-M systems for each species. Bar plots indicate the percentage of the most abundant MTase(s) found in each species assuming as inclusion criteria a minimum of 80% similarity in amino acid sequence and less than 20% difference in protein length. Stippled lines indicates a threshold of 80%, above which an MTase can be considered persistent. A core gene is denoted by 100%. (b) Pie-chart summarizing the percentages of species analyzed containing either persistent non-core (n.c.) MTases, core MTases, both, or none. (c) Pie-charts showing the breakdown of total, persistent, and core MTases per type. Taken from [Oliveira & Fang 2021].

functional roles in DNA uptake, biofilm formation, cell aggregation and motility [Vandenbussche *et al.* 2020, Beauchamp *et al.* 2017].

4.2.2 Research program

4.2.2.1 Methylome analysis and diversity

A collection of 50 *A. baumannii* and 100 *K. pneumoniae* clinical isolates stemming from the French National Reference center's strain collection will be sequenced using short- and long-read technologies (**Fig. 4.2a**). These isolates represent the main clones spreading in France, some of which belong to known high risk clones. Isolates of environmental and veterinarian origin will also be selected. Nanopore sequencing libraries will be prepared according to the '1D Genomic DNA by ligation' protocol provided by ONT. Reads will be base called using **Guppy**, mapped to corresponding references using BWA-MEM, and strand separated according to the initial alignment (with **Rsamtools**). Events will be associated with genomic positions with **Nanopolish eventalign**, and normalized across reads by correcting signal scaling and shifting. Detection of methylated sites will be performed by combining consecutive *P* values with Fisher's method (sumlog function) in sliding windows of 5 bp, and smoothing the signal along the genome. To validate the methylation data obtained by ONT, I will use SMRT-seq. The **RS_HGAP3** protocol will be used for de novo genome assembly, followed by the use of **custom scripts** for genome finishing and annotation. The **RS_Modification_and_Motif_Analysis** will be used for de novo methylation motif discovery. A **custom script** will be used to examine each motif to ensure its reliable methylation state. Methylation motif exceptionality will be performed using **RMES**. Analysis of motif abundance will be performed using a wavelet-based **multi-scale representation** of genomic signals. To further characterize methylation motif sites I will perform whole-genome alignment of the isolates using **progressiveMauve** and classify each of the former as either orthologous conserved (no SNPs or indels) or orthologous variable (with SNPs and/or indels). All this data will be integrated with the repertoire of DNA MTases, R-M systems, and other defense systems (defensome). Briefly, identification of R-M systems will be performed as previously [Oliveira *et al.* 2014b], and rely on curated Hidden Markov Model (HMM) profiles **already available**. For the remaining defensome I will proceed as previously [Oliveira *et al.* 2020]. Briefly, CRISPR repeats will be identified using the **CRISPR Recognition Tool**. For cas gene identification, I will obtain Cas protein family HMMs from the TIGRFAM database and PFAM families annotated as **Cas families**. Toxin-Antitoxin (T-A) will be detected with **TAFinder**. Genes pertaining to abortive infection systems, Bacteriophage Exclusion (BREX), DISARM, and other recently found systems, will be identified using recently published HMM profiles [Doron *et al.* 2018].

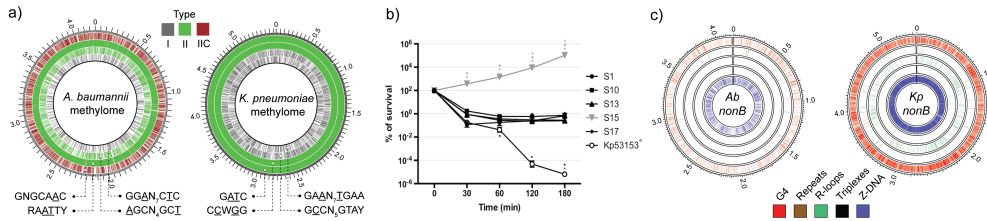


Figure 4.2: (a) The *A. baumannii* and *K. pneumoniae* methylomes. Shown are the positions of all methylation motif sites in two representative clinical isolates, colored according to MTase type. The corresponding sequences were obtained by the [RS_Modification_and_Motif_Analysis](#) tool from Pacbio. Further comparative epigenome analysis of a larger cohort of samples, will allow a comprehensive view of methylation target diversity. (b) Example of *K. pneumoniae* serum bactericidal assay. The average of survival (%) of multiple strains in human serum is plotted against the incubation time ($n = 4$, mean \pm SEM). All strains were compared to the strain S17. $*P < .05$; $**P < .01$; $***P < .001$, t test. (c) The landscape of *A. baumannii* and *K. pneumoniae* non-B propensity computed with in-house-developed pipelines. Multiple non-B forms (particularly G4, R-loops, and Z-DNA) were found to overlap methylation motifs in (a). ssDNA sequencing will allow confirming the presence of such non-B forms, and clarify their role as topological constraints influencing the access of the MTase’s catalytic site to overlapping methylation motifs.

4.2.2.2 Functional relevance of bacterial methylomes

To delve into the functional relevance of methylation at RAATTY and CCWGG sites, I will compare the transcriptome of WT strains with those of MTase mutants. Briefly, I plan to introduce point and frameshift mutation by using Transient Mutator-Multiplex Automated Genome Engineering shown to be efficient in *K. pneumoniae* [Gallagher *et al.* 2014]. For *A. baumannii*, point mutations will be introduced by homologous recombination using sacB and apramycin resistance for counter selection [Amin *et al.* 2013]. Mutants and WT of c.a. 20 clinically-relevant strains (10 *A. baumannii* and 10 *K. pneumoniae*) will be tested phenotypically by a progressive approach for multiple virulence/adaptation traits that may be impacted by methylome abrogation. Antibiotic susceptibility testing, susceptibility to anti-septic, and fitness in rich and minimal media (MH, LB and M9) will be conducted. More specific tests relevant to hospital and host adaptation will be performed, such as analysis to the survival to stresses, to desiccation, long periods in distilled water, in human urine, blood, and serum, the capacity to form biofilm (known to be impacted in homologous mutants [Vandenbussche *et al.* 2020, Beauchamp *et al.* 2017], and lung epithelium cell adherence (Fig. 4.2b). If deemed appropriate, *in vivo* virulence assays using *Galleria mellonella* and the mouse model of infection (systemic and pulmonary) will also be tested [Rubio *et al.* 2021]. For strains capable of intracellular replication, I will monitor this phenotype by microscopy [Rubio *et al.* 2021].

For RNA-seq, RNA will be extracted from three biological replicates of cultures of WT and MTase mutant strains, DNase-treated, rRNA-depleted and converted to cDNA. Read quality will be checked using [FastQC](#). I will use Trimmomatic to remove adapters and low-quality reads. Subsequently, rRNA sequences will be filtered from the dataset using SortMeRNA on the basis of the SILVA 16S and 23S rRNA databases and the Rfam 5S rRNA database. The resulting non-rRNA reads will be mapped to the reference genomes using BWA-MEM. The resulting BAM files will be sorted and indexed using SAMTOOLS, and read assignment will be performed using featureCounts. Normalization and differential-expression testing will be performed with the Bioconductor package DESeq2. Functional classification of genes will be performed using the [DAVID online database](#). The reproducibility of DAVID's functional classification will be tested with [Blast2GO](#) and [Panther](#). Briefly, for Blast2GO, I will run BLASTX searches of both species' genomes against the entire GenBank bacterial protein database. The output will be loaded into Blast2GO, and mapping, annotation and enrichment analysis will be performed as previously [indicated](#). For Panther, I will download the most recent [HMM library](#) and annotate the *A. baumannii* and *K. pneumoniae* proteomes with [pantherScore2.1.pl](#). Both input and background gene lists will be formatted to the Panther Generic Mapping File type as described [here](#). Genomic regions with significant high density of methylation sites nearby genes that are significantly mis-regulated in a given MTase mutant (typically seen in promoters), will be targeted for site directed-mutagenesis in order to gauge the impact of each methylation site mutation.

4.2.2.3 Interplay between methylation, DNA regulatory elements and DNA topology

Nearly all occurrences (mostly >95%; often >99%) of a motif recognized by an active MTase in a given prokaryotic genome are methylated. Previous bacterial methylome studies that analyzed one or few genomes had insufficient statistical power to perform a systematic interrogation of non-methylated motifs sites. Building on the large collection of methylomes, I will be able to perform a systematic detection and analysis of such positions as described in 4.2.2.1. I will then test the hypothesis that the on/off switch of DNA methylation in these species can contribute to epigenetic regulation as a result of competitive binding between MTases and other DNA-binding proteins (e.g., transcription factors, TFs). Prediction of TF binding sites (TFBSs) will be performed by retrieving *A. baumannii* and *K. pneumoniae* regulatory sites in FASTA format from the RegPrecise database [[Novichkov et al. 2013](#)]. These will be converted to position-specific scoring matrices using in-house-developed scripts. Matches between these matrices and our genome dataset will be performed using MAST, and filtered on the basis of *P* value. If any predicted TF overlaps non-methylated sites, occupancy maps will be experimentally validated by chromatin immunoprecipitation sequencing (ChIP-seq). Beyond TFBSs, we currently lack a holistic genome-scale view of how DNA methylation at a given target motif may be affected by DNA topology. In this regard, I will test the hypothesis

that non-B DNA secondary conformations (e.g.: Z-DNA, G-quadruplexes, DNA repeats) may hinder (or alternatively ease) the access of the MTase's catalytic site to its recognition motif, and also act as important epigenetic regulators. Our propensity mapping of non-B conformations in *A. baumannii* and *K. pneumoniae* already revealed a profuse variety of structures (**Fig. 4.2c**). Such mapping was performed with dedicated bioinformatic tools [Vlahovicek *et al.* 2003, Chiu *et al.* 2016, Bedrat *et al.* 2016, Jenjaroenpun *et al.* 2015, Achaz *et al.* 2007, Wang *et al.* 2004, Buske *et al.* 2012, Ho *et al.* 1986], and in-house developed scripts capable of performing a whole-genome screening of all currently known non-B DNA structures in an accurate and time-effective fashion (e.g.: the complete analysis of an *E. coli* genome takes ~ 2 hours of elapsed wall clock time). I will build on such data and experimentally validate such propensity via ssDNA sequencing [Kouzine *et al.* 2017]. Briefly, the latter uses potassium permanganate and an S1 nuclease treatment to introduce double-stranded breaks in DNA regions enriched for unpaired bases *in vivo*. These breaks are then labelled with biotin, and the DNA is sonicated, streptavidin-selected, and sequenced. The statistical significance of the overlaps between non-B DNA regions and methylated / non-methylated sites will be evaluated in R. The statistical significance of the overlaps between non-B DNA regions and methylated / non-methylated sites will be evaluated in R.

4.3 Interplay between diversification of methylation systems, their target specificity and genetic mobility

4.3.1 Introduction

I have recently observed an overabundance of MTases (both solitary and belonging to R-M systems) in MGEs [Oliveira *et al.* 2014b]. Several of these MTases were non-specific (i.e.: recognize degenerate target sites), which increases the spectrum of possible methylation signatures. Such property might be exploited to overcome common host restriction barriers employing, for example, Type I and III restriction enzymes as the principal barrier to control against infection. In other words, these MTases may serve as antidotes against R-M systems and thereby facilitate infection of new hosts and competition with other MGEs. Heterogeneous methylation (either as an ON \leftrightarrow OFF switch or through modifications of the target motif sequence), can be caused by spontaneous mutations in MTase coding genes by slipped-strand mispairing of simple sequence repeats (SSRs) or recombination between inverted repeats (the so called phasevarions or phase-variable regulons) [Casadesús & Low 2013]. Such heterogeneous methylation may lead to changes in gene expression and generate phenotypic plasticity within isogenic populations, thus aiding in the rapid adaptation to environmental shifts. While several studies have described multiple strategies of methylation systems' diversification, we currently lack a detailed understanding of the interplay between such diversification, methylation specificity, and genetic mobility. In particular, in this second research line I

propose to:

- Test the hypothesis that MGEs exploit methylation systems with distinct degrees of recognition target specificity to shape the former's routes of transfer.
- Perform a large-scale estimation of the recombination potential of all currently known methylation systems, and test the association of such potential with HGT, its mechanisms and its vectors.

For the first point, I specifically propose to disentangle between two scenarios (not mutually exclusive): *i*) that MGEs favor MTases with degenerate/multispecific target recognition sites to provide defense against multiple R-M systems and allow for HGT to take place across a broader host range of individuals; or *ii*) that MGEs favor MTases recognizing very specific target motifs to establish preferential routes of genetic flux across well-defined lineages (**Fig. 4.3**). Confounding variables such as family and size of MGE and type of methylation system will be accounted for.

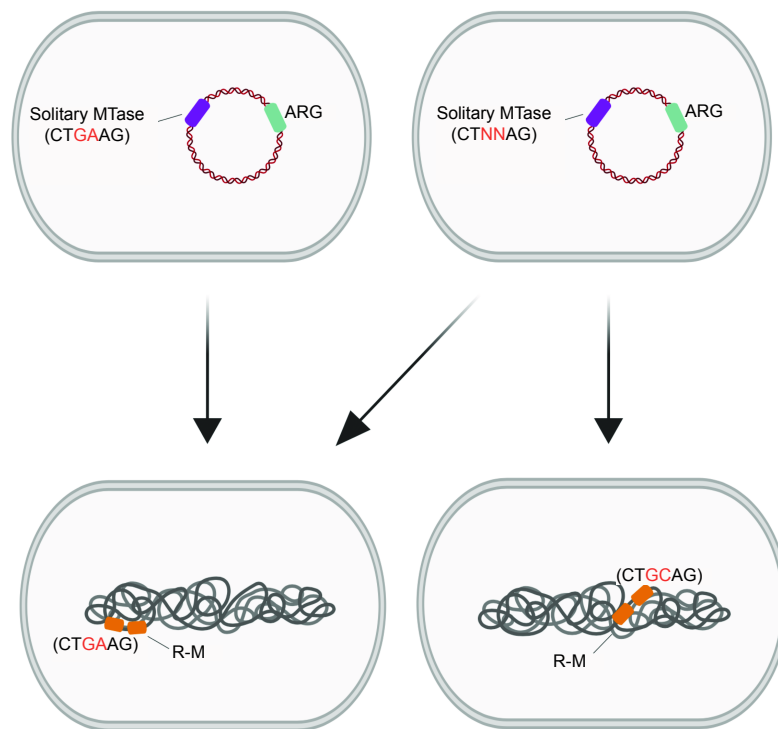


Figure 4.3: MGEs carrying MTases with non-specific (i.e.: degenerate target sites) have a broader host range.

For the second point, I will build upon recent analyses [Atack *et al.* 2020, Atack *et al.* 2018], and perform the most comprehensive and up-to-date characterization of the distribution and recombination potential of short direct repeats and SSR tracts in Type I (*hsdS*, *hsdM*, and *hsdR* genes), Type II and Type III (*mod*

genes) R-M systems, as well as inverted repeats in *hsdS* genes. Altogether, this research line will help understanding, for example, why certain plasmids have higher transfer ranges, and consequently lead to a broader dissemination of antibiotic resistance genes (ARGs). Or why certain phages are capable of infecting bacterial cells more effectively, become integrated within complex cell circuits, and eventually contribute to functional innovation. It will ultimately allow establishing a link between genetic mobility and the recombination potential of phasevarians.

4.3.2 Research program

This research line builds on the large collection of Pacbio and Nanopore methylomes available on the literature and on the REBASE database (<http://rebase.neb.com/cgi-bin/ssonelist?m2>). For example, at the time of writing, REBASE contained a total of 4,500 methylomes and 9,301 MTases whose target specificity was validated by third-generation sequencing. With the help of dedicated in-house developed scripts, all associated metadata (genome nucleotide and protein sequences, R-M types and methylation motifs) will be retrieved, and the latter classified as either multispecific (containing at least 1 ambiguous base) or specific (no ambiguous bases). Classification of MTases as belonging to intact R-M systems or as solitary will be performed on the basis of a guilty-by-association strategy [Oliveira *et al.* 2014b]. Identification and classification of MGEs will be carried out as previously [Oliveira *et al.* 2014b]. Briefly, prophages will be detected using PhageFinder [Fouts 2006] under strict mode, and PHASTER [Arndt *et al.* 2016] under default settings. I will take the common hits obtained by both programs, as well as those very few cases (commonly 10% of the hit list) corresponding to complete prophages predicted by just one of the programs. All elements smaller than 18 kb, or lacking matches to core phage proteins (e.g. terminase, capsid, head, tail proteins) will be removed. Integrons will be searched with IntegronFinder [Cury *et al.* 2016] under default settings. The identification of genes encoding the functions related to conjugation in ICEs will be performed as previously described [Cury *et al.* 2017]. Briefly, an element will be considered as conjugative when it contains the following components of the conjugative system: a VirB4/TraU ATPase, a relaxase, a coupling ATPase (T4CP), and a minimum number of mating pair formation (MPF) type-specific genes: two for types MPF_{FA} and MPF_{FATA}, or three for the others (types F, T, and G). In the case of IMEs, they will be identified by the fact that they encode relaxases but lack a complete conjugative transfer system, which is encoded in trans by another mobile element. Delimitation of ICEs and IMEs will be performed considering flanking core genes as upper bounds for their extremities. With such data at hand, I will be able to finally test the significance of the association between MTases, their methylation specificity, and mechanisms of mobility, while controlling for confounding factors such as MGE family, MTase type, host taxa, and genome size. Since MGEs are preferential reservoirs of ARGs, I will also test the hypothesis that degenerate MTases preferentially co-localize with MGEs harboring ARGs. For chromosomally-encoded MGEs harbouring MTases, I will investigate their genomic

context and organization, and test if these MGEs preferentially cluster in hotspots as previously performed [Oliveira *et al.* 2017]. Given the growing evidence showing that many bacterial species explore genetic mechanisms in genes of R-M systems (phasevarions) capable of modifying the transcriptome of the cell via epigenetic control, I will estimate the recombination potential of such phasevarions and correlate it with mechanisms of genetic mobility. In particular, I will use an evolutionary failure mode calculator [Jack *et al.* 2015] to compute for each R-M / MTase containing hypermutable SSRs and close repeats, their mutation rate relative to a sequence of the same length that does not include any predicted mutational hotspots. Such approach will ultimately allow testing if phasevarion-containing R-M systems / MTases carried by MGEs, present a recombination potential that significantly departs from that found in similar systems outside MGEs.

4.4 The anti-phage defensome of complex microbial populations

4.4.1 Introduction

Bacteria and archaea are at risk for both cell death and genomic invasion by a diverse set of genetic parasites (e.g. : phages), and as a result have developed an array of sophisticated lines of active defense that can collectively be referred to as the prokaryotic anti-phage defensome. Based on their action modes, the defensome components can be divided into two major groups: immunity and programmed cell death. The immunity group comprises: *i*) R-M systems that target specific sequences on the invading phage [Oliveira *et al.* 2014b, Oliveira *et al.* 2016]; *ii*) CRISPR-Cas [Koonin *et al.* 2017] which provide acquired immunity through memorization of previous phage attacks; *iii*) DNA Phosphorothioation (PT) systems [Wang *et al.* 2011] that replace non-bridging oxygen by sulphur on the DNA sugar-phosphate backbone; and *iv*) additional systems such as BREX [Goldfarb *et al.* 2015], prokaryotic Argonautes (pAgos) [Makarova *et al.* 2009], DISARM [Ofir *et al.* 2018], DRUANTIA, GABIJA, and ZORYA [Doron *et al.* 2018] whose mechanisms of action are not yet clear. On the other hand, the programmed cell death group includes: *i*) T-A systems which play roles in phage defense, virulence, or as DNA maintenance modules [Unterholzner *et al.* 2013]; and *ii*) Abortive Infection (ABI) systems that lead to cell death or metabolic arrest upon infection [Dy *et al.* 2014]. The past few decades have revealed extensive insights on how biotic interactions, such as competition, symbioses, HGT, and predation, play a role in the distribution and diversity of microbial communities across multiple biomes. The presence of phages in these communities adds an extra layer of complexity as it can lead to escalation, (similar to an arms-race), where each party invests in a greater number of weapons and/or defenses, resulting in mutual directional natural selection. And while the abundance, distribution, and diversity of anti-phage defense systems has been well characterized on genomic data, we currently lack a holistic view of the defensome across distinct

environmental and host-associated microbial populations. In this third research line I aim to perform a large-scale mapping of the defensomes of high-quality MAGs reconstructed from environmental (e.g.: oceans) and human-associated (e.g.: gut) metagenomic datasets. (**Fig. 4.4**). I will then use comparative genomics to evaluate the variability and distribution of defensome components, and test their association with mechanisms of genetic mobility. The presence of previously unidentified genes enriched in defense islands will also be assessed. While the first two research lines described in sections 4.2 and 4.3 mainly focused on methylation systems (part of R-Ms, one of the most well-studied type of defense system), here I seek to extend our insight into the abundance, genomic diversity and interplay between multiple defense systems. Ultimately, this project will deepen our understanding on the extent to which phage-host interactions play a role in shaping diversity and structure of complex microbial communities, as well as pinpoint previously unappreciated mechanisms by which immunity spreads.

4.4.2 Research program

High-quality (completeness >90%, contamination <5%, N50 >300 kb, at least 18 tRNAs) non-redundant prokaryotic MAGs will be downloaded from large-scale studies focusing on the [human gut microbiome](#) [Almeida *et al.* 2019], and [ocean metagenomes](#) [Delmont *et al.* 2021]. MAG annotation, quality checks, and taxonomy analyses will be performed with Prokka [Seemann 2014] and MAGpy [Stewart *et al.* 2019]. Identification of R-M systems will be performed as previously described [Oliveira *et al.* 2014b]. Briefly, curated reference protein sequences of Types I, II, IIC and III R-M systems and Type IV REases will be downloaded from the data set ‘gold standards’ of REBASE [Roberts *et al.* 2015]. All-against-all searches will be performed for REase and MTase standard protein sequences retrieved from REBASE using BLASTP (default settings, e-value < 10⁻³). The resulting e-values will be log-transformed and used for clustering into protein families by Markov Clustering (MCL). Each protein family will be aligned with MAFFT using the E-INS-i option, 1,000 cycles of iterative refinement, and offset 0. Alignments will be visualized in AliView and manually trimmed to remove poorly aligned regions at the extremities. HMM profiles will be built from each multiple sequence alignment (deprecated version available at the lab’s [Github page](#)) using the hmmbuild program from the HMMER suite (default parameters). Types I, II and III R-M systems will be identified by guilty-by-association, searching genes encoding the MTase and REase components at less than five genes apart. CRISPR repeats will be identified using the CRISPR Recognition Tool (CRT) [Bland *et al.* 2007] with default parameters. For CRISPR spacer homology search, I will consider as positive hits those with at least 80% identity. For cas gene identification, I will obtain Cas protein family HMMs from the TIGRFAM database [Haft *et al.* 2003] and PFAM families annotated as [Cas families](#). Genes pertaining to ABI systems will be searched with the PFAM profiles PF07751, PF08843, and PF14253. BREX systems will be searched using PFAM profiles for the core genes pglZ (PF08655) and brxC/pglY

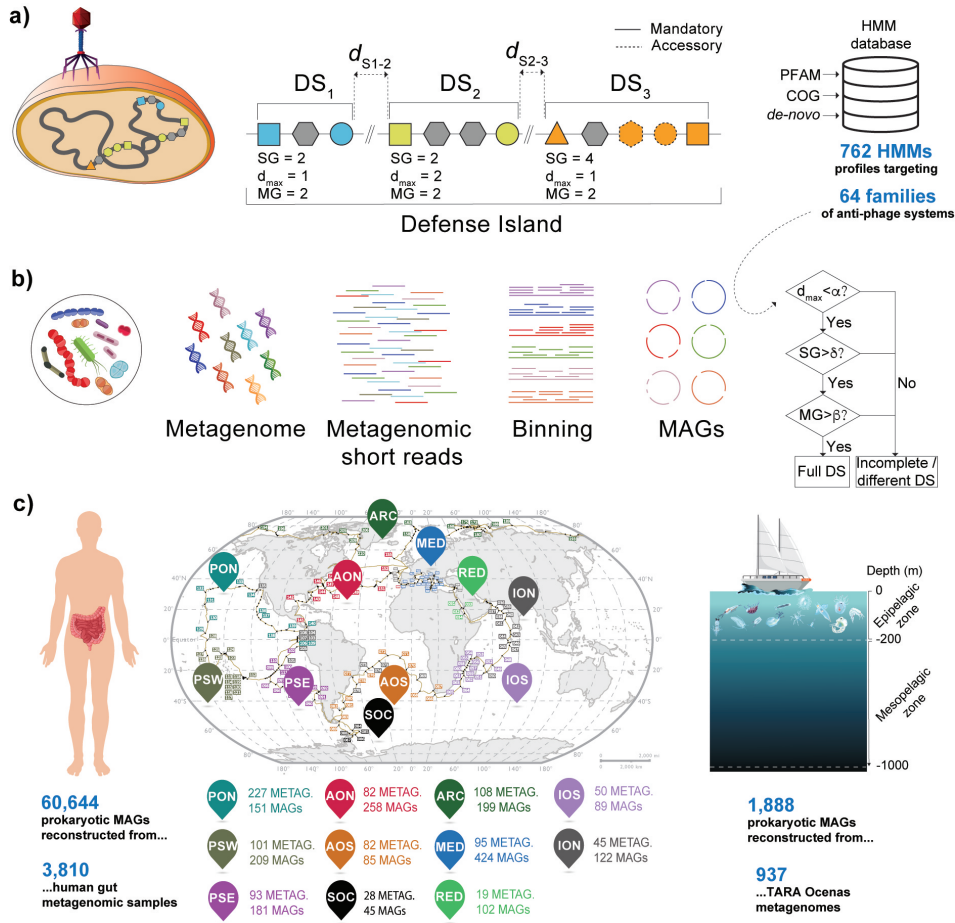


Figure 4.4: Defenseome analysis. (a) My approach relies on building the most complete HMM database of defense systems described with at least one experimental evidence of anti-phage function in the literature. The protein profiles used were retrieved from existing databases (PFAM, COG), taken from a recently developed tool [Tesson *et al.* 2021] or built from scratch when no suitable profiles existed. This led to an overall of 762 HMM profiles targeting 64 families of anti-phage systems. I then defined genetic organization rules based on the literature, allowing for two types of genetic components: "mandatory" and "accessory". Given the wide diversity of genetic organization of anti-phage systems, rules were written differently for different types of systems. Typically, for small systems (less than 3 proteins), the number of mandatory proteins required were strict whereas for bigger systems, the number of proteins required did not always require all components to be present. Shown is an example of a defense island containing three defense systems DS₁-DS₃, respectively characterized by a total sum of genes (SG), a maximum distance between defense genes (d_{max}), and a given number of mandatory genes (MG). (b) HMMs will be used to query high-quality near-complete MAGs, and the above organization rules will allow disentangling between complete or incomplete defense systems. (c) MAGs will be obtained from [Almeida *et al.* 2019] focusing on human gut microbiomes and [Delmont *et al.* 2021] focusing on ocean metagenomes recovered within the frame of the TARA project. A geographical representation of the collection stations of the TARA project is shown. ARC: Arctic Ocean; MED: Mediterranean Sea; RED: Red Sea, ION: Indian Ocean North; IOS: Indian Ocean South; SOC: Southern Ocean; AON: Atlantic Ocean North; AOS: Atlantic Ocean South; PON: Pacific Ocean North; PSE: Pacific South East; PSW: Pacific South West.

(PF10923), and specific PFAM profiles for each BREX type as indicated previously [Goldfarb *et al.* 2015]. DISARM systems will be identified using the PFAM signature domains (PF09369, PF00271, PF13091) belonging to the core gene triplet characteristic of this system [Ofir *et al.* 2018]. To search for pAgos I will build a dedicated HMM profile based on a list of 90 Ago-PIWI proteins [Makarova *et al.* 2009]. Searches for the ensemble of newly found antiphage systems will be performed using the list of PFAM profiles published by the authors [Doron *et al.* 2018]. Type II T-A systems will be detected using the TAFinder tool with default parameters. Computation prediction of putative novel defense systems will be performed as previously [Doron *et al.* 2018]. Briefly, PFAM annotations for bacterial and archaeal genes will be obtained from the Integrated Microbial Genomes database, and cross-referenced to the defense neighboring genes (± 10 genes) in the MAGs using `hmmsearch`. The protein coding sequences for neighboring genes for all family members will be clustered based on sequence homology with OrthoMCL with `blastp` parameters [-F 'm S' -v 100000 -b 100000 -e 1e-5 -m 8] and with MCL with inflation value of 1.1. Circus plots of phylogenetic abundance / distribution of anti-phage defensesome systems across MAGs, co-localization / co-occurrence with MGEs (including statistical significance), and genomic representation of novel defense systems will be performed with R.

4.4.3 Preliminary results

To illustrate the feasibility and potential of this third line of research, I performed some preliminary analyses on the TARA MAG dataset, whose results I show below. Briefly, I built upon 1,888 MAGs (94% bacteria and 6% archaea) reconstructed from 937 metagenomes derived from the TARA Oceans expedition. The latter were recovered from the surface and deep chlorophyll maximum layers from stations covering the Pacific, Atlantic, Indian, Arctic, and Southern oceans, as well as Mediterranean and Red seas. After applying my detection pipeline to the ensemble of TARA MAGs, I obtained a total of 2,286 complete defense systems and 103,636 defense genes belonging to 50 different anti-phage families (**Fig. 4.5a**). R-M systems were found to be the most abundant both in bacteria (with roughly 20%) and archaea (about 8%), presumably because they are among the very few capable of inducing post-segregational killing. The archaeal defensesome was globally less diverse and abundant than bacteria, which likely stems from the fact that HGT is substantially less frequent across the former compared to the latter. Interestingly, and contrary to previous observations that CRISPR-Cas systems are present in nearly all archaeal genomes, I did not find complete Cas systems in archaeal MAGs. Also, apart from R-M and Cas, single-gene defense systems such as NHI, PARIS (which can be easily shuttled by MGEs), or ABI systems that can work as T-A addition modules are favored, in detriment of larger defense systems such as BREX or DISARM. I also observed that such predominance of R-M (and to a large extent Cas) systems was largely independent of the geographical region. However, semi-enclosed oligotrophic seas such as the Mediterranean Sea, the Red sea, and oceans with low latitudinal

biodiversity patterns such as the Southern Ocean seem to depart from the other regions in terms of a lower abundance of defense systems, suggesting little challenge from phage infection (Fig. 4.5b).

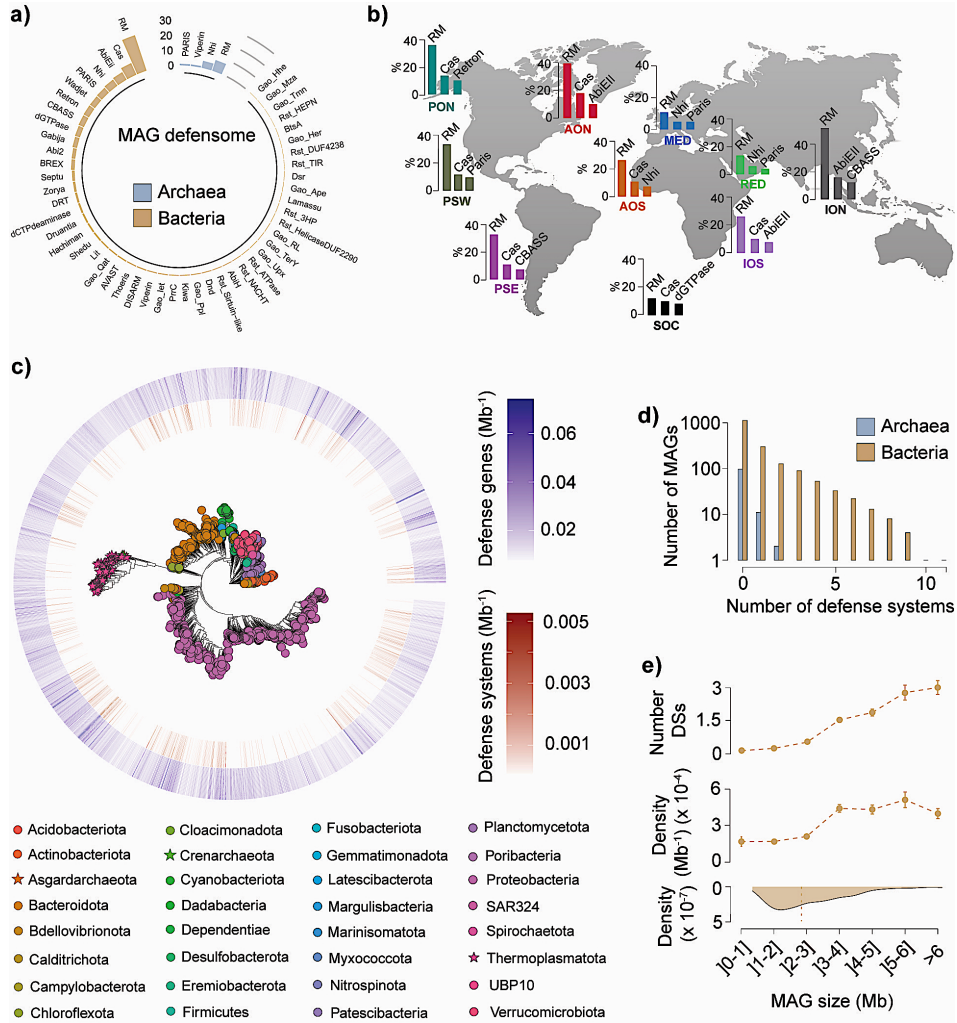


Figure 4.5: Overall view of the defensome of TARA MAGs. (a) Percentage of defense systems found in archaea and bacteria. (b) Percentage distribution of the top three defense systems found in MAGs across major oceans / seas analyzed. (c) Phylogenetic representation of the 1,888 MAGs, their corresponding clade, as well as density (per Mb) of complete defense systems (in red) or defense genes (in purple). (d) Variation of number of defense systems per MAG. (e) Variation of number and density (per MB) of defense systems (DSs) with MAG size (Mb).

The frequency of the defensome also varied widely among prokaryotic phyla with genome size, taxonomy and lifestyle (Fig. 4.5c). On average I found prokaryotic MAGs to encode 1.2 complete defense systems. The number of complete anti-phage mechanisms per genome varied widely from a minimum of zero (1,221 MAGs) to 9 (8

MAGs, belonging to clades typically engaging in high HGT, including Pseudomonadales and Burkholderiaceae) (**Fig. 4.5c,d**). More than 20% of MAGs contained more than 1 defense system. These statistics change completely when we look instead at the abundance of defense genes. In such case, no clade or MAG was entirely devoid of defensome (data not shown). All MAGs averaged in this case 55 defense genes, ranging from a minimum of 10 (typically in obligatory symbionts) to 700 in species engaging in high gene flux exchanges. Larger genomes typically engage more extensively in HGT, and are therefore expected to require a larger and more diversified defensome for protection. This explains the positive correlation between the total number of defense systems and MAG size, which was roughly linear from MAGs sizing 2 Mb onwards. When controlled for the genome size, we see that cells with larger MAGs are actually keeping the density of defense systems per Mb constant (**Fig. 4.5e**). I am currently working to understand why do prokaryotes keep such a large number and variety of incomplete anti-phage defense systems. Such abundance suggests that a larger than expected number of defense-associated genes may be interacting in previously unrecognized ways, or playing additional (non-defense) roles in the cell.

A major trend in the evolution of defense systems is their frequent clustering in distinct genomic regions that have been termed defense islands, by analogy with pathogenicity and symbiotic islands. Many islands combine diverse defense systems, and might also include genes involved in novel mechanisms. This prompted me to evaluate the abundance of defense islands in our dataset. I found a total of 405 defense islands (those containing at least 5 defense systems) with a median size of 15 genes (**Fig. 4.6a**). The most abundant defense systems at a global level (R-Ms and CRISPR-Cas), are also quite abundant in defense islands, but so are BREX and CBASS, two globally underrepresented defense systems (**Fig. 4.6b**). Such observations suggest that certain defense genes may be preferentially carried by MGEs that target chromosomal sinks (the so called defense islands). As rounds of MGE integration/degradation succeed, the remnant defense systems form clusters in the chromosome. The clustering of these systems may facilitate the evolution of functional interactions between them or co/regulation of gene expression. On (**Fig. 4.6c**) I show three examples of defense islands in which we can observe the high variability of defense system families mingled with 6% of other genes not directly related to defense. This prompted me to better look at these 6% of defense island 'dark matter'. Upon inspection of the GO functional classification of these 'non-defense' genes in defense islands, I observed a large variety of mobility-related genes (transposition, prophage genes, etc), which makes sense given that these islands are sinks of MGEs, but also motility-related genes (flagellum, cilium assembly), nitrogen fixation, tRNA methylation, among others (**Fig. 4.6d**). I finally reasoned that the number of anti-phage systems could also be influenced by the diversity of phages a prokaryote might encounter. To do so, I performed an in-depth search of prophages in the MAGs dataset and found 2,454 prophages in 1,888 genomes, with an average of 1.2 prophages per genome (**Fig. 4.6e**). Not more than 7 prophages were detected, and half of MAGs were devoid of prophages, pointing to a scenario where the marine

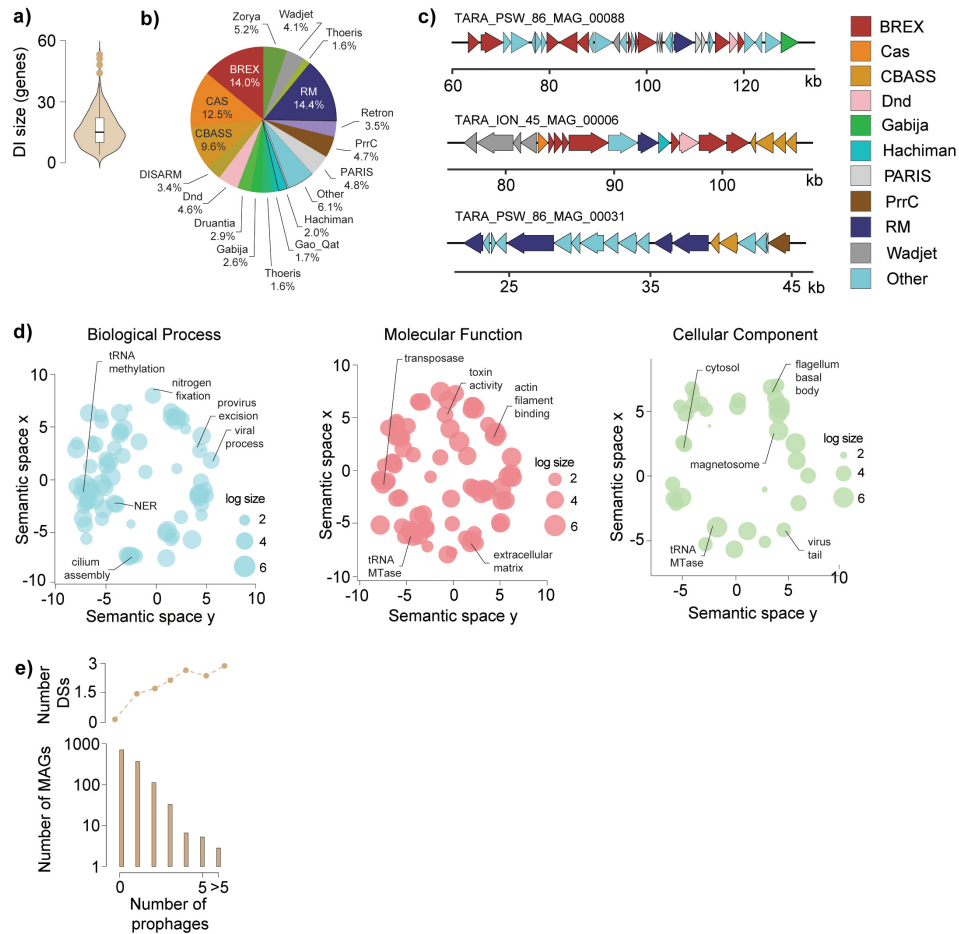


Figure 4.6: Overview of the composition of defense islands (those with at least 5 defense islands) in TARA MAGs. (a) Distribution of defense island size in TARA MAGs. (b) Percentage distribution of the defense systems found in TARA MAGs' defense islands. (c) Example of three defense islands with a rich diversity in defense systems/genes. (d) GO functional analysis of defense island genes not recognized as pertaining to known defense systems. (e) Variation of defense system number with number of prophages per MAG.

environment does not particularly favor these types of MGEs. Still, the number of prophages correlated positively with the number of defense systems, suggesting that the anti-viral arsenal of prokaryotes is also influenced by the number of prophages present in the genome.

4.5 Final remarks

The questions raised in sections 4.2 to 4.4 are timely and outstanding. Yet, there are several other equally exciting research avenues worth pursuing in the future (either by us or others) (Fig. 4.7). One of these questions concerns the metaepigenomic analyses of bacterial communities from different ecological niches. The latter will significantly deepen our understanding on the evolution of methylation systems, on the impacts of DNA methylation in shaping the composition of such niches, and in the understanding of the coevolutionary history of methylation systems and host genome.

In this regard, there is also increasing interest in studying the holoeptigenome, which by definition implies an epigenetic interaction between the host and its symbionts (the holobiont). Such interactions can affect key biological processes of both host and microorganisms and have the power to shape their coevolution. For example, dysbiosis and reduction of microbial diversity can change the proportion of metabolites acting as regulators of DNA and histone modifications in the host. Alternatively, the secretion or injection/translocation of nucleus-targeted effectors—termed nucleomodulins—from a bacterial pathogen into the host cytosol can subvert the host epigenome through interference with histone and DNA modifications, regulation of transcription, interference on the cell cycle, and regulation of cell signaling pathways. For example, the nucleomodulins Mhy from *Mycoplasma hyorhinitis* and Rv2966c from *Mycobacterium tuberculosis* are capable of acting as mammalian DNA MTases and regulate proliferation-specific pathways. Hence, it is foreseeable that the next years will bring additional research on nucleomodulin diversity in bacterial pathogens and a better understanding of the mechanisms used for nuclear trafficking and modulation of the host genome.

Another interesting research avenue is the one dealing with genetic assimilation. The latter essentially assumes that a stress-induced non-genetic change in phenotype can, during the course of selection and over multiple generations, become genetically encoded. This necessarily raises a few outstanding questions: (i) is this genetic assimilation aimed at maintaining stress-related epigenetic landscapes? (ii) are the observable changes in gene expression directly modulated by the acquisition of a particular subset of DNA methylation marks? Although recent studies have begun to provide insight into this topic, we will need to wait for further advances in long-read technologies applied to single-cell sequencing, in order to identify the missing pieces of what appears to be a complex puzzle of epigenetic-mediated persistence.

Beyond these promising routes of investigation, there is the important aspect concerning the harmonization of the research lines proposed in 4.2 to 4.4 with the

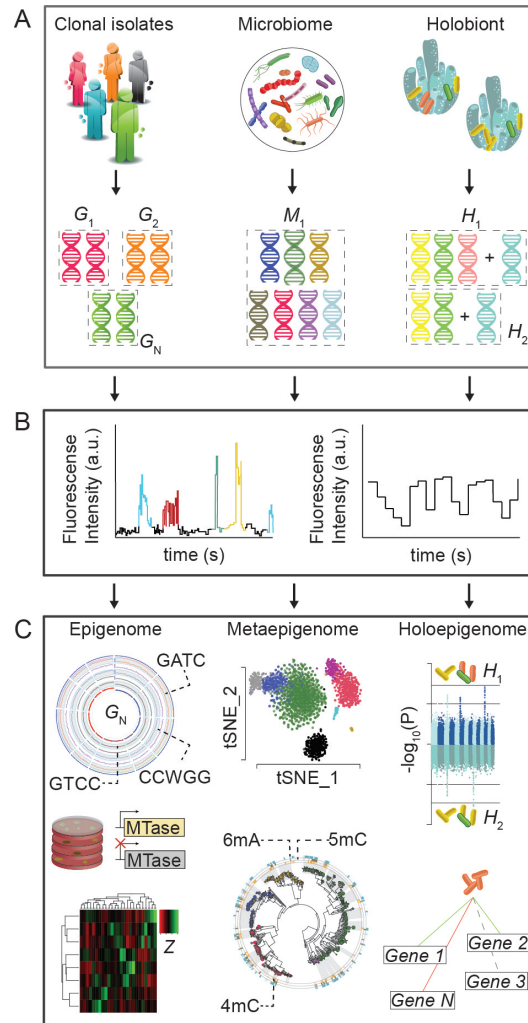


Figure 4.7: Approaches to study bacterial methylomes from clonal isolates, microbiomes, and holobionts. (A) Although the large abundance of methylomes profiled to date belong to genome (G) isolates, there is a growing interest in the analysis of microbiome (M) and holobiont (H) methylomes. (B) Recent progresses in “third-generation” sequencing technologies (e.g.: SMRT and nanopore sequencing) have enabled direct genome-wide detection of methylated positions and target motifs. (C) Relevant functional information on the epigenome can be obtained by targeted mutagenesis of DNA MTases. A comprehensive global transcriptome and functional profiling by RNA-seq offers the opportunity to further dissect the range of differentially expressed genes in a methylation-free strain. For metaepigenomes, nonlinear dimensionality reduction algorithms such as the t-distributed stochastic neighbor embedding (t-SNE), are a possible option to visualize and interpret methylation features across multiple metagenomic contigs. A phylogenetic representation of methylation systems’ density across several metagenome assembled genomes may also provide clues into the interplay between DNA methylation and factors unique to the environment of each community. In holoeigenomes, genome-wide analysis of CpG site methylation differences between multiple hosts (as shown in the Manhattan plots of P-values), may provide insight into the network of host genes whose expression is being significantly modulated by the presence of certain symbionts. Taken from [Oliveira 2021].

activity of SeqLab as sequencing platform, and my roles as lab head. Since its creation 25 years ago, the Genoscope – French National Sequencing Center has provided the scientific community with all the (meta)-omics expertise and data production capacities necessary to leverage multiple collaborative projects with high scientific impact. In particular, SeqLab houses an expanding portfolio of state-of-the-art sequencing equipment, including the Illumina NovaSeq 6000, the Oxford Nanopore Promethion, and the MGI DNBSEQ-600. We are also equipped with a suite of automated liquid-handling systems along with optical mapping devices, which allow for improved process streamlining and parallelization. Despite disposing of world-class facilities and bringing together a dedicated team of highly skilled technicians, SeqLab would greatly benefit from a more intense focus on providing value through research and development. Some examples include the identification of significant opportunities to improve processes across genomic pipelines, the development of novel sequencing strategies, and of new methods and procedures to maximize the efficiency, quality and throughput of all incoming machinery. In parallel, there is, in my opinion, great interest in developing lines of high-quality fundamental research that build upon SeqLab’s unique infrastructure and its surrounding ecosystem (e.g.: UMR 8030). This would allow the SeqLab to increase its international visibility, attract more funding, students/researchers and recognition. The research lines proposed in Chapter IV, are not only *i*) in line with key activities of the UMR 8030, regarding the interest in leveraging data collected from flagship projects such as that of TARA; but also represent *ii*) a first step that builds upon recent developments in third generation sequencing techniques and on my previously acquired know-how, to strategically implement a solid and long-term research activity in Microbial Epigenomics.

After my recruitment at Genoscope in September 2020, I immediately put in place a set of initiatives with the purpose of stimulating research activity at SeqLab. For example:

- I started providing basic formation in bioinformatics (handling and QC of sequencing data, genome assembly, variant calling, etc) to a few technicians;
- I started providing formation in cell culture and standard molecular biology techniques to a few technicians;
- I published an Opinion paper on my personal view of the future of the bacterial epigenomics field [Oliveira 2021] and edited a Springer Nature Methods in Molecular Biology book on *Computational Epigenomics end Epitranscriptomics* to be published in 2022;
- I started applying for multiples sources of funding to help kick-starting the research activities at SeqLab (e.g.: ANR, Genopole, H2020, etc);
- I started collaborating with key specialists in the fields of microbial genetics and epidemiology (see Chapter 2), whose research interests greatly complement ours;

- I oriented 4 master thesis (2 in 2021) and 2 in 2022 (ongoing);
- Research work from these students was selected for multiple oral communications in renowned national and international congresses;

Finally, I engaged in the writing of this HDR document. Not seeking personal recognition I must say. But having as main purpose that of allowing SeqLab to, for the first time, independently attract Ph.D. students and provide them with all the resources needed to develop meaningful and high-impact research. My personal long-term vision of SeqLab is not of a sequencing platform. But instead of a dynamic and internationally competitive R&D Genomics laboratory, with strong inner ties with the UMR 8030, but with a forward looking view towards addressing several of the biggest challenges in the field. Of course, looking back at predictions made just 10 years ago in the field of genomics, one should expect many additional unforeseen bottlenecks and setbacks that are just as difficult to predict now as they were back then. This is the reason why, with a profound sense of humility and responsibility, and grounded in a spirit of trust and collaboration, I challenge the SeqLab to step into the discomfort and dare to dream something greater.

Bibliography

- [Achaz *et al.* 2007] G. Achaz, F. Boyer, E. P. Rocha, A. Viari and E. Coissac. *Repeek, a tool to retrieve approximate repeats from large DNA sequences*. *Bioinformatics*, vol. 23, no. 1, pages 119–121, 2007. (Cited on page 66.)
- [Almeida *et al.* 2019] A. Almeida, A. L. Mitchell, M. Boland, S. C. Forster, G. B. Gloor, A. Tarkowska, T. D. Lawley and R. D. Finn. *A new genomic blueprint of the human gut microbiota*. *Nature*, vol. 568, no. 7753, pages 499–504, 2019. (Cited on pages 70 and 71.)
- [Amin *et al.* 2013] M. A. Amin, G. E. Richmond, P. Sen, T. H. Koh, L. J. V. Piddock and K. L. Chua. *A method for generating marker-less gene deletions in multidrug-resistant Acinetobacter baumannii*. *BMC Microbiol.*, vol. 13, no. 158, 2013. (Cited on page 64.)
- [Arber & Linn 1969] W. Arber and S. Linn. *DNA modification and restriction*. *Annu. Rev. Biochem.*, vol. 38, pages 467–500, 1969. (Cited on page 36.)
- [Ardissone *et al.* 2016] S. Ardisson, P. Redder, G. Russo, A. Frandi, C. Fumeaux, A. Patrignani, R. Schlapbach, L. Falquet and P. H. Viollier. *Cell cycle constraints and environmental control of local DNA hypomethylation in γ -Proteobacteria*. *PLoS Genet.*, vol. 12, no. 12, page e1006499, 2016. (Cited on page 54.)
- [Arndt *et al.* 2016] D. Arndt, J. R. Grant, A. Marcu, T. Sajed, A. Pon, Y. Liang and D. S. Wishart. *PHASTER: a better, faster version of the PHAST phage search tool*. *Nucleic Acids Res*, vol. 44, no. W1, pages 16–21, 2016. (Cited on page 68.)
- [Atack *et al.* 2015] J. M. Atack, Y. N. Srikhanta, K. L. Fox, J. A. Jurcisek, K. L. Brockman, T. A. Clark, M. Boitano, P. M. Power, F. E. Jen, A. G. McEwan, S. M. Grimmond, A. L. Smith, S. J. Barenkamp, J. Korlach, L. O. Bakaletz and M. P. Jennings. *A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable Haemophilus influenzae*. *Nat. Commun.*, vol. 6, page 7828, 2015. (Cited on page 50.)
- [Atack *et al.* 2018] J. M. Atack, Y. Yang, K. L. Seib, Y. Zhou and M. P. Jennings. *A survey of Type III restriction-modification systems reveals numerous, novel epigenetic regulators controlling phase-variable regulons; phasevarions*. *Nucleic Acids Res*, vol. 46, no. 7, pages 3532–3542, 2018. (Cited on page 67.)
- [Atack *et al.* 2020] J. M. Atack, C. Guo, T. Litfin, L. Yang, P. J. Blackall, Y. Zhou and M. P. Jennings. *Specificity Genes That Can Switch System Specificity by Recombination*. *mSystems*, vol. 5, no. 4, 2020. (Cited on page 67.)

- [Azzoni *et al.* 2007] A. R. Azzoni, S. C. Ribeiro, G. A. Monteiro and D. M. Prazeres. *The impact of polyadenylation signals on plasmid nuclease-resistance and transgene expression*. J Gene Med, vol. 9, no. 5, pages 392–402, 2007. (Cited on page 18.)
- [Bahloul *et al.* 1998] C. Bahloul, Y. Jacob, N. Tordo and P. Perrin. *DNA-based immunization for exploring the enlargement of immunological cross-reactivity against the lyssaviruses*. Vaccine, vol. 16, no. 4, pages 417–425, 1998. (Cited on page 20.)
- [Bai *et al.* 2007] R. K. Bai, S. M. Leal, D. Covarrubias, A. Liu and L. J. Wong. *Mitochondrial genetic background modifies breast cancer risk*. Cancer Res., vol. 67, no. 10, pages 4687–4694, 2007. (Cited on page 32.)
- [Balbontin *et al.* 2008] R. Balbontin, N. Figueroa-Bossi, J. Casadesus and L. Bossi. *Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in Salmonella enterica*. J. Bacteriol., vol. 190, no. 11, pages 4075–4078, 2008. (Cited on page 46.)
- [Baxter *et al.* 2004] M. A. Baxter, R. F. Wynn, S. N. Jowitt, J. E. Wraith, L. J. Fairbairn and I. Bellantuono. *Study of telomere length reveals rapid aging of human marrow stromal cells following in vitro expansion*. Stem Cells, vol. 22, no. 5, pages 675–682, 2004. (Cited on page 27.)
- [Beauchamp *et al.* 2017] J. M. Beauchamp, R. M. Leveque, S. Dawid and V. J. DiRita. *Methylation-dependent DNA discrimination in natural transformation of Campylobacter jejuni*. Proc. Natl. Acad. Sci. U.S.A., vol. 114, no. 38, pages E8053–E8061, 2017. (Cited on pages 63 and 64.)
- [Bedrat *et al.* 2016] A. Bedrat, L. Lacroix and J. L. Mergny. *Re-evaluation of G-quadruplex propensity with G4Hunter*. Nucleic Acids Res., vol. 44, no. 4, pages 1746–1759, 2016. (Cited on page 66.)
- [Bi & Liu 1996] X. Bi and L. F. Liu. *A replicational model for DNA recombination between direct repeats*. J. Mol. Biol., vol. 256, no. 5, pages 849–858, 1996. (Cited on page 20.)
- [Bland *et al.* 2007] C. Bland, T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides and P. Hugenholtz. *CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats*. BMC Bioinformatics, vol. 8, page 209, 2007. (Cited on page 70.)
- [Bonab *et al.* 2006] M. M. Bonab, K. Alimoghaddam, F. Talebian, S. H. Ghaffari, A. Ghavamzadeh and B. Nikbin. *Aging of mesenchymal stem cell in vitro*. BMC Cell Biol., vol. 7, page 14, 2006. (Cited on page 26.)

- [Boyd *et al.* 2009] E. F. Boyd, S. Almagro-Moreno and M. A. Parent. *Genomic islands are dynamic, ancient integrative elements in bacterial evolution*. Trends Microbiol., vol. 17, no. 2, pages 47–53, 2009. (Cited on page 46.)
- [Budroni *et al.* 2011] S. Budroni, E. Siena, J. C. Dunning Hotopp, K. L. Seib, D. Serruto, C. Nofroni, M. Comanducci, D. R. Riley, S. C. Daugherty, S. V. Angiuoli, A. Covacci, M. Pizza, R. Rappuoli, E. R. Moxon, H. Tettelin and D. Medini. *Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination*. Proc. Natl. Acad. Sci. U.S.A., vol. 108, no. 11, pages 4494–4499, 2011. (Cited on pages 41 and 43.)
- [Bunnell & Morgan 1998] B. A. Bunnell and R. A. Morgan. *Gene therapy for infectious diseases*. Clin. Microbiol. Rev., vol. 11, no. 1, pages 42–56, 1998. (Cited on page 18.)
- [Burrus *et al.* 2001] V. Burrus, C. Bontemps, B. Decaris and G. Guedon. *Characterization of a novel type II restriction-modification system, Sth368I, encoded by the integrative element ICESt1 of Streptococcus thermophilus CNRZ368*. Appl. Environ. Microbiol., vol. 67, no. 4, pages 1522–1528, 2001. (Cited on page 36.)
- [Buske *et al.* 2012] F. A. Buske, D. C. Bauer, J. S. Mattick and T. L. Bailey. *Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic data*. Genome Res., vol. 22, no. 7, pages 1372–1381, 2012. (Cited on page 66.)
- [Caplan 2007] A. I. Caplan. *Adult mesenchymal stem cells for tissue engineering versus regenerative medicine*. J. Cell. Physiol., vol. 213, no. 2, pages 341–347, 2007. (Cited on page 28.)
- [Casadesus & Low 2006] J. Casadesus and D. Low. *Epigenetic gene regulation in the bacterial world*. Microbiol. Mol. Biol. Rev., vol. 70, no. 3, pages 830–86, 2006. (Cited on page 50.)
- [Casadesús & Low 2013] J. Casadesús and D. A. Low. *Programmed heterogeneity: epigenetic mechanisms in bacteria*. J Biol Chem, vol. 288, no. 20, pages 13929–13935, 2013. (Cited on page 66.)
- [Chancey *et al.* 2015] S. T. Chancey, S. Agrawal, M. R. Schroeder, M. M. Farley, H. Tettelin and D. S. Stephens. *Composite mobile genetic elements disseminating macrolide resistance in Streptococcus pneumoniae*. Front Microbiol, vol. 6, page 26, 2015. (Cited on page 46.)
- [Chedin *et al.* 1994] F. Chedin, E. Dervyn, R. Dervyn, S. D. Ehrlich and P. Noirot. *Frequency of deletion formation decreases exponentially with distance between short direct repeats*. Mol. Microbiol., vol. 12, no. 4, pages 561–569, 1994. (Cited on page 23.)

- [Chen *et al.* 2004] S. L. Chen, W. W. Fang, F. Ye, Y. H. Liu, J. Qian, S. J. Shan, J. J. Zhang, R. Z. Chunhua, L. M. Liao, S. Lin and J. P. Sun. *Effect on left ventricular function of intracoronary transplantation of autologous bone marrow mesenchymal stem cell in patients with acute myocardial infarction.* Am. J. Cardiol., vol. 94, no. 1, pages 92–95, 2004. (Cited on page 28.)
- [Chiu *et al.* 2016] T. P. Chiu, F. Comoglio, T. Zhou, L. Yang, R. Paro and R. Rohs. *DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding.* Bioinformatics, vol. 32, no. 8, pages 1211–1213, 2016. (Cited on page 66.)
- [Coban *et al.* 2005] C. Coban, K. J. Ishii, M. Gursel, D. M. Klinman and N. Kumar. *Effect of plasmid backbone modification by different human CpG motifs on the immunogenicity of DNA vaccine vectors.* J. Leukoc. Biol., vol. 78, no. 3, pages 647–655, 2005. (Cited on page 18.)
- [Coene *et al.* 2011] E. D. Coene, C. Gadelha, N. White, A. Malhas, B. Thomas, M. Shaw and D. J. Vaux. *A novel role for BRCA1 in regulating breast cancer cell spreading and motility.* J. Cell Biol., vol. 192, no. 3, pages 497–512, 2011. (Cited on page 28.)
- [Cohen *et al.* 2016] N. R. Cohen, C. A. Ross, S. Jain, R. S. Shapiro, A. Gutierrez, P. Belenky, H. Li and J. J. Collins. *A role for the bacterial GATC methylome in antibiotic stress survival.* Nat. Genet., vol. 48, no. 5, pages 581–6, 2016. (Cited on page 50.)
- [Collins *et al.* 2018] J. Collins, C. Robinson, H. Danhof, C. W. Knetsch, H. C. van Leeuwen, T. D. Lawley, J. M. Auchtung and R. A. Britton. *Dietary trehalose enhances virulence of epidemic Clostridium difficile.* Nature, vol. 553, no. 7688, pages 291–294, 2018. (Cited on page 50.)
- [Cota *et al.* 2016] I. Cota, B. Bunk, C. Sproer, J. Overmann, C. Konig and J. Casadesus. *OxyR-dependent formation of DNA methylation patterns in OpvABOFF and OpvABON cell lineages of Salmonella enterica.* Nucleic Acids Res., vol. 44, no. 8, pages 3595–3609, 2016. (Cited on page 54.)
- [Croucher *et al.* 2011] N. J. Croucher, S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, B. Pichon, S. Baker, C. M. Parry, L. M. Lambertsen, D. Shahinas, D. R. Pillai, T. J. Mitchell, G. Dougan, A. Tomasz, K. P. Klugman, J. Parkhill, W. P. Hanage and S. D. Bentley. *Rapid pneumococcal evolution in response to clinical interventions.* Science, vol. 331, no. 6016, pages 430–434, 2011. (Cited on page 46.)
- [Croucher *et al.* 2014] N. J. Croucher, P. G. Coupland, A. E. Stevenson, A. Callendrello, S. D. Bentley and W. P. Hanage. *Diversification of bacterial genome*

- content through distinct mechanisms over different timescales.* Nat Commun, vol. 5, page 5471, 2014. (Cited on pages 41 and 43.)
- [Csuros 2010] M. Csuros. *Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood.* Bioinformatics, vol. 26, no. 15, pages 1910–1912, 2010. (Cited on page 42.)
- [Cury *et al.* 2016] J. Cury, T. Jové, M. Touchon, B. Néron and E. P. Rocha. *Identification and analysis of integrons and cassette arrays in bacterial genomes.* Nucleic Acids Res, vol. 44, no. 10, pages 4539–4550, 2016. (Cited on page 68.)
- [Cury *et al.* 2017] J. Cury, M. Touchon and E. P. C. Rocha. *Integrative and conjugative elements and their hosts: composition, distribution and organization.* Nucleic Acids Res, vol. 45, no. 15, pages 8943–8956, 2017. (Cited on page 68.)
- [Cyranoski 2010] D. Cyranoski. *Strange lesions after stem-cell therapy.* Nature, vol. 465, no. 7301, page 997, 2010. (Cited on page 27.)
- [Damas *et al.* 2012] J. Damas, J. Carneiro, J. Goncalves, J. B. Stewart, D. C. Samuels, A. Amorim and F. Pereira. *Mitochondrial DNA deletions are associated with non-B DNA conformations.* Nucleic Acids Res., vol. 40, no. 16, pages 7606–7621, Sep 2012. (Cited on page 33.)
- [Davis *et al.* 2013] B. M. Davis, M. C. Chao and M. K. Waldor. *Entering the era of bacterial epigenomics with single molecule real time DNA sequencing.* Curr. Opin. Microbiol., vol. 16, no. 2, pages 192–8, 2013. (Cited on page 50.)
- [Delmont *et al.* 2021] T. O. Delmont, J. J. Pierella Karlusich, I. Veseli, J. Fuessel, A. M. Eren, R. A. Foster, C. Bowler, P. Wincker and E. Pelletier. *Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean.* ISME J, 2021. (Cited on pages 70 and 71.)
- [Didelot *et al.* 2011] X. Didelot, R. Bowden, T. Street, T. Golubchik, C. Spencer, G. McVean, V. Sangal, M. F. Anjum, M. Achtman, D. Falush and P. Donnelly. *Recombination and population structure in Salmonella enterica.* PLoS Genet., vol. 7, no. 7, page e1002191, 2011. (Cited on page 41.)
- [Dimarino *et al.* 2013] A. M. Dimarino, A. I. Caplan and T. L. Bonfield. *Mesenchymal stem cells in tissue repair.* Front Immunol, vol. 4, page 201, 2013. (Cited on page 26.)
- [Doroghazi & Buckley 2010] J. R. Doroghazi and D. H. Buckley. *Widespread homologous recombination within and between Streptomyces species.* ISME J., vol. 4, no. 9, pages 1136–1143, 2010. (Cited on page 41.)

- [Doron *et al.* 2018] S. Doron, S. Melamed, G. Ofir, A. Leavitt, A. Lopatina, M. Keren, G. Amitai and R. Sorek. *Systematic discovery of antiphage defense systems in the microbial pangenome*. *Science*, vol. 359, no. 6379, 2018. (Cited on pages 63, 69 and 72.)
- [Dos Santos *et al.* 2010] F. Dos Santos, P. Z. Andrade, J. S. Boura, M. M. Abecasis, C. L. da Silva and J. M. Cabral. *Ex vivo expansion of human mesenchymal stem cells: a more effective cell proliferation kinetics and metabolism under hypoxia*. *J. Cell. Physiol.*, vol. 223, no. 1, pages 27–35, 2010. (Cited on page 28.)
- [Dupuis *et al.* 2013] M. E. Dupuis, M. Villion, A. H. Magadan and S. Moineau. *CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance*. *Nat Commun*, vol. 4, page 2087, 2013. (Cited on pages 36 and 37.)
- [Dy *et al.* 2014] R. L. Dy, R. Przybilski, K. Semeijn, G. P. Salmond and P. C. Fineran. *A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism*. *Nucleic Acids Res*, vol. 42, no. 7, pages 4590–4605, 2014. (Cited on page 69.)
- [Eckert & Yan 2000] K. A. Eckert and G. Yan. *Mutational analyses of dinucleotide and tetranucleotide microsatellites in Escherichia coli: influence of sequence on expansion mutagenesis*. *Nucleic Acids Res.*, vol. 28, no. 14, pages 2831–2838, 2000. (Cited on page 23.)
- [Efimenko *et al.* 2011] A. Efimenko, E. Starostina, N. Kalinina and A. Stolzing. *Angiogenic properties of aged adipose derived mesenchymal stem cells after hypoxic conditioning*. *J Transl Med*, vol. 9, page 10, 2011. (Cited on page 29.)
- [EMA 2011] EMA. *Reflection paper on stem cell-based medicinal products. Committee for Advanced Therapies (CAT)-European Medicines Agency 2011*. 2011. (Cited on page 27.)
- [Ershova *et al.* 2012] A. S. Ershova, A. S. Karyagina, M. O. Vasiliev, A. M. Lyashchuk, V. G. Lunin, S. A. Spirin and A. V. Alexeevski. *Solitary restriction endonucleases in prokaryotic genomes*. *Nucleic Acids Res.*, vol. 40, no. 20, pages 10107–10115, 2012. (Cited on page 40.)
- [Estibariz *et al.* 2018] Iratxe Estibariz, Annemarie Overmann, Florent Ailloud, Juliane Krebs, Christine Josenhans and Sebastian Suerbaum. *The core genome m5C methyltransferase JHP1050 (M.Hpy99III) plays an important role in orchestrating gene expression in Helicobacter pylori*. *bioRxiv*, 2018. (Cited on page 61.)
- [Estrada *et al.* 2012] J. C. Estrada, C. Albo, A. Benguria, A. Dopazo, P. Lopez-Romero, L. Carrera-Quintanar, E. Roche, E. P. Clemente, J. A. Enriquez,

- A. Bernad and E. Samper. *Culture of human mesenchymal stem cells at low oxygen tension improves growth and genetic stability by activating glycolysis*. Cell Death Differ., vol. 19, no. 5, pages 743–755, 2012. (Cited on page 27.)
- [Everitt *et al.* 2014] R. G. Everitt, X. Didelot, E. M. Batty, R. R. Miller, K. Knox, B. C. Young, R. Bowden, A. Auton, A. Votintseva, H. Larner-Svensson, J. Charlesworth, T. Golubchik, C. L. Ip, H. Godwin, R. Fung, T. E. Peto, A. S. Walker, D. W. Crook and D. J. Wilson. *Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus*. Nat Commun, vol. 5, page 3956, 2014. (Cited on page 46.)
- [Fang *et al.* 2012] G. Fang, D. Munera, D. I. Friedman, A. Mandlik, M. C. Chao, O. Banerjee, Z. Feng, B. Losic, M. C. Mahajan, O. J. Jabado, G. Deikus, T. A. Clark, K. Luong, I. A. Murray, B. M. Davis, A. Keren-Paz, A. Chess, R. J. Roberts, J. Korlach, S. W. Turner, V. Kumar, M. K. Waldor and E. E. Schadt. *Genome-wide mapping of methylated adenine residues in pathogenic Escherichia coli using single-molecule real-time sequencing*. Nat. Biotechnol., vol. 30, no. 12, pages 1232–9, 2012. (Cited on page 50.)
- [Fletcher *et al.* 2018] J. R. Fletcher, S. Erwin, C. Lanzas and C. M. Theriot. *Shifts in the gut metabolome and Clostridium difficile transcriptome throughout colonization and infection in a mouse model*. mSphere, vol. 3, no. 2, 2018. (Cited on page 56.)
- [Flusberg *et al.* 2010] B. A. Flusberg, D. R. Webster, J. H. Lee, K. J. Travers, E. C. Olivares, T. A. Clark, J. Korlach and S. W. Turner. *Direct detection of DNA methylation during single-molecule, real-time sequencing*. Nat. Methods, vol. 7, no. 6, pages 461–465, 2010. (Cited on page 50.)
- [Fouts 2006] D. E. Fouts. *PhageFinder: automated identification and classification of prophage regions in complete bacterial genome sequences*. Nucleic Acids Res., vol. 34, no. 20, pages 5839–5851, 2006. (Cited on page 68.)
- [Fraser *et al.* 2007] C. Fraser, W. P. Hanage and B. G. Spratt. *Recombination and the nature of bacterial speciation*. Science, vol. 315, no. 5811, pages 476–480, 2007. (Cited on page 41.)
- [Fukuda *et al.* 2008] E. Fukuda, K. H. Kaminska, J. M. Bujnicki and I. Kobayashi. *Cell death upon epigenetic genome methylation: a novel function of methyl-specific deoxyribonucleases*. Genome Biol., vol. 9, no. 11, page R163, 2008. (Cited on pages 36 and 37.)
- [Furuta *et al.* 2014] Y. Furuta, H. Namba-Fukuyo, T. F. Shibata, T. Nishiyama, S. Shigenobu, Y. Suzuki, S. Sugano, M. Hasebe and I. Kobayashi. *Methylome diversification through changes in DNA methyltransferase sequence specificity*. PLoS Genet., vol. 10, no. 4, page e1004272, 2014. (Cited on page 36.)

- [Gallagher *et al.* 2014] R. R. Gallagher, O. A. Lewis and F. J. Isaacs. *Rapid editing and evolution of bacterial genomes using libraries of synthetic DNA*. Nat. Protoc., vol. 9, pages 2301–2316, 2014. (Cited on page 64.)
- [Gao *et al.* 2016] F. Gao, S. M. Chiu, D. A. Motan, Z. Zhang, L. Chen, H. L. Ji, H. F. Tse, Q. L. Fu and Q. Lian. *Mesenchymal stem cells and immunomodulation: current status and future prospects*. Cell Death Dis, vol. 7, page e2062, 2016. (Cited on page 26.)
- [Gimble & Guilak 2003] J. Gimble and F. Guilak. *Adipose-derived adult stem cells: isolation, characterization, and differentiation potential*. Cytotherapy, vol. 5, no. 5, pages 362–369, 2003. (Cited on page 28.)
- [Gogarten *et al.* 2002] J. P. Gogarten, W. F. Doolittle and J. G. Lawrence. *Prokaryotic evolution in light of gene transfer*. Mol. Biol. Evol., vol. 19, no. 12, pages 2226–2238, 2002. (Cited on page 35.)
- [Goldfarb *et al.* 2015] T. Goldfarb, H. Sberro, E. Weinstock, O. Cohen, S. Doron, Y. Charpak-Amikam, S. Afik, G. Ofir and R. Sorek. *BREX is a novel phage resistance system widespread in microbial genomes*. EMBO J, vol. 34, no. 2, pages 169–183, 2015. (Cited on pages 69 and 72.)
- [Gonçalves *et al.* 2014] G. A. L. Gonçalves, P. H. Oliveira, A. G. Gomes, K. L. J. Prather, L. A. Lewis, D. M. F. Prazeres and G. A. Monteiro. *Evidence that the insertion events of IS2 transposition are biased towards abrupt compositional shifts in target DNA and modulated by a diverse set of culture parameters*. Appl. Microbiol. Biotechnol., vol. 98, no. 15, pages 6609–6619, 2014. (Cited on pages 11, 24 and 25.)
- [Gore *et al.* 2006] J. M. Gore, F. A. Ran and L. N. Ornston. *Deletion mutations caused by DNA strand slippage in Acinetobacter baylyi*. Appl. Environ. Microbiol., vol. 72, no. 8, pages 5239–5245, 2006. (Cited on page 23.)
- [Guynet *et al.* 2009] C. Guynet, A. Achard, B. T. Hoang, O. Barabas, A. B. Hickman, F. Dyda and M. Chandler. *Resetting the site: redirecting integration of an insertion sequence in a predictable way*. Mol. Cell, vol. 34, no. 5, pages 612–619, 2009. (Cited on page 26.)
- [Haft *et al.* 2003] D. H. Haft, J. D. Selengut and O. White. *The TIGRFAMs database of protein families*. Nucleic Acids Res, vol. 31, no. 1, pages 371–373, 2003. (Cited on page 70.)
- [He *et al.* 2013] M. He, F. Miyajima, P. Roberts, L. Ellison, D. J. Pickard, M. J. Martin, T. R. Connor, S. R. Harris, D. Fairley, K. B. Bamford, S. D’Arc, J. Brazier, D. Brown, J. E. Coia, G. Douce, D. Gerding, H. J. Kim, T. H. Koh, H. Kato, M. Senoh, T. Louie, S. Michell, E. Butt, S. J. Peacock, N. M. Brown, T. Riley, G. Songer, M. Wilcox, M. Pirmohamed, E. Kuijper, P. Hawkey,

- B. W. Wren, G. Dougan, J. Parkhill and T. D. Lawley. *Emergence and global spread of epidemic healthcare-associated Clostridium difficile*. Nat. Genet., vol. 45, no. 1, pages 109–113, 2013. (Cited on page 50.)
- [Ho *et al.* 1986] P. S. Ho, M. J. Ellison, G. J. Quigley, A. Rich and A. Rich. *A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences*. EMBO J., vol. 5, no. 10, pages 2737–2744, 1986. (Cited on page 66.)
- [Holzwarth *et al.* 2010] C. Holzwarth, M. Vaegler, F. Gieseke, S. M. Pfister, R. Handgretinger, G. Kerst and I. Muller. *Low physiologic oxygen tensions reduce proliferation and differentiation of human multipotent mesenchymal stromal cells*. BMC Cell Biol., vol. 11, page 11, 2010. (Cited on page 27.)
- [Horwitz *et al.* 2001] E. M. Horwitz, D. J. Prockop, P. L. Gordon, W. W. Koo, L. A. Fitzpatrick, M. D. Neel, M. E. McCarville, P. J. Orchard, R. E. Pyeritz and M. K. Brenner. *Clinical responses to bone marrow transplantation in children with severe osteogenesis imperfecta*. Blood, vol. 97, no. 5, pages 1227–1231, 2001. (Cited on page 28.)
- [Ingman & Gyllensten 2006] M. Ingman and U. Gyllensten. *mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences*. Nucleic Acids Res., vol. 34, no. Database issue, pages D749–751, 2006. (Cited on page 32.)
- [Initiative 2009] International Stem Cell Banking Initiative. *Consensus guidance for banking and supply of human embryonic stem cell lines for research purposes*. Stem Cell Rev., vol. 5, pages 301–314, 2009. (Cited on page 27.)
- [Ishikawa *et al.* 2009] K. Ishikawa, N. Handa and I. Kobayashi. *Cleavage of a model DNA replication fork by a Type I restriction endonuclease*. Nucleic Acids Res., vol. 37, no. 11, pages 3531–3544, 2009. (Cited on page 36.)
- [Jack *et al.* 2015] B. R. Jack, S. P. Leonard, D. M. Mishler, B. A. Renda, D. Leon, G. A. Suárez and J. E. Barrick. *Predicting the Genetic Stability of Engineered DNA Sequences with the EFM Calculator*. ACS Synth Biol, vol. 4, no. 8, pages 939–943, 2015. (Cited on page 69.)
- [Jallet *et al.* 1999] C. Jallet, Y. Jacob, C. Bahloul, A. Drings, E. Desmezieres, N. Tordo and P. Perrin. *Chimeric lyssavirus glycoproteins with increased immunological potential*. J. Virol., vol. 73, no. 1, pages 225–233, 1999. (Cited on page 20.)
- [Jenjaroenpun *et al.* 2015] P. Jenjaroenpun, T. Wongsurawat, S. P. Yenamandra and V. A. Kuznetsov. *QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences*. Nucleic Acids Res., vol. 43, no. W1, pages W527–534, 2015. (Cited on page 66.)

- [Kay 2011] M. A. Kay. *State-of-the-art gene-based therapies: the road ahead*. Nat. Rev. Genet., vol. 12, no. 5, pages 316–328, 2011. (Cited on page 18.)
- [Kim *et al.* 2009] J. Kim, J. W. Kang, J. H. Park, Y. Choi, K. S. Choi, K. D. Park, D. H. Baek, S. K. Seong, H. K. Min and H. S. Kim. *Biological characterization of long-term cultured human mesenchymal stem cells*. Arch. Pharm. Res., vol. 32, no. 1, pages 117–126, 2009. (Cited on page 29.)
- [Kita *et al.* 2003] K. Kita, H. Kawakami and H. Tanaka. *Evidence for horizontal transfer of the EcoT38I restriction-modification gene to chromosomal DNA by the P2 phage and diversity of defective P2 prophages in Escherichia coli TH38 strains*. J. Bacteriol., vol. 185, no. 7, pages 2296–2305, 2003. (Cited on page 36.)
- [Kobayashi *et al.* 1999] I. Kobayashi, A. Nobusato, N. Kobayashi-Takahashi and I. Uchiyama. *Shaping the genome–restriction-modification systems as mobile genetic elements*. Curr. Opin. Genet. Dev., vol. 9, no. 6, pages 649–656, 1999. (Cited on page 36.)
- [Koonin *et al.* 2017] E. V. Koonin, K. S. Makarova and F. Zhang. *Diversity, classification and evolution of CRISPR-Cas systems*. Curr. Opin. Microbiol., vol. 37, pages 67–78, 2017. (Cited on page 69.)
- [Korona *et al.* 1993] R. Korona, B. Korona and B. R. Levin. *Sensitivity of naturally occurring coliphages to Type I and Type II restriction and modification*. J. Gen. Microbiol., vol. 139 Pt 6, pages 1283–1290, 1993. (Cited on page 43.)
- [Koskiniemi *et al.* 2012] S. Koskiniemi, S. Sun, O. G. Berg and D. I. Andersson. *Selection-driven gene loss in bacteria*. PLoS Genet., vol. 8, no. 6, page e1002787, 2012. (Cited on page 45.)
- [Kouzine *et al.* 2017] F. Kouzine, D. Wojtowicz, L. Baranello, A. Yamane, S. Nelson, W. Resch, K. R. Kieffer-Kwon, C. J. Benham, R. Casellas, T. M. Przytycka and D. Levens. *Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome*. Cell Syst., vol. 4, no. 3, pages 344–356, 03 2017. (Cited on page 66.)
- [Krishnan *et al.* 2008] K. J. Krishnan, A. K. Reeve, D. C. Samuels, P. F. Chinnery, J. K. Blackwood, R. W. Taylor, S. Wanrooij, J. N. Spelbrink, R. N. Lightowlers and D. M. Turnbull. *What causes mitochondrial DNA deletions in human cells?* Nat. Genet., vol. 40, no. 3, pages 275–279, 2008. (Cited on page 33.)
- [Kulakauskas *et al.* 1995] S. Kulakauskas, A. Lubys and S. D. Ehrlich. *DNA restriction-modification systems mediate plasmid maintenance*. J. Bacteriol., vol. 177, no. 12, pages 3451–3454, 1995. (Cited on pages 36 and 37.)

- [Labrie *et al.* 2010] S. J. Labrie, J. E. Samson and S. Moineau. *Bacteriophage resistance mechanisms*. Nat. Rev. Microbiol., vol. 8, no. 5, pages 317–327, 2010. (Cited on page 35.)
- [Lanis *et al.* 2010] J. M. Lanis, S. Barua and J. D. Ballard. *Variations in TcdB activity and the hypervirulence of emerging strains of Clostridium difficile*. PLoS Pathog., vol. 6, no. 8, page e1001061, 2010. (Cited on page 50.)
- [Lazarus *et al.* 2005] H. M. Lazarus, O. N. Koc, S. M. Devine, P. Curtin, R. T. Maziarz, H. K. Holland, E. J. Shpall, P. McCarthy, K. Atkinson, B. W. Cooper, S. L. Gerson, M. J. Laughlin, F. R. Loberiza, A. B. Moseley and A. Bacigalupo. *Cotransplantation of HLA-identical sibling culture-expanded mesenchymal stem cells and hematopoietic stem cells in hematologic malignancy patients*. Biol. Blood Marrow Transplant., vol. 11, no. 5, pages 389–398, 2005. (Cited on page 28.)
- [Lebedeva *et al.* 2009] M. A. Lebedeva, J. S. Eaton and G. S. Shadel. *Loss of p53 causes mitochondrial DNA depletion and altered mitochondrial reactive oxygen species homeostasis*. Biochim. Biophys. Acta, vol. 1787, no. 5, pages 328–334, 2009. (Cited on page 32.)
- [Legras *et al.* 2008] A. Legras, A. Lievre, C. Bonaiti-Pellie, V. Cottet, A. Pariente, B. Nalet, J. Lafon, J. Faivre, C. Bonithon-Kopp, N. Goasguen, C. Penna, S. Olschwang, P. Laurent-Puig, P. Berthelemy, P. Cassan, M. Glikmanas, G. Gatineau-Sailliant, A. Courrier, D. Pillon, J. P. Michalet, J. P. Latrive, J. Guillan, A. Blanchi, B. Bour, T. Morin, F. Druart, J. L. Legoux, D. Labarriere, B. Naudy, D. Goldfain, A. Rotenberg, C. Bories, A. Andrieu, J. Doll and J. L. Staub. *Mitochondrial D310 mutations in colorectal adenomas: an early but not causative genetic event during colorectal carcinogenesis*. Int. J. Cancer, vol. 122, no. 10, pages 2242–2248, 2008. (Cited on page 29.)
- [Lessa *et al.* 2015] F. C. Lessa, L. G. Winston and L. C. McDonald. *Burden of extitClostridium difficile infection in the United States*. N. Engl. J. Med., vol. 372, no. 24, pages 2369–2370, 2015. (Cited on page 50.)
- [Lewis *et al.* 2012] L. A. Lewis, M. Astatke, P. T. Umekubo, S. Alvi, R. Saby, J. Afrose, P. H. Oliveira, G. A. Monteiro and D. M. Prazeres. *Protein-DNA interactions define the mechanistic aspects of circle formation and insertion reactions in IS2 transposition*. Mob. DNA, vol. 3, no. 1, page 1, 2012. (Cited on pages 11, 25 and 26.)
- [Li & Marban 2010] T. S. Li and E. Marban. *Physiological levels of reactive oxygen species are required to maintain genomic stability in stem cells*. Stem Cells, vol. 28, no. 7, pages 1178–1185, 2010. (Cited on page 27.)
- [Lim & van Oudenaarden 2007] H. N. Lim and A. van Oudenaarden. *A multistep epigenetic switch enables the stable inheritance of DNA methylation states*. Nat. Genet., vol. 39, no. 2, pages 269–275, 2007. (Cited on page 54.)

- [Loenen *et al.* 2014] W. A. Loenen, D. T. Dryden, E. A. Raleigh, G. G. Wilson and N. E. Murray. *Highlights of the DNA cutters: a short history of the restriction enzymes*. *Nucleic Acids Res.*, vol. 42, no. 1, pages 3–19, 2014. (Cited on page 36.)
- [Low *et al.* 2001] D. A. Low, N. J. Weyand and M. J. Mahan. *Roles of DNA adenine methylation in regulating bacterial gene expression and virulence*. *Infect. Immun.*, vol. 69, no. 12, pages 7197–204, 2001. (Cited on page 50.)
- [Lysaght & Campbell 2011] T. Lysaght and A. V. Campbell. *Regulating autologous adult stem cells: the FDA steps up*. *Cell Stem Cell*, vol. 9, no. 5, pages 393–396, 2011. (Cited on page 27.)
- [Mahillon & Chandler 1998] J. Mahillon and M. Chandler. *Insertion sequences*. *Microbiol. Mol. Biol. Rev.*, vol. 62, no. 3, pages 725–774, 1998. (Cited on page 24.)
- [Mairhofer *et al.* 2008] J. Mairhofer, I. Pfaffenzeller, D. Merz and R. Grabherr. *A novel antibiotic free plasmid selection system: advances in safe and efficient DNA therapy*. *Biotechnol. J.*, vol. 3, no. 1, pages 83–89, 2008. (Cited on page 18.)
- [Makarova *et al.* 2009] K. S. Makarova, Y. I. Wolf, J. van der Oost and E. V. Koonin. *Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements*. *Biol Direct*, vol. 4, page 29, 2009. (Cited on pages 69 and 72.)
- [Makarova *et al.* 2011] K. S. Makarova, Y. I. Wolf, S. Snir and E. V. Koonin. *Defense islands in bacterial and archaeal genomes and prediction of novel defense systems*. *J. Bacteriol.*, vol. 193, no. 21, pages 6039–6056, 2011. (Cited on page 36.)
- [Makarova *et al.* 2013] K. S. Makarova, Y. I. Wolf and E. V. Koonin. *Comparative genomics of defense systems in archaea and bacteria*. *Nucleic Acids Res.*, vol. 41, no. 8, pages 4360–4377, 2013. (Cited on page 36.)
- [Manso *et al.* 2014] A. S. Manso, M. H. Chai, J. M. Atack, L. Furi, M. De Ste Croix, R. Haigh, C. Trappetti, A. D. Ogunniyi, L. K. Shewell, M. Boitano, T. A. Clark, J. Kurlach, M. Blades, E. Mirkes, A. N. Gorban, J. C. Paton, M. P. Jennings and M. R. Oggioni. *A random six-phase switch regulates pneumococcal virulence via global epigenetic changes*. *Nat. Commun.*, vol. 5, page 5055, 2014. (Cited on page 50.)
- [Mira *et al.* 2001] A. Mira, H. Ochman and N. A. Moran. *Deletional bias and the evolution of bacterial genomes*. *Trends Genet.*, vol. 17, no. 10, pages 589–596, 2001. (Cited on page 45.)

- [Mochizuki *et al.* 2006] A. Mochizuki, K. Yahara, I. Kobayashi and Y. Iwasa. *Genetic addiction: selfish gene's strategy for symbiosis in the genome*. *Genetics*, vol. 172, no. 2, pages 1309–1323, 2006. (Cited on pages 36 and 37.)
- [Mouammine & Collier 2018] A. Mouammine and J. Collier. *The impact of DNA methylation in alphaproteobacteria*. *Mol. Microbiol.*, vol. 8, no. 12091, 2018. (Cited on page 61.)
- [Moxon *et al.* 2006] R. Moxon, C. Bayliss and D. Hood. *Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation*. *Annu. Rev. Genet.*, vol. 40, pages 307–333, 2006. (Cited on page 54.)
- [Mruk & Kobayashi 2014] I. Mruk and I. Kobayashi. *To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems*. *Nucleic Acids Res.*, vol. 42, no. 1, pages 70–86, 2014. (Cited on page 36.)
- [Muller & Varmus 1994] H. P. Muller and H. E. Varmus. *DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes*. *EMBO J.*, vol. 13, no. 19, pages 4704–4714, 1994. (Cited on page 24.)
- [Murray *et al.* 2012] I. A. Murray, T. A. Clark, R. D. Morgan, M. Boitano, B. P. Anton, K. Luong, A. Fomenkov, S. W. Turner, J. Korlach and R. J. Roberts. *The methylomes of six bacteria*. *Nucleic Acids Res.*, vol. 40, no. 22, pages 11450–62, 2012. (Cited on page 50.)
- [Naito *et al.* 1995] T. Naito, K. Kusano and I. Kobayashi. *Selfish behavior of restriction-modification systems*. *Science*, vol. 267, no. 5199, pages 897–899, 1995. (Cited on page 36.)
- [Nandi *et al.* 2015] T. Nandi, M. T. Holden, M. T. Holden, X. Didelot, K. Meher-shahi, J. A. Boddey, I. Beacham, I. Peak, J. Harting, P. Baybayan, Y. Guo, S. Wang, L. C. How, B. Sim, A. Essex-Lopresti, M. Sarkar-Tyson, M. Nelson, S. Smither, C. Ong, L. T. Aw, C. H. Hoon, S. Michell, D. J. Studholme, R. Titball, S. L. Chen, J. Parkhill and P. Tan. *Burkholderia pseudomallei sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles*. *Genome Res.*, vol. 25, no. 1, pages 129–141, 2015. (Cited on pages 41 and 43.)
- [Novichkov *et al.* 2013] P. S. Novichkov, A. E. Kazakov, D. A. Ravcheev, S. A. Leyn, G. Y. Kovaleva, R. A. Sutormin, M. D. Kazanov, W. Riehl, A. P. Arkin, I. Dubchak and D. A. Rodionov. *RegPrecise 3.0 - A resource for genome-scale exploration of transcriptional regulation in bacteria*. *BMC Genomics*, vol. 14, no. 745, 2013. (Cited on page 65.)
- [O'Connell Motherway *et al.* 2014] M. O'Connell Motherway, D. Watson, F. Bottacini, T. A. Clark, R. J. Roberts, J. Korlach, P. Garault, C. Chervaux,

- J. E. van Hylckama Vlieg, T. Smokvina and D. van Sinderen. *Identification of restriction-modification systems of *Bifidobacterium animalis* subsp. *lactis* CNCM I-2494 by SMRT sequencing and associated methylome analysis*. PLoS ONE, vol. 9, no. 4, page e94875, 2014. (Cited on page 36.)
- [Ofir *et al.* 2018] G. Ofir, S. Melamed, H. Sberro, Z. Mukamel, S. Silverman, G. Yaakov, S. Doron and R. Sorek. *DISARM is a widespread bacterial defence system with broad anti-phage activities*. Nat Microbiol, vol. 3, no. 1, pages 90–98, 2018. (Cited on pages 69 and 72.)
- [Oliveira & Fang 2021] P. H. Oliveira and G. Fang. *Conserved DNA methyltransferases: A window into fundamental mechanisms of epigenetic regulation in Bacteria*. Trends Microbiol., vol. 29, no. 1, pages 28–40, 2021. (Cited on pages 61 and 62.)
- [Oliveira & Mairhofer 2013] P. H. Oliveira and J. Mairhofer. *Marker-free plasmids for biotechnological applications - implications and perspectives*. Trends Biotechnol., vol. 31, no. 9, pages 539–547, 2013. (Cited on pages 11, 18 and 19.)
- [Oliveira *et al.* 2008] P. H. Oliveira, F. Lemos, G. A. Monteiro and D. M. Prazeres. *Recombination frequency in plasmid DNA containing direct repeats—predictive correlation with repeat and intervening sequence length*. Plasmid, vol. 60, no. 2, pages 159–165, 2008. (Cited on pages 11, 18, 22 and 23.)
- [Oliveira *et al.* 2009a] P. H. Oliveira, K. J. Prather, D. M. Prazeres and G. A. Monteiro. *Structural instability of plasmid biopharmaceuticals: challenges and implications*. Trends Biotechnol., vol. 27, no. 9, pages 503–511, 2009. (Cited on pages 11, 18 and 19.)
- [Oliveira *et al.* 2009b] P. H. Oliveira, D. M. Prazeres and G. A. Monteiro. *Deletion formation mutations in plasmid expression vectors are unfavored by runaway amplification conditions and differentially selected under kanamycin stress*. J. Biotechnol., vol. 143, no. 4, pages 231–238, 2009. (Cited on pages 11, 21 and 24.)
- [Oliveira *et al.* 2010] P. H. Oliveira, K. L. J. Prather, D. M. F. Prazeres and G. A. Monteiro. *Analysis of DNA repeats in bacterial plasmids reveals the potential for recurrent instability events*. Appl. Microbiol. Biotechnol., vol. 87, no. 6, pages 2157–2167, 2010. (Cited on pages 11, 21 and 22.)
- [Oliveira *et al.* 2011] P. H. Oliveira, K. L. J. Prather, D. M. F. Prazeres and G. A. Monteiro. *Mutation detection in plasmid-based biopharmaceuticals*. Biotechnol. J., vol. 6, no. 4, pages 378–391, 2011. (Cited on page 11.)
- [Oliveira *et al.* 2012] P. H. Oliveira, J. S. Boura, M. M. Abecasis, J. M. Gimble, C. L. da Silva and J. M. Cabral. *Impact of hypoxia and long-term cultivation*

- on the genomic stability and mitochondrial performance of ex vivo expanded human stem/stromal cells.* Stem Cell Res., vol. 9, no. 3, pages 225–236, 2012. (Cited on pages 11, 30 and 31.)
- [Oliveira *et al.* 2013] P. H. Oliveira, C. Lobato da Silva and J. M. S. Cabral. *An appraisal of human mitochondrial DNA instability: new insights into the role of non-canonical DNA structures and sequence motifs.* PLoS ONE, vol. 8, no. 3, page e59907, 2013. (Cited on pages 11, 34 and 35.)
- [Oliveira *et al.* 2014a] P. H. Oliveira, C. L. da Silva and J. M. Cabral. *Concise review: Genomic instability in human stem cells: current status and future challenges.* Stem Cells, vol. 32, no. 11, pages 2824–2832, 2014. (Cited on page 11.)
- [Oliveira *et al.* 2014b] P. H. Oliveira, M. Touchon and E. P. Rocha. *The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts.* Nucleic Acids Res., vol. 42, no. 16, pages 10618–31, 2014. (Cited on pages 11, 37, 38, 39, 40, 50, 61, 63, 66, 68, 69 and 70.)
- [Oliveira *et al.* 2016] P. H. Oliveira, M. Touchon and E. P. Rocha. *Regulation of genetic flux between bacteria by restriction-modification systems.* Proc. Natl. Acad. Sci. USA, vol. 113, no. 20, pages 5658–63, 2016. (Cited on pages 11, 42, 43, 44, 50 and 69.)
- [Oliveira *et al.* 2017] P. H. Oliveira, M. Touchon, J. Cury and E. P. C. Rocha. *The chromosomal organization of horizontal gene transfer in bacteria.* Nat. Commun., vol. 8, no. 1, page 841, 2017. (Cited on pages 11, 45, 46, 47, 48 and 69.)
- [Oliveira *et al.* 2020] P. H. Oliveira, J. W. Ribis, E. M. Garrett, D. Trzilova, A. Kim, O. Sekulovic, E. A. Mead, T. Pak, S. Zhu, G. Deikus, M. Touchon, M. Lewis-Sandari, C. Beckford, N. E. Zeitouni, D. R. Altman, E. Webster, I. Oussenko, S. Bunyavanich, A. K. Aggarwal, A. Bashir, G. Patel, F. Wallach, C. Hamula, S. Huprikar, E. E. Schadt, R. Sebra, H. van Bakel, A. Kasarskis, R. Tamayo, A. Shen and G. Fang. *Epigenomic characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis.* Nat. Microbiol., vol. 5, no. 1, pages 166–180, 2020. (Cited on pages 12, 51, 53, 55, 57, 60, 61 and 63.)
- [Oliveira 2010] P. H. Oliveira. *Structural instability in plasmid-based biopharmaceuticals.* PhD thesis, Instituto Superior Técnico, 2010. (Cited on page 11.)
- [Oliveira 2021] P. H. Oliveira. *Bacterial epigenomics: Coming of age.* mSystems, vol. 6, no. 4, pages e00747–21, 2021. (Cited on pages 12, 77 and 78.)
- [Overbeek *et al.* 1999] R. Overbeek, M. Fonstein, M. D’Souza, G. D. Pusch and N. Maltsev. *The use of gene clusters to infer functional coupling.* Proc.

- Natl. Acad. Sci. U.S.A., vol. 96, no. 6, pages 2896–2901, 1999. (Cited on page 45.)
- [Paredes-Sabja *et al.* 2014] D. Paredes-Sabja, A. Shen and J. A. Sorg. *Clostridium difficile* spore biology: sporulation, germination, and spore structural proteins. Trends Microbiol., vol. 22, no. 7, pages 406–416, 2014. (Cited on page 49.)
- [Peng *et al.* 2011] Z. Peng, C. Xie, Q. Wan, L. Zhang, W. Li and S. Wu. Sequence variations of mitochondrial DNA D-loop region are associated with familial nasopharyngeal carcinoma. Mitochondrion, vol. 11, no. 2, pages 327–333, 2011. (Cited on page 32.)
- [Perez-Martin & de Lorenzo 1997] J. Perez-Martin and V. de Lorenzo. Clues and consequences of DNA bending in transcription. Annu. Rev. Microbiol., vol. 51, pages 593–628, 1997. (Cited on page 33.)
- [Prather *et al.* 2006] K. L. Prather, M. C. Edmonds and J. W. Herod. Identification and characterization of IS1 transposition in plasmid amplification mutants of *E. coli* clones producing DNA vaccines. Appl. Microbiol. Biotechnol., vol. 73, no. 4, pages 815–826, 2006. (Cited on page 24.)
- [Prigione *et al.* 2010] A. Prigione, B. Fauler, R. Lurz, H. Lehrach and J. Adjaye. The senescence-related mitochondrial/oxidative stress pathway is repressed in human induced pluripotent stem cells. Stem Cells, vol. 28, no. 4, pages 721–733, 2010. (Cited on page 32.)
- [Rehman 2010] J. Rehman. Empowering self-renewal and differentiation: the role of mitochondria in stem cells. J. Mol. Med., vol. 88, no. 10, pages 981–986, 2010. (Cited on page 32.)
- [Ribeiro *et al.* 2008] S. C. Ribeiro, P. H. Oliveira, D. M. Prazeres and G. A. Monteiro. High frequency plasmid recombination mediated by 28 bp direct repeats. Mol. Biotechnol., vol. 40, no. 3, pages 252–260, 2008. (Cited on pages 11, 20 and 21.)
- [Ribeiro *et al.* 2010] S. C. Ribeiro, D. M. Prazeres and G. A. Monteiro. Evidence for the *in vivo* expression of a distant downstream gene under the control of *ColE1* replication origin. Appl. Microbiol. Biotechnol., vol. 86, no. 2, pages 671–679, 2010. (Cited on page 20.)
- [Roberts *et al.* 2003] R. J. Roberts, M. Belfort, T. Bestor, A. S. Bhagwat, T. A. Bickle, J. Bitinaite, R. M. Blumenthal, S. K. h. Degtyarev, D. T. Dryden, K. Dybvig, K. Firman, E. S. Gromova, R. I. Gumport, S. E. Halford, S. Hattman, J. Heitman, D. P. Hornby, A. Janulaitis, A. Jeltsch, J. Josephsen, A. Kiss, T. R. Klaenhammer, I. Kobayashi, H. Kong, D. H. Kruger, S. Lacks, M. G. Marinus, M. Miyahara, R. D. Morgan, N. E. Murray, V. Nagaraja, A. Piekarowicz, A. Pingoud, E. Raleigh, D. N. Rao, N. Reich, V. E.

- Repin, E. U. Selker, P. C. Shaw, D. C. Stein, B. L. Stoddard, W. Szybalski, T. A. Trautner, J. L. Van Etten, J. M. Vitor, G. G. Wilson and S. Y. Xu. *A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes*. *Nucleic Acids Res.*, vol. 31, no. 7, pages 1805–1812, 2003. (Cited on page 36.)
- [Roberts *et al.* 2013] G. A. Roberts, P. J. Houston, J. H. White, K. Chen, A. S. Stephanou, L. P. Cooper, D. T. Dryden and J. A. Lindsay. *Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant Staphylococcus aureus (MRSA) populations*. *Nucleic Acids Res.*, vol. 41, no. 15, pages 7472–7484, 2013. (Cited on page 41.)
- [Roberts *et al.* 2015] R. J. Roberts, T. Vincze, J. Posfai and D. Macelis. *REBASE—a database for DNA restriction and modification: enzymes, genes and genomes*. *Nucleic Acids Res.*, vol. 43, no. Database issue, pages D298–299, 2015. (Cited on pages 36 and 70.)
- [Rocha 2008] E. P. Rocha. *The organization of the bacterial genome*. *Annu. Rev. Genet.*, vol. 42, pages 211–233, 2008. (Cited on page 45.)
- [Rodriguez-Jimenez *et al.* 2008] F. J. Rodriguez-Jimenez, V. Moreno-Manzano, R. Lucas-Dominguez and J. M. Sanchez-Puelles. *Hypoxia causes downregulation of mismatch repair system and genomic instability in stem cells*. *Stem Cells*, vol. 26, no. 8, pages 2052–2062, 2008. (Cited on pages 27 and 29.)
- [Rowe-Magnus *et al.* 2003] D. A. Rowe-Magnus, A. M. Guerout, L. Biskri, P. Bouige and D. Mazel. *Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae*. *Genome Res.*, vol. 13, no. 3, pages 428–442, 2003. (Cited on page 40.)
- [Rubio *et al.* 2021] T. Rubio, Gagné S., C. D. Debruyne, C. Cluzel, D. Mongellaz, Rousselle P., S Gottig, H. Seifert, P.G. Higgins and S.P. Salcedo. *Incidence of an intracellular multiplication niche amongst Acinetobacter baumannii clinical isolates*. *bioRxiv*, 2021. (Cited on page 64.)
- [Sage *et al.* 2010] J. M. Sage, O. S. Gildemeister and K. L. Knight. *Discovery of a novel function for human Rad51: maintenance of the mitochondrial genome*. *J. Biol. Chem.*, vol. 285, no. 25, pages 18984–18990, 2010. (Cited on page 32.)
- [Samuels *et al.* 2004] D. C. Samuels, E. A. Schon and P. F. Chinnery. *Two direct repeats cause most human mtDNA deletions*. *Trends Genet.*, vol. 20, no. 9, pages 393–398, 2004. (Cited on page 33.)
- [Santos *et al.* 2011] F. d. Santos, P. Z. Andrade, M. M. Abecasis, J. M. Gimble, L. G. Chase, A. M. Campbell, S. Boucher, M. C. Vemuri, C. L. Silva and J. M. Cabral. *Toward a clinical-grade expansion of mesenchymal stem cells from human sources: a microcarrier-based culture system under xeno-free*

- conditions*. Tissue Eng Part C Methods, vol. 17, no. 12, pages 1201–1210, 2011. (Cited on page 28.)
- [Schneeberger *et al.* 2005] R. G. Schneeberger, K. Zhang, T. Tatarinova, M. Troukhan, S. F. Kwok, J. Drais, K. Klinger, F. Orejudos, K. Macy, A. Bhakta, J. Burns, G. Subramanian, J. Donson, R. Flavell and K. A. Feldmann. *Agrobacterium T-DNA integration in Arabidopsis is correlated with DNA sequence compositions that occur frequently in gene promoter regions*. Funct. Integr. Genomics, vol. 5, no. 4, pages 240–253, 2005. (Cited on page 24.)
- [Sebaihia *et al.* 2006] M. Sebaihia, B. W. Wren, P. Mullany, N. F. Fairweather, N. Minton, R. Stabler, N. R. Thomson, A. P. Roberts, A. M. Cerdeno-Tarraga, H. Wang, M. T. Holden, A. Wright, C. Churcher, M. A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S. Whithead, B. Dupuy, G. Dougan, B. Barrell and J. Parkhill. *The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome*. Nat. Genet., vol. 38, no. 7, pages 779–786, 2006. (Cited on page 50.)
- [Seekatz & Young 2014] A. M. Seekatz and V. B. Young. *Clostridium difficile and the microbiota*. J. Clin. Invest., vol. 124, no. 10, pages 4182–4189, 2014. (Cited on page 49.)
- [Seemann 2014] T. Seemann. *Prokka: rapid prokaryotic genome annotation*. Bioinformatics, vol. 30, no. 14, pages 2068–2069, 2014. (Cited on page 70.)
- [Sensebe *et al.* 2011] L. Sensebe, P. Bourin and K. Tarte. *Good manufacturing practices production of mesenchymal stem/stromal cells*. Hum. Gene Ther., vol. 22, no. 1, pages 19–26, 2011. (Cited on pages 27 and 28.)
- [Seshasayee *et al.* 2012] A. S. Seshasayee, P. Singh and S. Krishna. *Context-dependent conservation of DNA methyltransferases in bacteria*. Nucleic Acids Res., vol. 40, no. 15, pages 7066–7073, 2012. (Cited on page 40.)
- [Sharp *et al.* 1989] P. M. Sharp, D. C. Shields, K. H. Wolfe and W. H. Li. *Chromosomal location and evolutionary rate variation in enterobacterial genes*. Science, vol. 246, no. 4931, pages 808–810, 1989. (Cited on page 45.)
- [Smits *et al.* 2016] W. K. Smits, D. Lyras, D. B. Lacy, M. H. Wilcox and E. J. Kuijper. *Clostridium difficile infection*. Nat Rev Dis Primers, vol. 2, page 16020, 2016. (Cited on page 49.)
- [Stewart *et al.* 2019] R. D. Stewart, M. D. Auffret, T. J. Snelling, R. Roehle and M. Watson. *MAGpy: a reproducible pipeline for the downstream analysis of*

- metagenome-assembled genomes (MAGs)*. *Bioinformatics*, vol. 35, no. 12, pages 2150–2152, 2019. (Cited on page 70.)
- [Strauer *et al.* 2002] B. E. Strauer, M. Brehm, T. Zeus, M. Kosterling, A. Hernandez, R. V. Sorg, G. Kogler and P. Wernet. *Repair of infarcted myocardium by autologous intracoronary mononuclear bone marrow cell transplantation in humans*. *Circulation*, vol. 106, no. 15, pages 1913–1918, 2002. (Cited on page 28.)
- [Summers *et al.* 1993] D. K. Summers, C. W. Beton and H. L. Withers. *Multicopy plasmid instability: the dimer catastrophe hypothesis*. *Mol. Microbiol.*, vol. 8, no. 6, pages 1031–1038, 1993. (Cited on page 20.)
- [Szeverenyi *et al.* 1996] I. Szeverenyi, A. Hodel, W. Arber and F. Olsz. *Vector for IS element entrapment and functional characterization based on turning on expression of distal promoterless genes*. *Gene*, vol. 174, no. 1, pages 103–110, 1996. (Cited on page 24.)
- [Takahashi *et al.* 2011] N. Takahashi, S. Ohashi, M. R. Sadykov, Y. Mizutani-Ui and I. Kobayashi. *IS-linked movement of a restriction-modification system*. *PLoS ONE*, vol. 6, no. 1, page e16554, 2011. (Cited on page 36.)
- [Tarte *et al.* 2010] K. Tarte, J. Gaillard, J. J. Lataillade, L. Fouillard, M. Becker, H. Mossafa, A. Tchirkov, H. Rouard, C. Henry, M. Splingard, J. Dulong, D. Monnier, P. Gourmelon, N. C. Gorin and L. Sensebe. *Clinical-grade production of human mesenchymal stromal cells: occurrence of aneuploidy without transformation*. *Blood*, vol. 115, no. 8, pages 1549–1553, 2010. (Cited on page 28.)
- [Tesson *et al.* 2021] F. Tesson, A. Hervé, M. Touchon, C. d’Humières, J. Cury and A. Bernheim. *Systematic and quantitative view of the antiviral arsenal of prokaryotes*. *bioRxiv*, 2021. (Cited on page 71.)
- [Thomas & Nielsen 2005] C. M. Thomas and K. M. Nielsen. *Mechanisms of, and barriers to, horizontal gene transfer between bacteria*. *Nat. Rev. Microbiol.*, vol. 3, no. 9, pages 711–721, 2005. (Cited on page 35.)
- [Touchon *et al.* 2009] M. Touchon, C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat, S. Mangenot, V. Martinez-Jehanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Turret, B. Vacherie, D. Vallenet, C. Medigue, E. P. Rocha and E. Denamur. *Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths*. *PLoS Genet.*, vol. 5, no. 1, page e1000344, 2009. (Cited on page 46.)

- [Treangen & Rocha 2011] T. J. Treangen and E. P. Rocha. *Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes*. PLoS Genet., vol. 7, no. 1, page e1001284, 2011. (Cited on page 35.)
- [Tsai *et al.* 2011] C. C. Tsai, Y. J. Chen, T. L. Yew, L. L. Chen, J. Y. Wang, C. H. Chiu and S. C. Hung. *Hypoxia inhibits senescence and maintains mesenchymal stem cell properties through down-regulation of E2A-p21 by HIF-TWIST*. Blood, vol. 117, no. 2, pages 459–469, 2011. (Cited on pages 27 and 28.)
- [Tuppen *et al.* 2010] H. A. Tuppen, E. L. Blakely, D. M. Turnbull and R. W. Taylor. *Mitochondrial DNA mutations and human disease*. Biochim. Biophys. Acta, vol. 1797, no. 2, pages 113–128, 2010. (Cited on page 32.)
- [Uchijima *et al.* 1998] M. Uchijima, A. Yoshida, T. Nagata and Y. Koide. *Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T cell responses against an intracellular bacterium*. J. Immunol., vol. 161, no. 10, pages 5594–5599, 1998. (Cited on page 18.)
- [Ueyama *et al.* 2012] H. Ueyama, T. Horibe, S. Hinotsu, T. Tanaka, T. Inoue, H. Urushihara, A. Kitagawa and K. Kawakami. *Chromosomal variability of human mesenchymal stem cells cultured under hypoxic conditions*. J. Cell. Mol. Med., vol. 16, no. 1, pages 72–82, 2012. (Cited on pages 27 and 28.)
- [Unterholzner *et al.* 2013] S. J. Unterholzner, B. Poppenberger and W. Rozhon. *Toxin-antitoxin systems: Biology, identification, and application*. Mob Genet Elements, vol. 3, no. 5, page e26219, 2013. (Cited on page 69.)
- [Vaish 2007] M. Vaish. *Mismatch repair deficiencies transforming stem cells into cancer stem cells and therapeutic implications*. Mol. Cancer, vol. 6, page 26, 2007. (Cited on page 29.)
- [Valiente *et al.* 2014] E. Valiente, M. D. Cairns and B. W. Wren. *The Clostridium difficile PCR ribotype 027 lineage: a pathogen on the move*. Clin. Microbiol. Infect., vol. 20, no. 5, pages 396–404, 2014. (Cited on page 50.)
- [Vandenbussche *et al.* 2020] I. Vandenbussche, A. Sass, M. Pinto-Carbó, O. Manweiler, L. Eberl and T. Coenye. *DNA methylation epigenetically regulates gene expression in Burkholderia cenocepacia and controls biofilm formation, cell aggregation, and motility*. mSphere, vol. 5, no. 4, pages e00455–20, 2020. (Cited on pages 63 and 64.)
- [Vieira-Silva & Rocha 2010] S. Vieira-Silva and E. P. Rocha. *The systemic imprint of growth and its uses in ecological (meta)genomics*. PLoS Genet., vol. 6, no. 1, page e1000808, 2010. (Cited on page 45.)

- [Vlahovicek *et al.* 2003] K. Vlahovicek, L. Kajan and S. Pongor. *DNA analysis servers: plot.it, bend.it, model.it and IS*. Nucleic Acids Res., vol. 31, no. 13, pages 3686–3687, 2003. (Cited on page 66.)
- [Wang *et al.* 2004] H. Wang, M. Noordewier and C. J. Benham. *Stress-induced DNA duplex destabilization (SIDDD) in the E. coli genome: SIDDD sites are closely associated with promoters*. Genome Res., vol. 14, no. 8, pages 1575–1584, 2004. (Cited on page 66.)
- [Wang *et al.* 2011] L. Wang, S. Chen, K. L. Vergin, S. J. Giovannoni, S. W. Chan, M. S. DeMott, K. Taghizadeh, O. X. Cordero, M. Cutler, S. Timberlake, E. J. Alm, M. F. Polz, J. Pinhassi, Z. Deng and P. C. Dedon. *DNA phosphorothioation is widespread and quantized in bacterial genomes*. Proc Natl Acad Sci U S A, vol. 108, no. 7, pages 2963–2968, 2011. (Cited on page 69.)
- [Williams 2002] K. P. Williams. *Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies*. Nucleic Acids Res., vol. 30, no. 4, pages 866–875, 2002. (Cited on page 46.)
- [Wilmes *et al.* 2009] P. Wilmes, S. L. Simmons, V. J. Denef and J. F. Banfield. *The dynamic genetic repertoire of microbial communities*. FEMS Microbiol. Rev., vol. 33, no. 1, pages 109–132, 2009. (Cited on page 45.)
- [Wion & Casadesus 2006] D. Wion and J. Casadesus. *N6-methyl-adenine: an epigenetic signal for DNA-protein interactions*. Nat. Rev. Microbiol., vol. 4, no. 3, pages 183–192, 2006. (Cited on pages 50 and 54.)
- [Zhang *et al.* 2007] J. Zhang, D. Guo, Y. Chang, C. You, X. Li, X. Dai, Q. Weng, J. Zhang, G. Chen, X. Li, H. Liu, B. Han, Q. Zhang and C. Wu. *Non-random distribution of T-DNA insertions at various levels of the genome hierarchy as revealed by analyzing 13 804 T-DNA flanking sequences from an enhancer-trap mutant library*. Plant J., vol. 49, no. 5, pages 947–959, 2007. (Cited on page 24.)
- [Zidaric & Rupnik 2016] V. Zidaric and M. Rupnik. *Sporulation properties and antimicrobial susceptibility in endemic and rare Clostridium difficile PCR ribotypes*. Anaerobe, vol. 39, pages 183–188, 2016. (Cited on page 50.)

APPENDIX A

5 most important publications

The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts

Pedro H. Oliveira^{1,2,*}, Marie Touchon^{1,2} and Eduardo P.C. Rocha^{1,2}

¹Institut Pasteur, Microbial Evolutionary Genomics, Département Génomes et Génétique, Paris, France and ²CNRS, UMR3525, Paris, France

Received June 4, 2014; Revised July 29, 2014; Accepted July 30, 2014

ABSTRACT

The roles of restriction-modification (R-M) systems in providing immunity against horizontal gene transfer (HGT) and in stabilizing mobile genetic elements (MGEs) have been much debated. However, few studies have precisely addressed the distribution of these systems in light of HGT, its mechanisms and its vectors. We analyzed the distribution of R-M systems in 2261 prokaryote genomes and found their frequency to be strongly dependent on the presence of MGEs, CRISPR-Cas systems, integrons and natural transformation. Yet R-M systems are rare in plasmids, in prophages and nearly absent from other phages. Their abundance depends on genome size for small genomes where it relates with HGT but saturates at two occurrences per genome. Chromosomal R-M systems might evolve under cycles of purifying and relaxed selection, where sequence conservation depends on the biochemical activity and complexity of the system and total gene loss is frequent. Surprisingly, analysis of 43 pan-genomes suggests that solitary R-M genes rarely arise from the degradation of R-M systems. Solitary genes are transferred by large MGEs, whereas complete systems are more frequently transferred autonomously or in small MGEs. Our results suggest means of testing the roles for R-M systems and their associations with MGEs.

INTRODUCTION

The flow of genetic information between bacterial cells by horizontal gene transfer (HGT) drives bacterial evolution (1,2) and restriction-modification (R-M) systems are key moderators of this process (3,4). They are thought to be ubiquitous in bacteria and archaea (5), and operate like many poison-antidote systems: they typically encode a methyltransferase (MTase) function that modifies a particular sequence and a restriction endonuclease (REase) function that cleaves a DNA when its recognition sequence is

unmethylated (6–8). The three classical types of R-M systems differ in their molecular structure, sequence recognition, cleavage position and cofactor requirements (9) (Supplementary Figure S1). Type I systems are complex heterooligomers either comprising one DNA sequence specificity (S), two REase and two MTase subunits with restriction and modification activities, or two MTase and one S subunits with modification activity only. Type II systems encoded on separate genes are composed of one homodimeric or homotetrameric REase and one monomeric MTase, and in most cases are able to operate separately and independently from each other at least *in vitro*. Some Type II systems, particularly Types IIB, IIG, IIL, and some IIH (collectively termed IIC) encode both restriction and modification domains within the same protein (10,11). Type III systems are heterotrimers or heterotetramers of products of two genes, *res* and *mod*, involved in restriction and modification, respectively. Both subunits are required for restriction, whereas Mod is sufficient to produce a modification. Finally, Type IV ‘restriction systems’, as opposed to R-M systems, are composed of one or two REases that cleave modified recognition sites (12).

R-M systems are major players in the co-evolutionary interaction between mobile genetic elements (MGEs) and their hosts. Closely related strains have different systems and distantly related species sometimes have similar systems, suggesting frequent HGT. This leads to weak phylogenetic association between systems and taxa (13–16), and typically one needs to compare strains within species to observe ortholog systems (17,18). Incoming DNA is unlikely to be modified in a way compatible with the R-M systems of the new host and will be degraded. This has led to very early proposals that R-M systems are bacterial innate immune systems (19), since they effectively allow self- from non-self discrimination. R-M systems might preferentially cluster with and stabilize other antivirus defense systems (toxin-antitoxin, abortive infection) in the so-called defense islands, i.e. discrete DNA segments that include a plethora of defense systems (5,20). In some cases, different defense systems have been shown to operate synergistically in order to increase the overall resistance to phage infection (21). Currently, it remains unclear, and is a matter of active re-

*To whom correspondence should be addressed. Tel: +33 1 40 61 33 53; Fax: +33 1 45 68 87 27; Email: pcpco@gmail.com

search, the extent to which such co-localization occurs, its underlying mechanisms and if/how it translates into a functional cooperation between systems.

Some R-M systems can also propagate horizontally in a selfish way. Incoming DNA carrying an R-M system induces 'genetic addiction' to the host by post-segregational killing (22). This behavior leads to the stabilization of MGEs against challenge by competitor elements as long as the R-M system is present (23–25). Accordingly, genes encoding R-M systems have been reported to move between prokaryotic genomes within MGEs such as plasmids (23,26,27), prophages (27,28), insertion sequences/transposons (24,27), integrative conjugative elements (ICEs) (27,29) and integrons (27,30,31). In this regard, a mutual benefit is established between MGEs and R-M systems; the former facilitating horizontal transfer and the latter stabilizing it (31).

In other cases yet, the biological significance of some R-M systems remains obscure [reviewed in (32)]. For example, Type III systems are known to undergo phase variation (33) and Types I and II to affect the expression of certain genes (16,34), which might confer a fitness advantage to the host under certain environmental conditions. The processes underlying birth, death, pseudogenization (genetic degradation) or modification of the function of R-M systems also remain poorly understood, even though they are thought to be at the origin of 'solitary' ('orphan') REases and MTases (35,36), hybrid systems originated by the fusion of R-M components (37,38) or movement of DNA sequence recognition domains between different R-M systems (16,39,40).

The study of R-M systems is at a key point in time. On the one hand, a number of studies have enlarged the known scope of activity of these systems in bacterial cells (20,21,41), and a single resource, REBASE (42), has regrouped most of this information. On the other hand, the recent availability of tools to characterize bacterial methylomes is opening new perspectives on the effect of R-M systems in bacterial epigenetics (16,43). However, there is a lack of recent studies on some of the original questions put forward regarding R-M systems. How abundant are these systems? Which are more abundant? How rapidly do they evolve? How many systems are actually in MGEs? Which MGEs? Is there an association between R-M systems and different mechanisms of genetic mobility? In this work, we have used a comparative genomics approach to answer these questions. For this we have precisely identified MGEs and genes encoding mechanisms of transfer and protection against transfer. With such data at hand, we could characterize the associations between R-M systems and genetic mobility.

MATERIALS AND METHODS

Data

We analyzed 2393 chromosomes and 1813 plasmids representing 2261 fully sequenced prokaryotic genomes (2117 bacterial and 144 archaeal) and 831 complete phage genomes. These sequences and their annotations were retrieved from Genbank Refseq (<ftp://ftp.ncbi.nih.gov/genomes>, last accessed in February 2013). We used the definition of phages, chromosomes and plasmids of GenBank.

We excluded genes indicated in the GenBank files as partial genes, as well as those lacking a stop codon or having one inside the reading frame. Curated reference protein sequences of Types I, II, IIC and III R-M systems and Type IV REases were downloaded from the data set 'gold standards' of REBASE (42) (last accessed in January 2013).

Clustering analyses and construction of protein profiles

All-against-all searches were performed for REase and MTase standard protein sequences retrieved from REBASE using BLASTP (default settings, e -value $<10^{-3}$), and the resulting e -values were log transformed and used for clustering into protein families by Markov Clustering v10–201 (44). To regulate the granularity of clustering, we have modified the inflation parameter (I) by increments of 0.2 in the range of 1.0 to 10.0 and proceeded with values of $I = 1.2–1.4$. In this process, we excluded proteins that were either redundant or very divergent in sequence length. Each protein family was aligned with MAFFT v7.0.17 (45) using the E-INS-i option, 1000 cycles of iterative refinement and offset 0. Alignments were visualized in SEAVIEW v4.4.0 (46) and manually trimmed to remove poorly aligned regions at the extremities. Hidden Markov Model (HMM) profiles were then built from each multiple sequence alignment using the hmmbuild program from the HMMER v3.0 suite (47) (default parameters). Type II MTases were retrieved using the PFAM-A profiles PF01555.12, PF02086.9, PF00145.1 and PF07669.5 (last accessed in February 2013). Types II and IV REases are very divergent and do not produce good multiple alignments (48), which precludes their use to build protein profiles. In these cases BLASTP was used to scan the genomes for homologs (default settings, e -value $<10^{-3}$ and minimum coverage alignment of 50%). For Type IV REases and Type IIC systems, the control by co-localization with other genes of the system is not possible. To check the quality of our identifications, we compared them with the predictions of REBASE. For this, we sampled 10% of the replicons containing Type IIC systems or Type IV REases. Next, we queried REBASE for the total numbers of Type IIC systems and Type IV REases for each of these replicons. We have excluded REBASE hits corresponding to R-M systems interrupted by mobile elements or harboring any frameshifts. We found that for Type IIC systems our predictions and the ones from REBASE were practically identical (only 1.7% of the predicted Type IIC systems were not present in REBASE, whereas only 2.2% of REBASE predictions were lacking in our data set). For Type IV REases, 16.1% of our hits lacked in REBASE and 13.0% of REBASE predictions lacked in our list. We inquired on what would take to change our method to identify more REBASE predictions and found that if we do not require a minimum coverage of the alignment we could recover 98.4% of the REBASE predictions. Nevertheless, since some of these alignments are quite poor, we opted by keeping the coverage criterion at the cost of risking missing a small part of the systems.

Identification of R-M systems and solitary R-M components

Types I, II and III R-M systems were identified by searching genes encoding the MTase and REase components at

less than four genes apart. The output was subsequently curated in order to eliminate multiple occurrences of the same R-M system, for example as a result of the presence of two REase or MTase genes pertaining to the same R-M system. R-M systems containing more than one specificity (S) gene were considered as a single system. Situations involving ambiguous identifications may also occur, for example between REases of Types II and IV, or between Type IIC systems and other MTases or REases. In these cases, the R-M type was defined on the basis of the corresponding genomic context (presence or not of a linked REase or MTase) and on the output of the analysis of the system using REBASE. Type IIC R-M systems were defined as those including a gene encoding both a MTase and a REase function with similarity to Type IIC MTases and REases. An R-M system was defined as 'complete' if both REase and MTase were present. For the contextual analysis of R-M systems, we have considered two independent R-M systems as co-localized if their distance was below 10 genes. Genes of functionally linked components of R-M systems are typically co-localized, and although there are no apparent impediments for the existence of a functional link if the genes are distantly located in a genome (49), there are only very few distantly located R-M systems currently known to be functional (eventually as a result of biased searches), many of them being of Type I (40,50,51). We have considered a REase or MTase as 'solitary', if no cognate MTase or REase was found at a distance of less than 10 genes away, in a similar way to what was performed by others (35). At the current state of knowledge, one cannot use comparative genomics to infer a functional link between non-co-localized REase and MTase genes even though some such cases have been reported (36,52). Given the quick rate of R-M systems gain and loss described in this work, it is unlikely that many R-M systems could have their components encoded in distant regions in the genome.

Analysis of substitution rates

All-against-all BLASTP searches were performed on the sets of putative R-M systems scanned in the genomes (default settings, e -value $<10^{-3}$). Clustering was performed using the SILIX package v1.2.8 (<http://lbbe.univ-lyon1.fr/SiLiX>, last accessed in April 2013) (53) using a minimum identity threshold of 80% and default values for the remaining parameters. Singletons were eliminated from our data set. The remaining protein sequences (putative orthologs) were reverse-translated to the corresponding DNA sequences using PAL2NAL v14 (54). Pairwise rates of non-synonymous substitutions (dN), synonymous substitutions (dS) and ω (dN/dS) were computed using the yn00 program of the PAML package v4.4b (55) implementing the Yang and Nielsen method (56). Estimations yielding $dS > 1$ (corresponding to situations of substitution saturation and representing 16.1% of the total data) were discarded to improve the quality of estimation of ω .

Identification and classification of prophages, conjugative elements, integrons and CRISPR-Cas systems

The identification and classification of prophages was performed as in (57). This corresponds to the genomes of

temperate phages integrated in the bacterial chromosome and is therefore a data set different from the genomes of phages from GenBank, which were sequenced from virions and most often correspond to virulent phages. The identification of genes encoding the functions related to conjugation in ICEs and in integrative mobilizable elements (IMEs) was obtained as in (58). ICEs (also called conjugative transposons) encode the entire machinery required for conjugation between cells. IMEs encode relaxases but lack a complete conjugative transfer system, which is encoded in *trans* by another mobile element. These conjugative elements are very abundant in bacterial genomes (58). The presence of integrons was based on the simultaneous detection of tyrosine recombinases (PFAM family profile PF00589) and of the conserved specific region of integron integrases (a dedicated profile was built using HMMER) (59). Clustered regularly interspaced short palindromic repeats (CRISPRs) were identified following the methodology published in (60). Briefly CRISPRs were identified using the CRISPR Recognition Tool (61) using default parameters. For purposes of protospacer identification, (i.e. sequences from invading genetic elements that are incorporated into CRISPR loci after infection), BLASTN was used for similarity searches between CRISPR spacer sequences and R-M genes of complete systems ($n = 7764$) or R-M solitary genes ($n = 6446$) (default settings, e -value $<10^{-5}$). Matches showing at least 90% of identity and less than 10% difference in sequence length between query and hit were retained. Clusters of *cas* genes were identified using MacSyFinder Abby S. S. *et al.*, (submitted for publication, <https://github.com/gem-pasteur/macsfinder>).

Detection of competence systems

We gathered representative proteins pertaining to competence systems of experimentally studied Gram-positive and Gram-negative model systems, from which multiple alignments and HMM profiles were built. For Gram-positive bacteria, this was performed for the DNA-binding receptor ComEA, the cytoplasmic membrane protein ComEC, the adenosine triphosphate-binding protein ComFA, the traffic NTPase ComGA, the polytopic membrane protein ComGB, the major pseudopilin ComGC and the prepilin peptidase ComC. A PFAM profile was extracted for the DNA processing protein A DprA (PF02481.10). Protein profiles for Gram-negative bacteria were taken from the bacterial secretion system detection tool integrated in MacSyFinder. These include the outer-membrane/tip-located adhesin PilC, the prepilin peptidase PilD, inner-membrane pilus-associated protein PilM, pilot protein PilP, secretin PilQ, PilT/PilU ATPases, minor pilin PilV and major pilins PilA and PilE. PFAM profiles were extracted for the inner-membrane pilus-associated proteins PilN (PF05137.8) and PilO (PF04350.8). Loci encoding the natural transformation machinery were defined as having a maximum distance between two consecutive genes of five and a minimum number of six genes. On the basis of evidence gathered from the literature, we have considered the PilU (62) and ComC (63) components as facultative.

Identification of core- and pan-genomes

We built core-genomes for each of the 43 species having at least seven complete genomes available in Genbank RefSeq (Supplementary Table S1) as in (64). A preliminary set of orthologs was identified as bidirectional-best-hits using end-gap free global alignment, between the proteome of a pivot and each of the other strain proteomes. Hits with less than 80% similarity in amino acid sequence or more than 20% difference in protein length were discarded. For every pairwise comparison, this list of orthologs was then refined taking into account the conservation of gene neighborhood. Because, (i) few genome rearrangements are observed at these short evolutionary distances (65) and (ii) HGT is frequent (2), genes outside conserved blocks of synteny are likely to be xenologs or paralogs. Hence, we combined the previously homology analysis with the classification of these genes as either syntenic or nonsyntenic, for positional orthology determination. Thus, positional orthologs were defined as bi-directional best hits adjacent to at least four other pairs of bi-directional best hits within a neighborhood of 10 genes (five upstream and five downstream). These parameters (four genes being less than half of the diameter of the neighborhood) allow retrieving orthologs on the edge of rearrangement breakpoints and therefore render the analysis robust to the presence of rearrangements. The core-genome of each species was defined as the intersection of pairwise lists of positional orthologs. It thus consists in the genes present in all genomes of a species and, therefore, can also be used to compare the genomic localization of homologous proteins between strains (see below). Pan-genomes are the full complement of genes in the species and were built by clustering homologous proteins into families for each of the 43 species. We determined the lists of putative homologs between pairs of genomes (including plasmids) with BLASTP and used the e -values ($<10^{-4}$) to cluster them using SILIX. SILIX parameters were set such that a protein was homolog to another in a given family if the aligned part had at least 80% (stringent pan-genome) or 40% of identity (relaxed pan-genome) and if it included more than 80% of the smallest protein.

Genomic location of chromosomal R-M systems

Integration regions containing a given R-M system were defined as the regions flanked by the two consecutive core genes that include the system (see above). These regions correspond to single/multiple integration and/or deletion events. They can be strain-specific (recent integration) or shared by different strains (ancestral acquisition). When these regions corresponded to rearrangement breakpoints, i.e. to cases where the flanking core genes in the focal genome were not consecutive core genes in at least another genome of the same clade, they were ignored and excluded from further analysis. We observed such cases for 13 out of 850 R-M systems (less than 2%) and 33 out of 1153 (less than 3%) solitary R-M genes.

RESULTS

Abundance and distribution of R-M systems in genomes

We identified a total of 4743 R-M systems in 2261 prokaryotic genomes (Figure 1A and Supplementary Table S2). Type II systems are the most intensely studied and are also the most abundant (42.4%). Type IIC systems, in which the REase and MTase are part of the same polypeptide (Supplementary Figure S1), account for more than a third (38.8%) of all Type II R-M systems. Our results point to an average value of 0.54 Type II (excluding Type IIC) R-M systems per genome, which is considerably higher than the ratio of 0.35 previously found in the literature (35). Type I are the second most abundant, corresponding to $\sim 29.5\%$ of all R-M systems. Type IV (methylation targeted) REases were found to be much more abundant (19.9%) than Type III (8.2%). We found similar trends in the relative amounts of R-M systems when the analysis was performed in chromosomes and plasmids separately (Supplementary Figure S2A and B). It should be noted that we used different appropriated methods to search for R-M systems. These include BLAST-based methods (REases of Type II and Type IV) and HMM-based methods (all other components). HMM-based methods are more sensitive (66), and we cannot exclude the possibility that we have under-estimated the number of Type IV REases (see the Materials and Methods section for comparisons with REBASE). This large number of Type IV REases is somewhat surprisingly, given the very few studies devoted to them.

The frequency of R-M systems varies widely among bacterial large phyla (Figure 1B; see also Supplementary Figure S2C for clades with less than 10 species in the genome data set). Some clades, such as Alphaproteobacteria or Chlamydiae, have very few systems even when controlled for genome size (less than one system per Mb). One could argue that the presence of many bacteria with small genomes rarely engaging in HGT would lead to fewer events of acquisition of R-M systems and weaker selection for these systems in these clades. However, this does not seem the case for Alphaproteobacteria, as this clade typically shows lower numbers and densities of R-M systems than the remaining Proteobacteria irrespectively of genome size (Supplementary Figure S3A). Hence, taxonomy and genome size may both be important variables determining the number of R-M systems in genomes. For Chlamydiae, the reduced size and number of genomes available does not allow to disentangle between the effects of the two variables. On the other extreme of R-M systems abundance, Epsilonproteobacteria contain by very far the highest number and density of R-M systems. While starting to analyze this question we met with previous observations showing that the genomes of *Helicobacter* harbor an exceptionally high number of R-M systems compared to other genera (average number of R-M systems in *Helicobacter* = 11.8 per genome) (16,32,67,68). We removed these genomes from further analyses because they have been studied before and indeed strongly inflate the statistics in the genome size range [1.5–2.2] Mb. This reduced the size of the genome data set by only 2.3% to 2210 genomes. Tenericutes, which include wall-less bacteria such as *Mycoplasmata*, show the second highest density of R-M

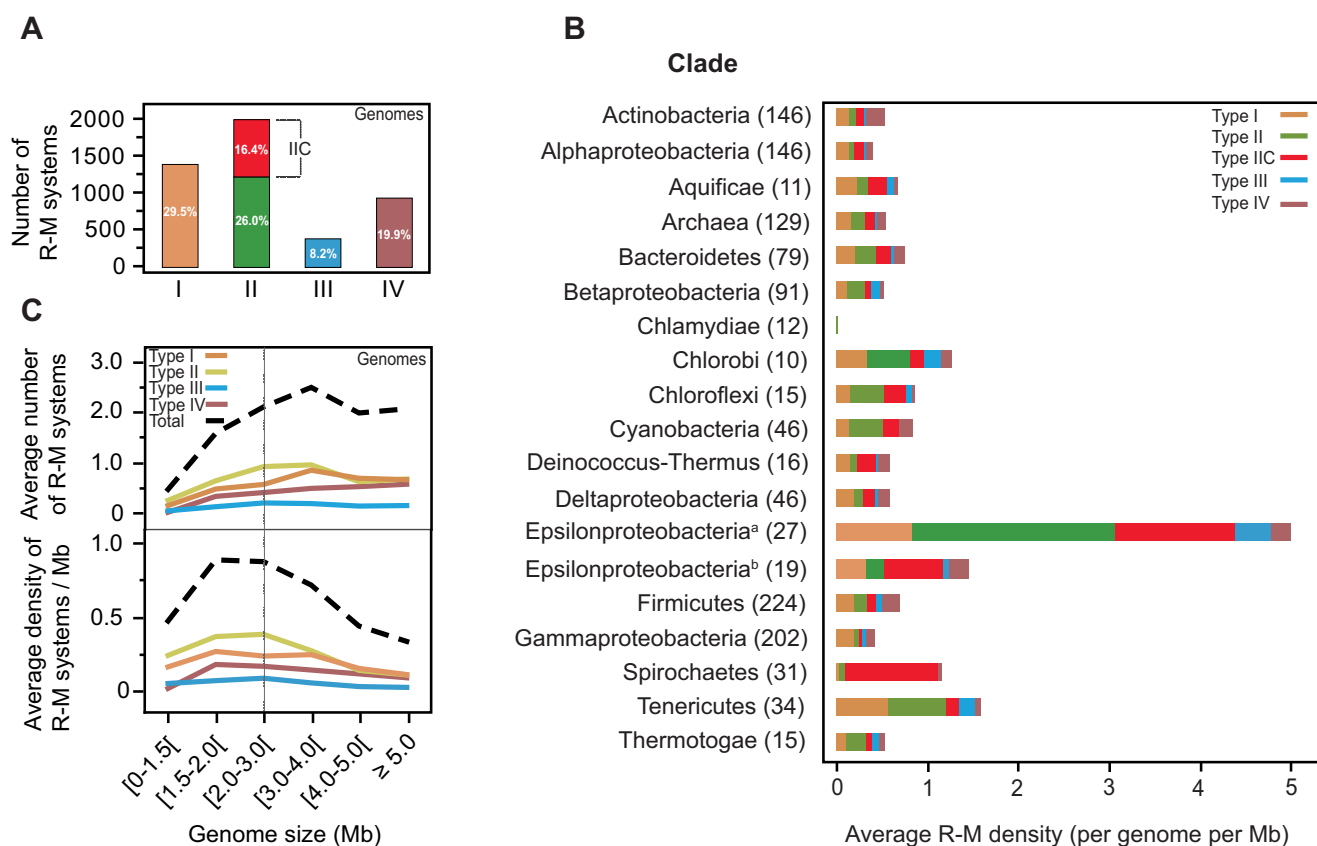


Figure 1. Quantification and distribution of R-M systems in 2261 prokaryotic genomes. (A) Amount of Types I, II, IIC, III R-M systems and Type IV REases found in genomes. Corresponding percentages are indicated. (B) Average R-M density (per genome per Mb) according to clade. The largest peak on R-M density observed for Epsilonproteobacteria^(a) results from the presence of multiple systems particularly among *Helicobacter* species. For comparison, we also show the density for Epsilonproteobacteria without *Helicobacter*^(b). Only clades with at least 10 different species were considered for comparison. The number of species within each clade is indicated next to its name. (C) Distribution of the average number of R-M systems per genome (upper graph) and average density per genome per Mb (bottom graph) according to genome size (Mb). Stippled line separates the regions having small and large genomes. Genomes of *Helicobacter* were not included to avoid obtaining extremely inflated values in the [1.5–2.2[Mb genome size range.

systems. This is surprising, because this clade includes almost only genomes with sizes below 2 Mb (average genome size = 0.892 Mb). Recent reports show clear evidence of frequent horizontal transfer in this phylum (69,70), which might favor the acquisition of R-M systems. Overall, 74.2% of the genomes harbor R-M systems, and no clade is entirely devoid of them. R-M systems are therefore ubiquitous, but the patterns of their distribution are very diverse and depend on genome size, taxonomy and lifestyle.

R-M systems have been shown to be more abundant in larger genomes (20,32). We observed a positive correlation between the total number of R-M systems and genome size (Spearman's $\rho = 0.2256$, $P < 10^{-4}$) (Figure 1C). However, this correlation is more complex than previously suggested. For small genomes (<2 Mb), there is a quick increase in the number (Spearman's $\rho = 0.4758$, $P < 10^{-4}$) and density (Spearman's $\rho = 0.3810$, $P < 10^{-4}$) of R-M systems with genome size. For larger genomes (≥ 2 Mb), the average number of R-M systems is nearly independent of genome size (Spearman's $\rho = -0.0284$, $P > 0.2$) and kept around two per genome. In other words, the number of R-M systems rises with genome size, but the effect saturates for genomes

larger than 2 Mb. Accordingly, the density of R-M systems decreases with increasing genome size for this group (Spearman's $\rho = -0.3434$, $P < 10^{-4}$). These correlations are qualitatively similar when including *Helicobacter* data (Supplementary Figure S3B) or when excluding the very abundant Type II systems from the analysis (Supplementary Figure S3C). Similar trends in the distribution of R-M systems were also observed when the analysis was performed in chromosomes and plasmids separately (Supplementary Figure S4).

R-M systems are over-represented in naturally competent organisms

Horizontal transfer can take place by natural transformation which relies on a complex membrane machinery that, with few known exceptions, is closely related to Type IV pili and Type II secretion systems (71). *Helicobacter pylori* is competent and has many R-M systems, which has led to the suggestion that bacteria amenable to natural transformation have more R-M systems (32). However, *H. pylori* is one of the exceptions and uses a derivative of a Type IV secretion system for transformation (72). As mentioned

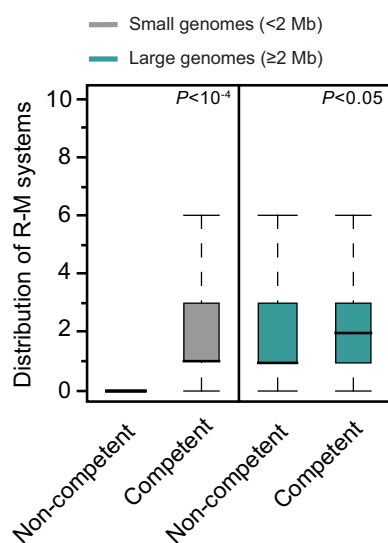


Figure 2. Box plots of the co-occurrence of R-M systems and natural competence machinery in small (<2 Mb) and large (≥ 2 Mb) genomes. Error bars represent standard deviations. Mann–Whitney–Wilcoxon test P value is indicated next to the box plots.

above, it is also an outlier in the distribution of R-M systems. To test the statistical association between the ability for natural transformation and the number of R-M systems, we predicted the presence of the machinery of transformation from genome data. For this, we built protein profiles for the components of DNA uptake systems and used them to query the genomes of 904 Proteobacteria and 463 Firmicutes with MacSyFinder (see the Materials and Methods section for further details). The results of this analysis must be taken with care, since many bacteria encoding the full competence machinery have not yet been proven to be competent, presumably because the conditions where competence is expressed have not yet been found (73). Our results show that indeed genomes encoding the repertoire of genes involved in natural transformation have more R-M systems ($P < 10^{-4}$; Mann–Whitney–Wilcoxon test). This result remains valid when we split the data set into small (<2 Mb) ($P < 10^{-4}$; Mann–Whitney–Wilcoxon test) and large (≥ 2 Mb) ($P < 0.05$; Mann–Whitney–Wilcoxon test) genomes (Figure 2). These findings confirm the association between the presence of HGT and the abundance of R-M systems in genomes. It remains to be understood if naturally transformable bacteria over-represent R-M systems because these systems may be frequently acquired by natural transformation and stabilized by post-segregational killing or because these systems are particularly advantageous for naturally transformable bacteria.

The genetic mobility of R-M systems

We showed that the abundance of R-M systems is associated with the existence of HGT even in naturally transformable bacteria where transfer does not require selfish MGEs. This raises the question of the role of R-M systems in MGEs, independently of their effect on HGT. R-M systems were discovered by their ability to prevent phage infection, but some

phages also encode R-M systems (74). We searched for R-M systems in 831 complete phage genomes available from GenBank and only identified nine R-M systems (seven of Type II and two of Type IV) (Supplementary Figure S5A). Accordingly, the density of R-M systems in phages is lower than those found in any other type of replicon considered in our analysis (Supplementary Figure S5A). This suggests that R-M systems are rarely associated with phages and that these are rarely vectors of their horizontal transfer. To evaluate the association between the ecology of phages and R-M systems, we also studied the R-M systems present in prophages. These are temperate phages that we identified integrated in the chromosome (see the Materials and Methods section). Interestingly, the density of R-M systems in our data set of prophages was 8-fold higher than in the data set of GenBank phage sequences (most of which are virulent phages) ($P < 10^{-4}$; Chi-square test) (Supplementary Figure S5A). This suggests that temperate phages leading to successful lysogens are more likely to encode R-M systems than virulent phages.

Several works have shown that R-M systems stabilize plasmids in cells by their addictive behavior (22,23,25,27). Surprisingly, we found very few systems in plasmids when compared to chromosomes (219 versus 3802). Plasmids are smaller and we do find that the density of R-M systems is around five times higher in plasmids than in chromosomes ($P < 10^{-4}$; Chi-square test) (Supplementary Figure S5A). Nevertheless, only 10.5% of the plasmids encode R-M systems (Type IV REases included), whereas 69% of the chromosomes do so. More than half of the plasmids lack genes associated with conjugation (75). The rarity of R-M systems in plasmids might be the consequence of the presumably lower genetic mobility of these elements. To test this hypothesis, we divided the plasmids into two classes: plasmids encoding the conjugation machinery or at least the relaxase that allows them to be mobilized *in trans* by another conjugation machinery (MOB⁺, 44.6% from total), and plasmids lacking even the relaxase (MOB⁻, 55.4%). More MOB⁺ than MOB⁻ plasmids were found to contain R-M systems (respectively 113 versus 75, $P < 10^{-4}$; Chi-square test) (Figure 3A). MOB⁺ plasmids also have higher density of R-M systems than MOB⁻, showing that larger replicon size is not enough to explain the higher abundance of R-M systems we observe in the former plasmids (respectively 2.89 and 2.23 per Mb, $P < 10^{-4}$; Chi-square test). The abundance of R-M systems in plasmids is therefore linked with their ability to transfer horizontally by conjugation. Plasmids account for a very small percentage of all R-M systems even if they over-represent R-M systems relative to the size of their replicons.

We then aimed at identifying if some types of R-M systems were over- or under-represented in certain types of MGEs. For this, we computed the observed/expected (O/E) ratios of the number of each type of R-M system present within plasmids, prophages and ICES/IMEs. In this test, the null statistical hypothesis is that the relative distribution of the types of R-M systems is similar in chromosomes and MGEs. The distribution of the different types of R-M systems in all these MGEs was different from that of the chromosome (all $P < 10^{-4}$; Chi-square test) (Figure 3B). Type III systems appear as particularly over-represented in ICES/IMEs ($P < 10^{-4}$; Chi-square test). Type IV REases

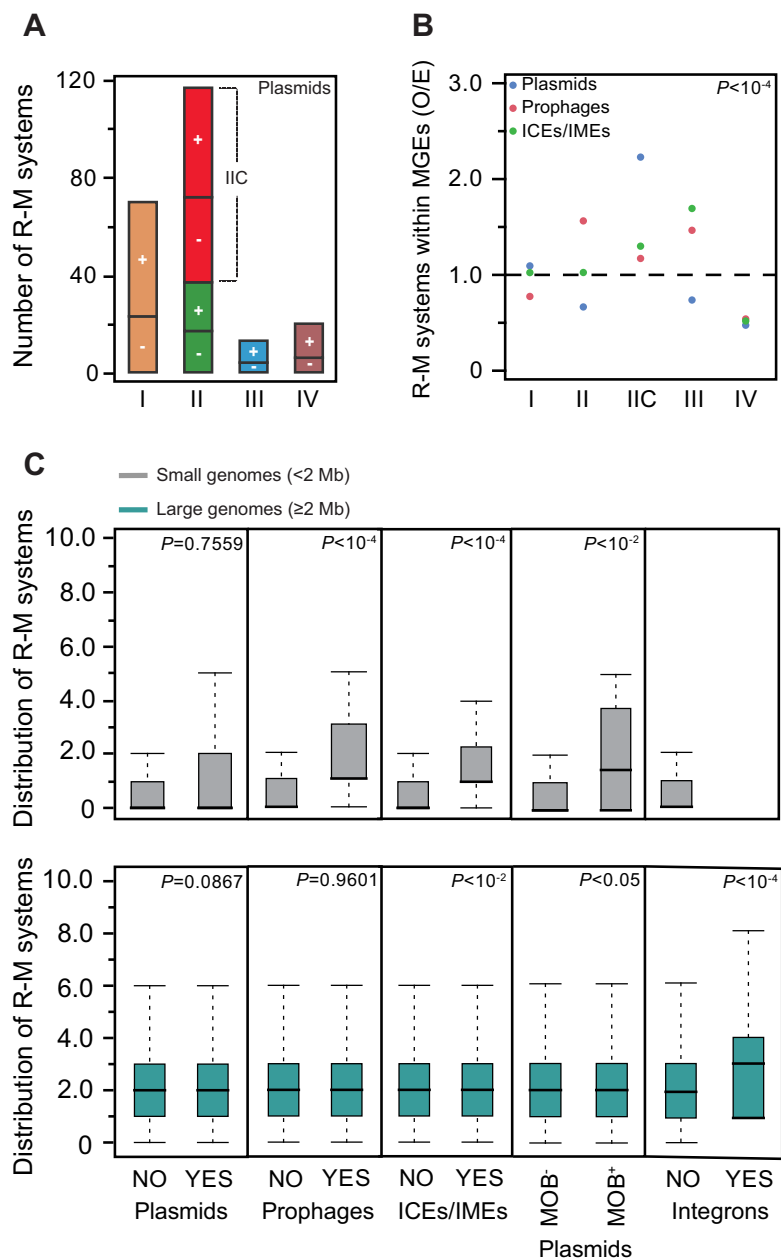


Figure 3. Quantification and distribution of R-M systems in MGEs. (A) Amount of Types I, II, IIC, III R-M systems and Type IV REases found in plasmids. The latter were classified according to their transmissibility: plasmids encoding the entire conjugation machinery or at least the relaxase (MOB⁺, shown as +), and plasmids lacking even the relaxase (MOB⁻, shown as -). (B) Observed/expected (O/E) ratios of R-M systems in plasmids, prophages and ICEs/IMEs. Expected values were obtained by multiplying the total number of each type of R-M system by the fraction of R-M systems assigned to each MGE. (C) Co-occurrence of R-M systems and MGEs. Box plots of the genomic co-occurrence of R-M systems with plasmids, prophages, ICEs/IMEs and integrons in small (<2 Mb) and large (≥ 2 Mb) genomes. Error bars represent standard deviations. Mann–Whitney–Wilcoxon test *P* values are indicated.

are under-represented in all MGEs, which is consistent with their role in defense against invading epigenetic DNA methylation systems (41). The most compact systems (Type IIC) are overabundant in all MGEs, and especially in plasmids. Accordingly, genomes known to harbor a very large number of plasmids (e.g. those from Spirochaetes) include many Type IIC systems (as seen before in Figure 1B). In fact, Type IIC systems found in Spirochaete plasmids comprise roughly 26% of all Type IIC systems detected in plas-

mids. In prophages, R-M systems are rare, but we still observed a significant over-representation of Type II systems. These systems are able to stabilize genetic elements by post-segregational killing and they could favor the stabilization of the lysogenic state as well as protect the host against infection by other phages.

Since MGEs appear to carry few R-M systems, we decided to quantify the association between the number of R-M systems and the presence/absence of the different MGEs

in genomes. Large genomes (≥ 2 Mb) show no strong association between the number of R-M systems and the presence or absence of plasmids (independently of being MOB⁻ or MOB⁺), prophages and ICEs/IMEs (Figure 3C and Supplementary Figure S5B). Only integrons are more likely to be found in large genomes with more R-M systems. This latter observation is in agreement with previous works suggesting that R-M systems stabilize super-integrons (76). Among small genomes (< 2 Mb), the number of R-M systems is positively correlated with the presence of prophages, ICEs/IMEs and MOB⁺ plasmids (Figure 3C). These results are consistent with the ones on the abundance of R-M systems in prokaryotic genomes, even when integron-, prophage- and plasmid-associated R-M systems are excluded (Supplementary Figure S5C). Overall, these data suggest that the abundance of R-M systems is indeed associated with genome size and the presence of MGEs because they are both associated with higher rates of horizontal transfer.

Co-occurrence and co-localization of R-M systems

The previous results suggest that bacteria enduring extensive HGT are more likely to encode R-M systems. We therefore analyzed the association of R-M systems with other systems dedicated to the control of MGEs. We started by analyzing the co-occurrence of the different types of R-M systems in genomes. We found significant co-occurrences for Type I R-M systems and Type IV REases ($P < 10^{-3}$; Chi-square test), Types I and III ($P < 10^{-4}$; Chi-square test) and Type IIC and Type IV REases ($P < 10^{-3}$; Chi-square test). Co-occurrence could result from selection for a diversity of R-M systems in a genome or from the presence of the so-called defense islands (5,20). To test between these hypotheses we computed the number of R-M systems occurring at less than 10 genes apart in genomes. We found that only 10.0% of the R-M systems were this close in genomes (11.8% in genomes encoding at least two R-M systems). When we increased the neighborhood to 50 genes, the frequency of co-occurring systems only increased to 18.3%. Therefore, while we confirm previous observations of clustering of R-M systems in genomes (5,20), this affects a relatively small number of systems. We then analyzed systematically the pairs of close R-M systems (< 10 genes apart). Type IV REases were found to often co-localize with other R-M systems (average inter-system distance = 3.2 ± 1.8 genes) ($P < 10^{-4}$; Chi-square test) (Figure 4A). Co-localization was particularly striking with Type I systems ($P < 10^{-4}$; Chi-square test) (Figure 4A), as previously observed (41). With the exceptions of Type IIC with Type IV, and Type II with itself, the remaining systems did not show significant co-localization patterns.

Co-occurrence of R-M systems with other defense systems

CRISPR-Cas systems provide acquired immunity against viruses and other MGEs, being present in most archaeal ($\sim 90\%$) and in a significant portion ($\sim 40\%$) of the bacterial species for which genomes are available (77,78). CRISPRs are arrays of 24–28-bp direct repeats, separated by short unique sequences acquired from past infections (spacers) localized close to a cluster of *cas* genes, which form the basis

for their specificity in the immune response. It has been recently reported that Type II CRISPR-Cas immune systems and Type II R-M systems work synergistically to prevent infection by phages (21). Indeed, we found that genomes encoding R-M systems are more likely to encode CRISPR-Cas systems ($P < 10^{-4}$; Mann–Whitney–Wilcoxon test), for both large and small genomes (both $P < 10^{-4}$; Mann–Whitney–Wilcoxon test) (Figure 4B). Many prokaryotes also encode homologs of the Argonaute (ARGO)-PIWI family of proteins, which have been recently found to be a bacterial defense system against MGEs (79,80). However, there is no significant co-occurrence of ARGOs and R-M systems in genomes ($P > 0.05$; Chi-square test, for both large and small genomes), even if ARGOs and CRISPR-Cas significantly co-occur ($P < 10^{-4}$; Chi-square test). It should be noted that only 5.3% of all CRISPR-Cas systems co-occur in a proximity of ± 25 genes of R-M systems and ARGO systems. The observation that co-occurrence of systems in the genomes is rarely associated with close co-localization in the genome suggests that co-occurrence is not caused by co-transfer, co-regulation in neighboring operons or co-occurrence in ‘defense islands’.

Since R-M systems are also involved in plasmid control, we hypothesized that CRISPR-Cas systems might target incoming R-M systems to prevent infection by plasmids. We tested this hypothesis by searching for sequence similarity between 80 685 CRISPR spacers identified in 1068 genomes containing such systems and our data set of R-M genes. We found only nine spacers (0.01% of the initial subset) with $\geq 90\%$ coverage and identity with R-M systems (one single spacer with 100% identity). The same analysis performed with solitary REases and MTases resulted in only 14 (90% identity) and 9 (100% identity) spacers (see the Materials and Methods section and Supplementary Table S3 for details). These results show that R-M systems are rarely targeted by CRISPR spacers.

Evolution of R-M systems

To gain insight into the evolutionary history of R-M systems, we built pan-genomes for a set of 43 bacterial species (Supplementary Table S1). Pan-genomes can be defined as the total gene repertoire of a set of strains of a given species, being composed of a core-genome harboring genes present in all strains, and an accessory (or dispensable) genome containing the remaining genes. We found only $\sim 4\%$ of the R-M systems in the core-genome (Supplementary Figure S6). This tendency was common to all R-M types. The large majority of gene families ($\sim 80\%$) are present in less than 1/3 of the strains (non-persistent genes), suggesting that they have been recently horizontally transferred.

The rapid turnover of R-M systems in bacteria fits previous suggestions that R-M systems are rapidly lost because they are not selected for (35). To clarify this point in a population genetics setup, we inferred the pairwise non-synonymous (dN) and synonymous (dS) substitution rates as well as the dN/dS ratio for all REase and MTase genes of each type of R-M system. A dN/dS ratio exceeding 1 is expected if natural selection promotes diversification in the protein sequence (adaptive or diversifying selection), whereas a ratio below 1 is expected when the major-

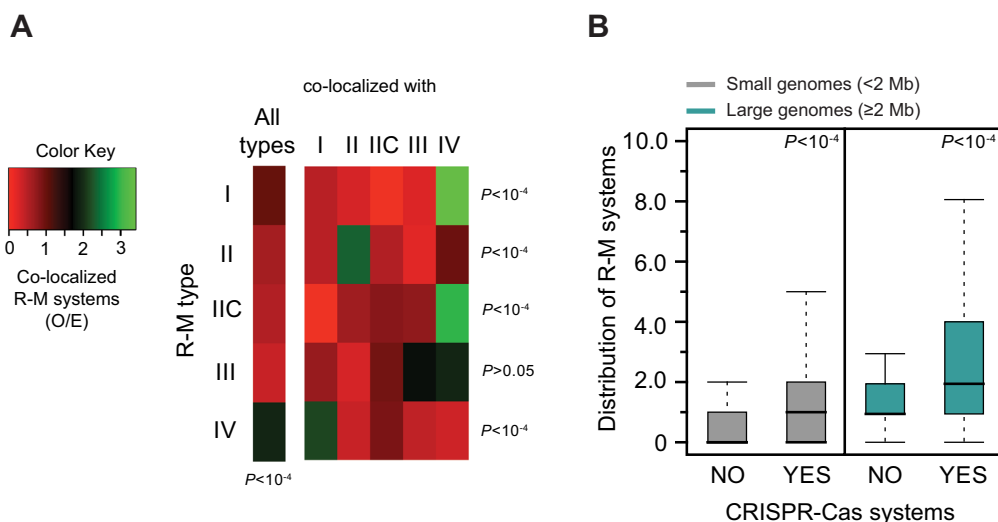


Figure 4. Associations between R-M and other defense systems. (A) Contextual analysis of R-M systems. Observed/expected (O/E) ratios of co-localized R-M systems (all types) and individual R-M types. Expected values were obtained by multiplying the total number of each type of R-M system by the fraction of R-M systems assigned to each MGE. (B) Box plots representing the co-occurrence of R-M and CRISPR-Cas systems in small (<2 Mb) and large (≥2 Mb) genomes. Error bars represent standard deviations. Mann–Whitney–Wilcoxon test P values are indicated.

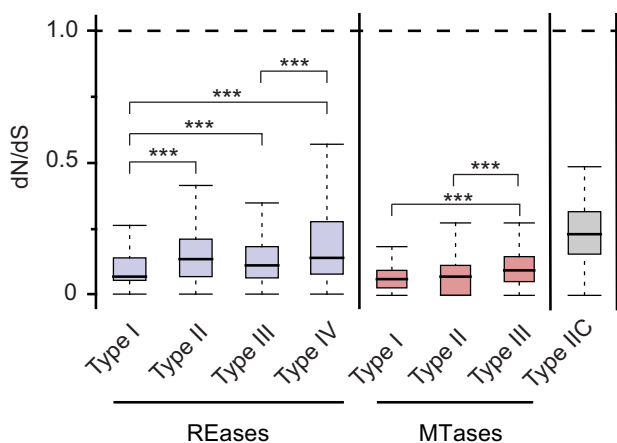


Figure 5. Evolution of R-M systems. Variation in dN/dS between REases, MTases and Type IIC systems. Error bars represent standard deviations. Significance was determined by computing Mann–Whitney–Wilcoxon test P values. For the sake of simplicity, we only show intra-REase and intra-MTase statistics. All remaining pairwise statistical comparisons are significant at $P < 10^{-3}$ with the exception of Type I REases-Type III MTases ($P < 0.05$) and Type III REases-Type III MTases ($P < 10^{-2}$). *** $P < 10^{-3}$.

ity of non-synonymous changes are deleterious and systematically purged by natural selection (purifying selection). A dN/dS close to 1 suggests that most mutations are not under selection (neutrality). Our results show that both MTases and REases in all types of R-M systems are under strong purifying selection (dN/dS \ll 1; Figure 5 and Supplementary Figure S7). The observation of purifying selection on MTases is expected as long as the REase is active. Yet this is clear evidence against hypotheses that REases could evolve neutrally or under selection for inactivation once in prokaryotic genomes.

Not all enzymes evolve at similar rates (Figure 5). MTases show stronger purifying selection than REases, i.e. lower dN/dS, for all systems. This might explain why traditionally MTases are more conserved and easier to identify by sequence similarity (32). The dN/dS values of the different systems are also variable. We observed the highest dN/dS values for Types IIC and IV and the lowest for Type I. Interestingly, this suggests an association between the structural complexity of R-M systems and the intensity of purifying selection. Type IV and Type II proteins do not interact with other proteins and the former and Type IIC act alone, i.e. the entire system is composed of a single protein. The ensuing weaker structural constraints, the simpler co-evolution of restriction sites, and one single R-M protein might allow faster evolution. On the opposite extreme, Type I systems, which constitute the largest protein complexes and are thus presumably more structurally constrained, show the lowest dN/dS values.

Solitary R-M genes

Previous works have found many solitary MTases and some solitary REases in genomes, suggesting they result from the genetic degradation of intact systems. Partial loss of R-M systems would lead to solitary MTases or REases that could eventually become domesticated by the host genome (35,36). Our observation confirms that components of solitary and complete systems are homologous, since we found more than 5000 solitary components in genomes using the same protein profiles and REBASE data that we used for the complete systems (Table 1). Yet demonstration of homology is not sufficient to suggest that solitary genes derive from complete systems. We have shown above that the vast majority of complete systems were not on the core-genome, suggesting they were very often acquired after the last speciation event. Therefore, to test the hypothesis that solitary components derive from complete systems one needs

to show that this occurs at the level of the species. For this, we computed the number of protein families in the pan-genomes including complete R-M systems, solitary genes or both. If solitary genes often resulted from the degradation of complete systems, one would expect to find both types of elements in many pan-genome families. Instead, from the 525 (358) protein families of the pan-genomes encoding a complete R-M system (solitary component), only 9% (13%) also included a protein encoded by a solitary gene in another genome (note that sequencing and annotation errors will tend to inflate these numbers). Intriguingly, this observation suggests that most solitary R-M genes do not originate from the recent degradation of complete R-M systems.

If solitary R-M genes were derived from genetic degradation of R-M systems arising by horizontal transfer then they should be encoded in MGEs ongoing genetic degradation. Hence, MGEs carrying solitary systems should be smaller than those encoding complete R-M systems. To test this hypothesis, we computed the distance between the two flanking core genes surrounding solitary R-M genes and complete R-M systems. We found that for ~80% of the species containing non-persistent R-M proteins, this distance was smaller for complete systems than for solitary proteins ($P < 10^{-2}$; Binomial test) (Figure 6A). Accordingly, solitary elements are very abundant in large MGEs such as phages and conjugative elements, whereas we showed above that complete systems are rare in these elements. In phages we identified 155 solitary genes (Table 1 and Figure 6B), even though phages encode very few R-M systems. Plasmids containing only solitary R-M proteins are larger than those carrying complete R-M systems (median sizes of 106 and 50 kb, respectively, $P < 10^{-4}$, Mann–Whitney–Wilcoxon test) (Figure 6C). These results strongly suggest that R-M systems and solitary proteins are transferred independently through distinct MGEs and/or transfer mechanisms. Additionally, this implies that solitary components of R-M systems are not systematically part of an ongoing genetic degradation process. Instead, it suggests solitary components are acquired as such by bacterial chromosomes.

To inquire on the nature of such adaptive functions, we compared the numbers of solitary REases and MTases. Solitary REases and MTases might be components of R-M systems encoded apart in the genome. Several of these have been found (35,36,81). In this case, one would expect a similar number of solitary REases and MTases. If solitary elements arose from random degradation of complete R-M systems, one would also expect to observe nearly as many solitary MTases as REases in the genome. On the other hand, solitary MTases and REases are known to have different impact in the cell fitness (36,52). Mutations leading to loss of the MTase while keeping the REase functional and expressed may be lethal if the restriction site is available in the genome. Accordingly, it has been shown that in bacteria solitary MTases are much more abundant than solitary REases (35). It is possible that the remaining REases are either associated with MTases in *trans* or are not expressed. Phages also show a strong predominance of solitary MTases over REases (143 against 12, $P < 0.05$, Chi-square test) (Table 1 and Figure 6B). Interestingly, phages over-represent solitary MTases (REases) four (two) times more than bacterial chromosomes ($P < 10^{-4}$ and $P < 10^{-2}$,

respectively, Chi-square test). This suggests selection for the presence of solitary R-M genes in phages, and especially MTases. MTases with broad sequence specificities might be selected to avoid restriction by the host at the moment of infection. However, solitary MTases were found to be much more abundant in temperate than in virulent phages ($P < 10^{-3}$; Chi-square test), and more abundant within prophages effectively present in the genomes than in the temperate phages of GenBank ($P < 10^{-2}$; Chi-square test) (Table 1). This suggests that solitary MTases have some additional adaptive role in lysogeny.

DISCUSSION

In line with many previous studies, we have shown here that R-M systems are nearly ubiquitous in Prokaryotes. We have also observed that their absolute abundance varies widely between phyla, within the limits of such an analysis using a database of bacterial genomes that is biased toward Proteobacteria and Firmicutes. Analysis of metagenomic data may allow in the future assessing if variations also occur between environments or clades that yet lack sequenced genomes. In the present data set, only the smallest genomes systematically lack R-M systems. These genomes typically correspond to sexually isolated endosymbiotic bacteria enduring very little or no HGT. This suggests that the link between genome size and HGT for small genomes is caused by the decreased frequency of HGT in many of the smaller genomes. Accordingly, the small genomes of bacteria that endure HGT, like the Tenericutes (including *Mycoplasma*), show high densities of R-M systems. Interestingly, several *Mycoplasma* encode R-M systems engaging in phase variation, which should increase their diversity (82). In the genome size range up to 2 Mb, one finds a variety of bacteria with increasing rates of horizontal transfer. Bacteria with genomes of ~2 Mb like *Helicobacter*, *Haemophilus* or *Neisseria* are known to engage extensively in HGT. Larger genomes are expected to follow the same general trend within bacteria (2). For these genomes we did not observe a tendency of more R-M systems in larger genomes. Instead, the average number of R-M systems is kept approximately constant and equal to two systems per genome. R-M systems have been shown to decrease the probability of infection by naïve phages by values around 10^{-5} with strong variations from 10^{-2} to 10^{-7} (21,83,84). If the rate above is multiplicative, then two R-M systems will decrease the probability of phage infection by an average factor around 10^{-10} . This is a simplified calculation assuming unbiased phage genome sequences (85) and lack of modification of the phage DNA (86,87), but under these conditions, additional R-M systems should provide little additional capacity of defense against MGEs. Moreover, they will increase the cost of genome methylation and the probability of accidental restriction of the chromosome. Hence, the presence of large numbers of R-M systems in some genomes likely requires additional explanations, for example the R-M addictive behavior and/or high rates of transfer of R-M systems.

Our work suggests that the distribution and evolution of R-M systems differs between types. We found a large number of Type IIC systems. To test if these might have arisen

Table 1. Numbers and densities (per element per Mb) of solitary MTases and REases found in chromosomes, plasmids, prophages and phages (temperate and virulent)

	# Solitary MTases (dens./element/Mb)	# Solitary REases (dens./element/Mb)
Chromosomes ^a (2342)	3966 (0.593)	731 (0.116)
Plasmids (<i>n</i> = 1787)	292 (1.58)	23 (0.182)
Prophages (<i>n</i> = 2827)	717 (5.44)	15 (0.096)
Phages (<i>n</i> = 831)	143 (2.47)	12 (0.230)
Temperate phages (<i>n</i> = 311)	56 (3.55)	5 (0.290)
Virulent phages (<i>n</i> = 520)	87 (1.85)	7 (0.196)

^aResults shown for chromosomes do not include R-M systems located in prophages. *Helicobacter* genomes were not considered.

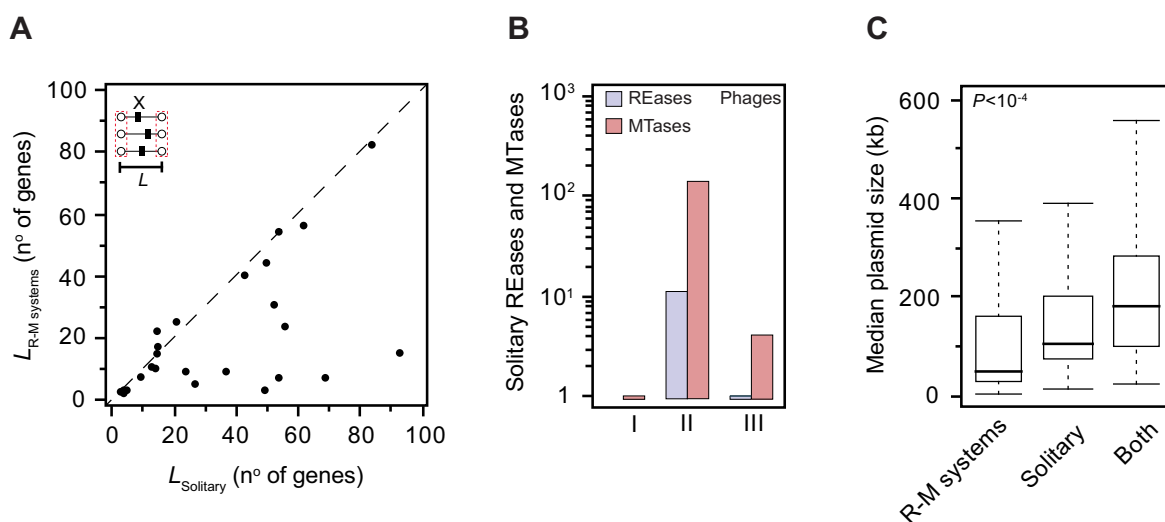


Figure 6. Comparative analysis of complete R-M systems and solitary components. (A) Median size of regions (L , expressed in the number of genes) harboring complete volatile R-M systems versus solitary volatile R-M elements (indicated as X). Stippled line corresponds to the identity. (B) The number of solitary REases and MTases in phages. Over 90% of the total hits were found to correspond to solitary MTases. (C) Median plasmid size (kb) for plasmids containing only complete R-M systems, solitary components or both. Mann–Whitney–Wilcoxon test P value is indicated next to the box plots.

by fusion of head-to-tail-oriented REase and MTase genes in a single uninterrupted hybrid polypeptide we identified the Type IIC systems sharing medium to high similarity (>50%) with Type II systems encoded in two genes. Only 0.9% of the Type IIC systems have such level of similarity with other Type II systems (found mainly in Firmicutes), which suggests that such fusions are rare. Type IIC systems are more compact and their sequence specificity is linked in the same peptide with the REase and MTase functions. This might allow them to evolve new specificities faster (88), which is compatible with the observed rapid sequence evolution. Both rapid evolution and compactness might explain why they are more often encoded in MGEs. It is also possible that type IIC systems are more efficient at establishing in new hosts.

Some types of R-M systems co-occur with others more often than expected. Type IV REases are often encoded close to Type I R-M systems (Figure 4A). This conformation allows the degradation of unmethylated DNA recognized by the Type I and of modified DNA recognized by the Type IV system. Most Type I MTases methylate adenines to *N*6-methyladenines (m6A) (89). On the other hand, several known Type IV REases do not recognize m6A (12). The

complementarity between the two systems might favor their clustering. It might also favor the evolution of broad substrate specificities in Type IV REases as long as this does not lead to the degradation of the DNA modified by the co-localized Type I system. The absence of selection for very specific sequence recognition might lead to more relaxed selection for Type IV REases. This is in agreement with the higher dN/dS ratios observed for these REases (Figure 5).

A number of works have shown that R-M systems can be stabilized in genomes and have an impact on the host genome composition (90–92). We observed relatively few R-M systems in plasmids, some in prophages, and practically none in phages. On the other hand, all these MGEs encode a large number of solitary R-M genes, notably MTases. While these results suggest that R-M systems may be used by MGEs to stabilize their presence in hosts, this occurs rarely in our data set. In contrast, we found that MGEs very often encode solitary MTases. These may serve as antidotes against R-M systems and thereby facilitate infection of new hosts and competition with other MGEs. Solitary MTases were suggested to result from complete R-M systems by loss of the gene encoding the REase (35). However, families of pan-genomes either include solitary genes

or complete systems, but rarely both. Also, solitary MTases are typically transferred by larger MGEs than complete R-M systems. These results suggest that solitary genes arrive more frequently in genomes by HGT than by *in situ* genetic degradation of complete systems. Therefore, complete R-M systems and solitary R-M genes are largely independent sets of genes.

One intriguing question that remains to be clarified in more detail is how can R-M systems show such a high turnover in genomes when they are so poorly represented in MGEs? This question can be subdivided into two more specific ones: what is the source of the new R-M systems? Why are they so often lost? One possible explanation for the first question relies on the ability of certain R-M systems to behave as mobile units *per se*, sometimes generating extensive genomic rearrangements upon their insertion (27,93,94). Here we have observed that the majority of the acquired regions containing R-M systems are typically small (Figure 6A), suggesting that R-M mobility may be less dependent on MGEs and more dependent, for example, on the existence of small genomic integration hotspots. It is also possible that R-M systems frequently exploit other mechanisms such as natural transformation, vesicles, nanotubes, gene transfer agents or generalized transduction in order to move between genomes (73,95). This possibility is backed up by our data showing a higher number of R-M systems in competent than in non-competent bacterial hosts (Figure 2).

The frequent loss of R-M systems is apparently inconsistent with the observed strong selection against non-synonymous changes. These two observations may be reconciled if the selection pressure on the system fluctuates in time, i.e. if R-M systems alternate periods of strong purifying selection and periods of relaxed selection. The former would lead to the purge of non-synonymous changes and low dN/dS. The latter would lead to rapid gene loss. Relaxed selection might occur when there are many other R-M systems in the genome, especially if these have the same sequence specificity. In this case, there is competition between R-M systems resulting in relaxed selection for their maintenance in the genome. It could also occur when there is strong selection for HGT, e.g. in moments of stress, or when population cycles lead to the fixation of slightly deleterious changes, e.g. small population sizes or selective sweeps.

Our genome comparison analysis provides new insight into the intricate relationships between MGEs, R-M systems and other cell defensive systems. We found that such relationships are complex and depend on the type of R-M system and MGE involved. As a further intriguing novel feature, we observed that solitary R-M components and complete systems are essentially independent sets of genes. The growing access to bacterial methylome data will allow for a more comprehensive understanding of methylation specificity and how it affects bacteria genetic diversification and protection from MGEs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

The authors wish to thank Sophie Abby (Institut Pasteur) for help with MacSyFinder, and the anonymous reviewers for suggestions on the manuscript.

FUNDING

European Research Council [EVOMOBILOME n°281605]. Funding for open access charge. European Research Council [EVOMOBILOME n°281605].

Conflict of interest statement. None declared.

REFERENCES

- Gogarten, J.P., Doolittle, W.F., and Lawrence, J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, **19**, 2226–2238.
- Treangen, T.J. and Rocha, E.P. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.*, **7**, e1001284.
- Thomas, C.M. and Nielsen, K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.*, **3**, 711–721.
- Labrie, S.J., Samson, J.E., and Moineau, S. (2010) Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.*, **8**, 317–327.
- Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
- Mruk, I. and Kobayashi, I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.*, **42**, 70–86.
- Loenen, W.A., Dryden, D.T., Raleigh, E.A., Wilson, G.G., and Murray, N.E. (2014) Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res.*, **42**, 3–19.
- Ishikawa, K., Handa, N., and Kobayashi, I. (2009) Cleavage of a model DNA replication fork by a Type I restriction endonuclease. *Nucleic Acids Res.*, **37**, 3531–3544.
- Roberts, R.J., Belfort, M., Bestor, T., Bhagwat, A.S., Bickle, T.A., Bitinaite, J., Blumenthal, R.M., Degtyarev, S., Dryden, D.T., and Dybvig, K. (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.
- Janulaitis, A., Vaisvila, R., Timinskas, A., Klimasauskas, S., and Butkus, V. (1992) Cloning and sequence analysis of the genes coding for Eco571 type IV restriction-modification enzymes. *Nucleic Acids Res.*, **20**, 6051–6056.
- Morgan, R.D., Dwinell, E.A., Bhatia, T.K., Lang, E.M., and Luyten, Y.A. (2009) The Mmel family: type II restriction-modification enzymes that employ single-strand modification for host protection. *Nucleic Acids Res.*, **37**, 5208–5221.
- Loenen, W.A. and Raleigh, E.A. (2014) The other face of restriction: modification-dependent enzymes. *Nucleic Acids Res.*, **42**, 56–69.
- Jeltsch, A. and Pingoud, A. (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.*, **42**, 91–96.
- Aras, R.A., Small, A.J., Ando, T., and Blaser, M.J. (2002) *Helicobacter pylori* interstrain restriction-modification diversity prevents genome subversion by chromosomal DNA from competing strains. *Nucleic Acids Res.*, **30**, 5391–5397.
- Nobusato, A., Uchiyama, I., and Kobayashi, I. (2000) Diversity of

- restriction-modification gene homologues in *Helicobacter pylori*. *Gene*, **259**, 89–98.
16. Furuta, Y., Namba-Fukuyo, H., Shibata, T.F., Nishiyama, T., Shigenobu, S., Suzuki, Y., Sugano, S., Hasebe, M., and Kobayashi, I. (2014) Methylome diversification through changes in DNA methyltransferase sequence specificity. *PLoS Genet.*, **10**, e1004272.
 17. Vale, F.F., Megraud, F., and Vitor, J.M. (2009) Geographic distribution of methyltransferases of *Helicobacter pylori*: evidence of human host population isolation and migration. *BMC Microbiol.*, **9**, 193.
 18. Budroni, S., Siena, E., Dunning Hotopp, J.C., Seib, K.L., Serruto, D., Nofroni, C., Comanducci, M., Riley, D.R., Daugherty, S.C., and Angiuoli, S.V. *et al.* Budroni, S., Siena, E., Dunning Hotopp, J.C., Seib, K.L., Serruto, D., Nofroni, C., Comanducci, M., Riley, D.R., Daugherty, S.C., and Angiuoli, S.V. (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 4494–4499.
 19. Arber, W. and Linn, S. (1969) DNA modification and restriction. *Annu. Rev. Biochem.*, **38**, 467–500.
 20. Makarova, K.S., Wolf, Y.I., Snir, S., and Koonin, E.V. (2011) Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J. Bacteriol.*, **193**, 6039–6056.
 21. Dupuis, M.E., Villion, M., Magadan, A.H., and Moineau, S. (2013) CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat. Commun.*, **4**, 2087.
 22. Naito, T., Kusano, K., and Kobayashi, I. (1995) Selfish behavior of restriction-modification systems. *Science*, **267**, 897–899.
 23. Kulakauskas, S., Luby, A., and Ehrlich, S.D. (1995) DNA restriction-modification systems mediate plasmid maintenance. *J. Bacteriol.*, **177**, 3451–3454.
 24. Takahashi, N., Ohashi, S., Sadykov, M.R., Mizutani-Ui, Y., and Kobayashi, I. (2011) IS-linked movement of a restriction-modification system. *PLoS ONE*, **6**, e16554.
 25. Mochizuki, A., Yahara, K., Kobayashi, I., and Iwasa, Y. (2006) Genetic addiction: selfish gene's strategy for symbiosis in the genome. *Genetics*, **172**, 1309–1323.
 26. Betlach, M., Hershfield, V., Chow, L., Brown, W., Goodman, H., and Boyer, H.W. (1976) A restriction endonuclease analysis of the bacterial plasmid controlling the *ecoRI* restriction and modification of DNA. *Fed. Proc.*, **35**, 2037–2043.
 27. Furuta, Y. and Kobayashi, I. (2013) Restriction-modification systems as mobile genetic elements. In: *Bacterial Integrative Mobile Genetic Elements*, Roberts, AP and Mullany, P. Eds. Landes Bioscience, Austin, TX. pp. 85–103.
 28. Kita, K., Kawakami, H., and Tanaka, H. (2003) Evidence for horizontal transfer of the EcoT381 restriction-modification gene to chromosomal DNA by the P2 phage and diversity of defective P2 prophages in *Escherichia coli* TH38 strains. *J. Bacteriol.*, **185**, 2296–2305.
 29. Burrus, V., Bontemps, C., Decaris, B., and Guedon, G. (2001) Characterization of a novel type II restriction-modification system, Sth3681, encoded by the integrative ICESt1 of *Streptococcus thermophilus* CNRZ368. *Appl. Environ. Microbiol.*, **67**, 1522–1528.
 30. Rowe-Magnus, D.A., Guerout, A.M., Ploncard, P., Dychinco, B., Davies, J., and Mazel, D. (2001) The evolutionary history of chromosomal super-integrations provides an ancestry for multiresistant integrations. *Proc. Natl Acad. Sci. U.S.A.*, **98**, 652–657.
 31. Kobayashi, I., Nobusato, A., Kobayashi-Takahashi, N., and Uchiyama, I. (1999) Shaping the genome—restriction-modification systems as mobile genetic elements. *Curr. Opin. Genet. Dev.*, **9**, 649–656.
 32. Vasu, K. and Nagaraja, V. (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol. Mol. Biol. Rev.*, **77**, 53–72.
 33. Srikhanta, Y.N., Fox, K.L., and Jennings, M.P. (2010) The phasevarion: phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes. *Nat. Rev. Microbiol.*, **8**, 196–206.
 34. Vitoriano, I., Vitor, J.M., Oleastro, M., Roxo-Rosa, M., and Vale, F.F. (2013) Proteome variability among *Helicobacter pylori* isolates clustered according to genomic methylation. *J. Appl. Microbiol.*, **114**, 1817–1832.
 35. Seshasayee, A.S., Singh, P., and Krishna, S. (2012) Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.*, **40**, 7066–7073.
 36. Ershova, A.S., Karyagina, A.S., Vasiliev, M.O., Lyashchuk, A.M., Lunin, V.G., Spirin, S.A., and Alexeevski, A.V. (2012) Solitary restriction endonucleases in prokaryotic genomes. *Nucleic Acids Res.*, **40**, 10107–10115.
 37. Mokrishcheva, M.L., Solonin, A.S., and Nikitin, D.V. (2011) Fused *eco29kIR*- and *M* genes coding for a fully functional hybrid polypeptide as a model of molecular evolution of restriction-modification systems. *BMC Evol. Biol.*, **11**, 35.
 38. Liang, J. and Blumenthal, R.M. (2013) Naturally-occurring, dually-functional fusions between restriction endonucleases and regulatory proteins. *BMC Evol. Biol.*, **13**, 218.
 39. Furuta, Y. and Kobayashi, I. (2012) Movement of DNA sequence recognition domains between non-orthologous proteins. *Nucleic Acids Res.*, **40**, 9218–9232.
 40. Furuta, Y., Kawai, M., Uchiyama, I., and Kobayashi, I. (2011) Domain movement within a gene: a novel evolutionary mechanism for protein diversification. *PLoS ONE*, **6**, e18819.
 41. Fukuda, E., Kaminska, K.H., Bujnicki, J.M., and Kobayashi, I. (2008) Cell death upon epigenetic genome methylation: a novel function of methyl-specific deoxyribonucleases. *Genome Biol.*, **9**, R163.
 42. Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2010) REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.
 43. Watson, M., O.C.M., Bottacini, D., Clark, F., Roberts, T.A., Korch, R.J., Garault, J., Chervaux, P., van Hylckama, C., Vlieg, J.E., and Smokvina, T. *et al.* Watson, M., O.C.M., Bottacini, D., Clark, F., Roberts, T.A., Korch, R.J., Garault, J., Chervaux, P., van Hylckama, C., Vlieg, J.E., and Smokvina, T. (2014) Identification of restriction-modification systems of *Bifidobacterium animalis* subsp. *lactis* CNCM I-2494 by SMRT sequencing and associated methylome analysis. *PLoS ONE*, **9**, e94875.
 44. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
 45. Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
 46. Gouy, M., Guindon, S., and Gascuel, O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.*, **27**, 221–224.
 47. Finn, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
 48. Pingoud, A., Wilson, G.G., and Wende, W. (2014) Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res.*, **42**, 7489–7527.

49. Kobayashi, I. Pingoud, A. Ed. Kobayashi, I. Kobayashi, I. (2004) Restriction-modification systems as minimal life forms. In: *Restriction Endonucleases*, Pingoud, A. Ed. Springer-Verlag, Berlin, Germany. pp. 19–62.
50. Schouler, C., Gautier, M., Ehrlich, S.D., and Chopin, M.C. Schouler, C., Gautier, M., Ehrlich, S.D., and Chopin, M.C. (1998) Combinational variation of restriction modification specificities in *Lactococcus lactis*. *Mol. Microbiol.*, **28**, 169–178.
51. Tsuru, T., Kawai, M., Mizutani-Ui, Y., Uchiyama, I., and Kobayashi, I. Tsuru, T., Kawai, M., Mizutani-Ui, Y., Uchiyama, I., and Kobayashi, I. (2006) Evolution of paralogous genes: reconstruction of genome rearrangements through comparison of multiple genomes within *Staphylococcus aureus*. *Mol. Biol. Evol.*, **23**, 1269–1285.
52. Takahashi, N., Naito, Y., Handa, N., and Kobayashi, I. Takahashi, N., Naito, Y., Handa, N., and Kobayashi, I. (2002) A DNA methyltransferase can protect the genome from postdisturbance attack by a restriction-modification gene complex. *J. Bacteriol.*, **184**, 6100–6108.
53. Miele, V., Penel, S., and Duret, L. Miele, V., Penel, S., and Duret, L. (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics*, **12**, 116.
54. Suyama, M., Torrents, D., and Bork, P. Suyama, M., Torrents, D., and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.
55. Yang, Z. Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
56. Yang, Z. and Nielsen, R. Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, **17**, 32–43.
57. Bobay, L.M., Touchon, M., and Rocha, E.P. Bobay, L.M., Touchon, M., and Rocha, E.P. (2013) Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. *PLoS Genet.*, **9**, e1003825.
58. Guglielmini, J., Quintais, L., Garcillan-Barcia, M.P., de la Cruz, F., and Rocha, E.P. Guglielmini, J., Quintais, L., Garcillan-Barcia, M.P., de la Cruz, F., and Rocha, E.P. (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.*, **7**, e1002222.
59. Cambray, G., Guerout, A.M., and Mazel, D. Cambray, G., Guerout, A.M., and Mazel, D. (2010) Integrons. *Annu. Rev. Genet.*, **44**, 141–166.
60. Touchon, M. and Rocha, E.P. Touchon, M. and Rocha, E.P. (2010) The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE*, **5**, e11126.
61. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholz, P. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholz, P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
62. Seitz, P. and Blokesch, M. Seitz, P. and Blokesch, M. (2013) DNA-uptake machinery of naturally competent *Vibrio cholerae*. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 17987–17992.
63. Petersen, F.C., Tao, L., and Scheie, A.A. Petersen, F.C., Tao, L., and Scheie, A.A. (2005) DNA binding-uptake system: a link between cell-to-cell communication and biofilm formation. *J. Bacteriol.*, **187**, 4392–4400.
64. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., and Bouvet, O. et al. Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., and Bouvet, O. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.*, **5**, e1000344.
65. Rocha, E.P., Touchon, M., and Feil, E.J. Rocha, E.P., Touchon, M., and Feil, E.J. (2006) Similar compositional biases are caused by very different mutational effects. *Genome Res.*, **16**, 1537–1547.
66. Eddy, S.R. Eddy, S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
67. Xu, Q., Morgan, R.D., Roberts, R.J., and Blaser, M.J. Xu, Q., Morgan, R.D., Roberts, R.J., and Blaser, M.J. (2000) Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proc. Natl Acad. Sci. U.S.A.*, **97**, 9671–9676.
68. Lin, L.F., Posfai, J., Roberts, R.J., and Kong, H. Lin, L.F., Posfai, J., Roberts, R.J., and Kong, H. (2001) Comparative genomics of the restriction-modification systems in *Helicobacter pylori*. *Proc. Natl Acad. Sci. U.S.A.*, **98**, 2740–2745.
69. Marena, M., Barbe, V., Gourgues, G., Mangenot, S., Sagne, E., and Citti, C. Marena, M., Barbe, V., Gourgues, G., Mangenot, S., Sagne, E., and Citti, C. (2006) A new integrative conjugative element occurs in *Mycoplasma agalactiae* as chromosomal and free circular forms. *J. Bacteriol.*, **188**, 4137–4141.
70. Sirand-Pugnet, P., Lartigue, C., Marena, M., Jacob, D., Barre, A., Barbe, V., Schenowitz, C., Mangenot, S., Couloux, A., and Segurens, B. et al. Sirand-Pugnet, P., Lartigue, C., Marena, M., Jacob, D., Barre, A., Barbe, V., Schenowitz, C., Mangenot, S., Couloux, A., and Segurens, B. (2007) Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. *PLoS Genet.*, **3**, e75.
71. Chen, I. and Dubnau, D. Chen, I. and Dubnau, D. (2004) DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.*, **2**, 241–249.
72. Hofreuter, D., Odenbreit, S., and Haas, R. Hofreuter, D., Odenbreit, S., and Haas, R. (2001) Natural transformation competence in *Helicobacter pylori* is mediated by the basic components of a type IV secretion system. *Mol. Microbiol.*, **41**, 379–391.
73. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.P. (2014) Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.*, **12**, 181–196.
74. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A., and van Sinderen, D. Murphy, J., Mahony, J., Ainsworth, S., Nauta, A., and van Sinderen, D. (2013) Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.*, **79**, 7547–7555.
75. Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P., and de la Cruz, F. Smillie, C., Garcillan-Barcia, M.P., Francia, M.V., Rocha, E.P., and de la Cruz, F. (2010) Mobility of plasmids. *Microbiol. Mol. Biol. Rev.*, **74**, 434–452.
76. Rowe-Magnus, D.A., Guerout, A.M., Biskri, L., Bouige, P., and Mazel, D. Rowe-Magnus, D.A., Guerout, A.M., Biskri, L., Bouige, P., and Mazel, D. (2003) Comparative analysis of superintegrons: engineering extensive genetic diversity in the Vibrionaceae. *Genome Res.*, **13**, 428–442.
77. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J. van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.*, **34**, 401–407.
78. Sorek, R., Kunin, V., and Hugenholz, P. Sorek, R., Kunin, V., and Hugenholz, P. (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.
79. Makarova, K.S., Wolf, Y.I., van der Oost, J., and Koonin, E.V. Makarova, K.S., Wolf, Y.I., van der Oost, J., and Koonin, E.V. (2009) Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct.*, **4**, 29.
80. Swarts, D.C., Jore, M.M., Westra, E.R., Zhu, Y., Janssen, J.H., Snijders, A.P., Wang, Y., Patel, D.J., Berenguer, J., and Brouns, S.J. et al. Swarts, D.C., Jore, M.M., Westra, E.R., Zhu, Y., Janssen, J.H., Snijders, A.P., Wang, Y., Patel, D.J., Berenguer, J., and Brouns, S.J. (2014) DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*, **507**, 258–261.
81. Banerjee, S. and Chowdhury, R. Banerjee, S. and Chowdhury, R. (2006) An orphan DNA (cytosine-5)-methyltransferase in *Vibrio cholerae*. *Microbiology*, **152**, 1055–1062.
82. Dybvig, K., Sitaraman, R., and French, C.T. Dybvig, K., Sitaraman, R., and French, C.T. (1998) A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 13923–13928.
83. Blumenthal, R.M. and Cheng, X. Blumenthal, R.M. and Cheng, X. (2002) Restriction-modification systems. In: *Modern Microbial Genetics*, Streips, UN and Yasbin, RE. Eds. Blumenthal, R.M. and Cheng, X. 2nd ed. Blumenthal, R.M. and Cheng, X. (2002) Restriction-modification systems. In: *Modern Microbial Genetics*, Streips, UN and Yasbin, RE. Eds. Wiley, NY. pp. 177–225.
84. Korona, R., Korona, B., and Levin, B.R. Korona, R., Korona, B., and Levin, B.R. (1993) Sensitivity of naturally occurring coliphages

- to type I and type II restriction and modification. *J. Gen. Microbiol.*, **139**(Pt 6), 1283–1290.
85. Krüger, D.H. and Bickle, T.A. Krüger, D.H. and Bickle, T.A. (1983) Bacteriophage survival. Multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts. *Microbiol. Rev.*, **47**, 345–360.
86. Warren, R.A. Warren, R.A. (1980) Modified bases in bacteriophage DNAs. *Annu. Rev. Microbiol.*, **34**, 137–158.
87. Korona, R. and Levin, B.R. Korona, R. and Levin, B.R. (1993) Phage-mediated selection for restriction-modification. *Evolution*, **47**, 565–575.
88. Rimseliene, R., Maneliene, Z., Lubys, A., and Janulaitis, A. Rimseliene, R., Maneliene, Z., Lubys, A., and Janulaitis, A. (2003) Engineering of restriction endonucleases: using methylation activity of the bifunctional endonuclease Eco57I to select the mutant with a novel sequence specificity. *J. Mol. Biol.*, **327**, 383–391.
89. Loenen, W.A., Dryden, D.T., Raleigh, E.A., and Wilson, G.G. Loenen, W.A., Dryden, D.T., Raleigh, E.A., and Wilson, G.G. (2014) Type I restriction enzymes and their relatives. *Nucleic Acids Res.*, **42**, 20–44.
90. Qian, L. and Kussell, E. Qian, L. and Kussell, E. (2012) Evolutionary dynamics of restriction site avoidance. *Phys. Rev. Lett.*, **108**, 158105.
91. Rocha, E.P., Danchin, A., and Viari, A. Rocha, E.P., Danchin, A., and Viari, A. (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.*, **11**, 946–958.
92. Handa, N., Nakayama, Y., Sadykov, M., and Kobayashi, I. Handa, N., Nakayama, Y., Sadykov, M., and Kobayashi, I. (2001) Experimental genome evolution: large-scale genome rearrangements associated with resistance to replacement of a chromosomal restriction-modification gene complex. *Mol. Microbiol.*, **40**, 932–940.
93. Furuta, Y., Abe, K., and Kobayashi, I. Furuta, Y., Abe, K., and Kobayashi, I. (2010) Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. *Nucleic Acids Res.*, **38**, 2428–2443.
94. Furuta, Y., Kawai, M., Yahara, K., Takahashi, N., Handa, N., Tsuru, T., Oshima, K., Yoshida, M., Azuma, T., and Hattori, M. *et al.* Furuta, Y., Kawai, M., Yahara, K., Takahashi, N., Handa, N., Tsuru, T., Oshima, K., Yoshida, M., Azuma, T., and Hattori, M. (2011) Birth and death of genes linked to chromosomal inversion. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 1501–1506.
95. Lang, A.S., Zhaxybayeva, O., and Beatty, J.T. Lang, A.S., Zhaxybayeva, O., and Beatty, J.T. (2012) Gene transfer agents: phage-like elements of genetic exchange. *Nat. Rev. Microbiol.*, **10**, 472–482.

Regulation of genetic flux between bacteria by restriction–modification systems

Pedro H. Oliveira^{a,b,1}, Marie Touchon^{a,b}, and Eduardo P. C. Rocha^{a,b}

^aMicrobial Evolutionary Genomics, Institut Pasteur, 75015 Paris, France; and ^bCNRS, UMR 3525, 75015 Paris, France

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved April 5, 2016 (received for review March 2, 2016)

Restriction–modification (R–M) systems are often regarded as bacteria’s innate immune systems, protecting cells from infection by mobile genetic elements (MGEs). Their diversification has been recently associated with the emergence of particularly virulent lineages. However, we have previously found more R–M systems in genomes carrying more MGEs. Furthermore, it has been suggested that R–M systems might favor genetic transfer by producing recombinogenic double-stranded DNA ends. To test whether R–M systems favor or disfavor genetic exchanges, we analyzed their frequency with respect to the inferred events of homologous recombination and horizontal gene transfer within 79 bacterial species. Genetic exchanges were more frequent in bacteria with larger genomes and in those encoding more R–M systems. We created a recognition target motif predictor for Type II R–M systems that identifies genomes encoding systems with similar restriction sites. We found more genetic exchanges between these genomes, independently of their evolutionary distance. Our results reconcile previous studies by showing that R–M systems are more abundant in promiscuous species, wherein they establish preferential paths of genetic exchange within and between lineages with cognate R–M systems. Because the repertoire and/or specificity of R–M systems in bacterial lineages vary quickly, the preferential fluxes of genetic transfer within species are expected to constantly change, producing time-dependent networks of gene transfer.

homologous recombination | horizontal gene transfer | bacterial evolution

Prokaryotes evolve rapidly by acquiring genetic information from other individuals, often through the action of mobile genetic elements (MGEs) such as plasmids or phages (1). In bacterial population genetics, the events of gene transfer are usually termed horizontal gene transfer (HGT) when they result in the acquisition of new genes and homologous recombination (HR) when they result in allelic replacements. The distinction between the two evolutionary mechanisms (HGT and HR) is not always straightforward: incoming DNA may integrate the host genome by double crossovers at homologous regions, leading to allelic replacements in these regions and to the acquisition of novel genes in the intervening ones. HR takes place only between highly similar sequences, typically within species (2). As a result, it usually involves the exchange of few polymorphisms, eventually in multiple regions, between cells (3). It may also result in no change if the recombining sequences are identical, which leaves no traces and cannot be detected by sequence analysis. HGT may occur between distant species, resulting in the acquisition of many genes in a single event. The replication and maintenance of MGEs have fitness costs to the bacterial host and have led to the evolution of cellular defense systems. These systems can sometimes be counteracted by MGEs, leading to evolutionary arms races.

Restriction–modification (R–M) systems are some of the best known and the most widespread bacterial defense systems (4). They encode a methyltransferase (MTase) function that modifies particular DNA sequences in function of the presence of target recognition sites and a restriction endonuclease (REase) function that cleaves them when they are unmethylated (5). R–M systems are traditionally classified into three main types. Type II systems are by far the most abundant and the best studied (6).

With the exception of the subType IIC, they comprise MTase and REase functions encoded on separate genes and are able to operate independently from each other. R–M systems severely diminish the infection rate by MGEs and have been traditionally seen as bacteria’s innate immune systems (7). However, successful infection of a few cells generates methylated MGEs immune to restriction that can invade the bacterial population (8). Hence, R–M systems are effective as defense systems during short periods of time and especially when they are diverse across a population (9, 10). In particular, it has been suggested that they might facilitate colonization of new niches (11). Type II R–M systems are also addictive modules that can propagate selfishly in populations (12). Both roles of R–M systems, as defense or selfish systems, may explain why they are very diverse within species (13, 14). Accordingly, R–M systems endure selection for diversification and are rapidly replaced (15, 16).

Several recent large-scale studies of population genomics have observed more frequent HR within than between lineages (17, 18). This suggests that HR might favor the generation of cohesive population structures within bacterial species (19). Specific lineages of important pathogens that have recently changed their R–M repertoires show higher sexual isolation, such as *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Burkholderia pseudomallei*, and *Staphylococcus aureus* (20–22). For example, a Type I R–M system decreased transfer to and from a major methicillin-resistant *S. aureus* lineage (23). Diversification of R–M target recognition sites could thus reduce transfer between lineages with different systems while establishing preferential gene fluxes between those with R–M systems recognizing the same target motifs (cognate R–M). However, these results can be confounded by evolutionary distance: closely related genomes are more likely to encode similar R–M systems, inhabit the same environments (facilitating transfer between cells), and have

Significance

The role of restriction–modification (R–M) as bacteria’s innate immune system, and a barrier to sexual exchange, has often been challenged. Recent works suggested that the diversification of these systems might have driven the evolution of highly virulent bacterial lineages. Here, we showed that R–M systems were more abundant in species enduring more DNA exchanges and that within-species flux of genetic material was higher when cognate systems were present. Presumably, bacteria enduring frequent infections by mobile elements select for the presence of more numerous R–M systems, but rapid diversification of R–M systems leads to varying patterns of sexual exchanges between bacterial lineages.

Author contributions: P.H.O. and E.P.C.R. designed research; P.H.O., M.T., and E.P.C.R. analyzed data; and P.H.O., M.T., and E.P.C.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

¹To whom correspondence should be addressed. Email: pcpcho@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603257113/-DCSupplemental.

similar sequences (that recombine at higher rates). The advantages conferred by new genes might be higher when transfer takes place between more similar genetic backgrounds.

Here, we aimed at testing the effect of R-M systems on the genetic flux in bacterial populations. We concentrated on Type II R-M systems because they are the best studied, very frequent, and those for which we could predict sequence specificity. We inferred genome-wide counts of HR and HGT and tested their association with the frequency of R-M systems encoded in the genomes. We then made a more precise test of the key hypothesis that bacteria carrying similar R-M systems establish highways of gene transfer, independently of phylogenetic proximity and clade-specific traits.

Results

Quantification of Homologous Recombination, HGT, and Their Covariates.

We analyzed a dataset of 79 core genomes and pangenomes (SI Methods) corresponding to a total of 884 complete genomes. These clades were based on taxonomy, i.e., the genomes of a named species were put together. They spanned many different bacterial phyla (Fig. 1A and SI Methods). The pangenomes varied between 466 and 18,302 gene families (Dataset S1), and correlated with genome size

(Spearman's $\rho = 0.89$, $P < 10^{-4}$) and phylogenetic depth, defined as the average root-to-tip distances in the clade phylogenetic tree (SI Methods and Dataset S2) (Spearman's $\rho = 0.42$, $P < 10^{-4}$). Hence, our dataset represents a large diversity of bacteria in terms of taxonomy, genome size, and intraspecies diversity.

HR is notoriously difficult to quantify accurately (24). We used five different programs to detect HR in the core genome (SI Methods). These programs detect different types of signals, and together they should provide a thorough assessment of HR. Among the 79 core genomes, we found an average of 329 (NSS), 374 (MaxCHI), 264 (PHI), 504 (Geneconv), and 1,035 (Clonal-FrameML, CFML) HR events per core genome (Datasets S1 and S3). Even if the different methods provided different numbers of events, their results were highly correlated (average Spearman's $\rho = 0.84$, all comparisons $P < 10^{-4}$). Accordingly, we focused our analysis on the results of Geneconv, which provides the positions of recombination tracts and directions of transfer necessary for the last part of this study.

We used Count (25) to infer the events of HGT from the patterns of presence and absence of gene families in the species' trees (SI Methods and Dataset S4). We identified 236,894 events of gene transfer in the 79 pangenomes (Dataset S1). These events were very

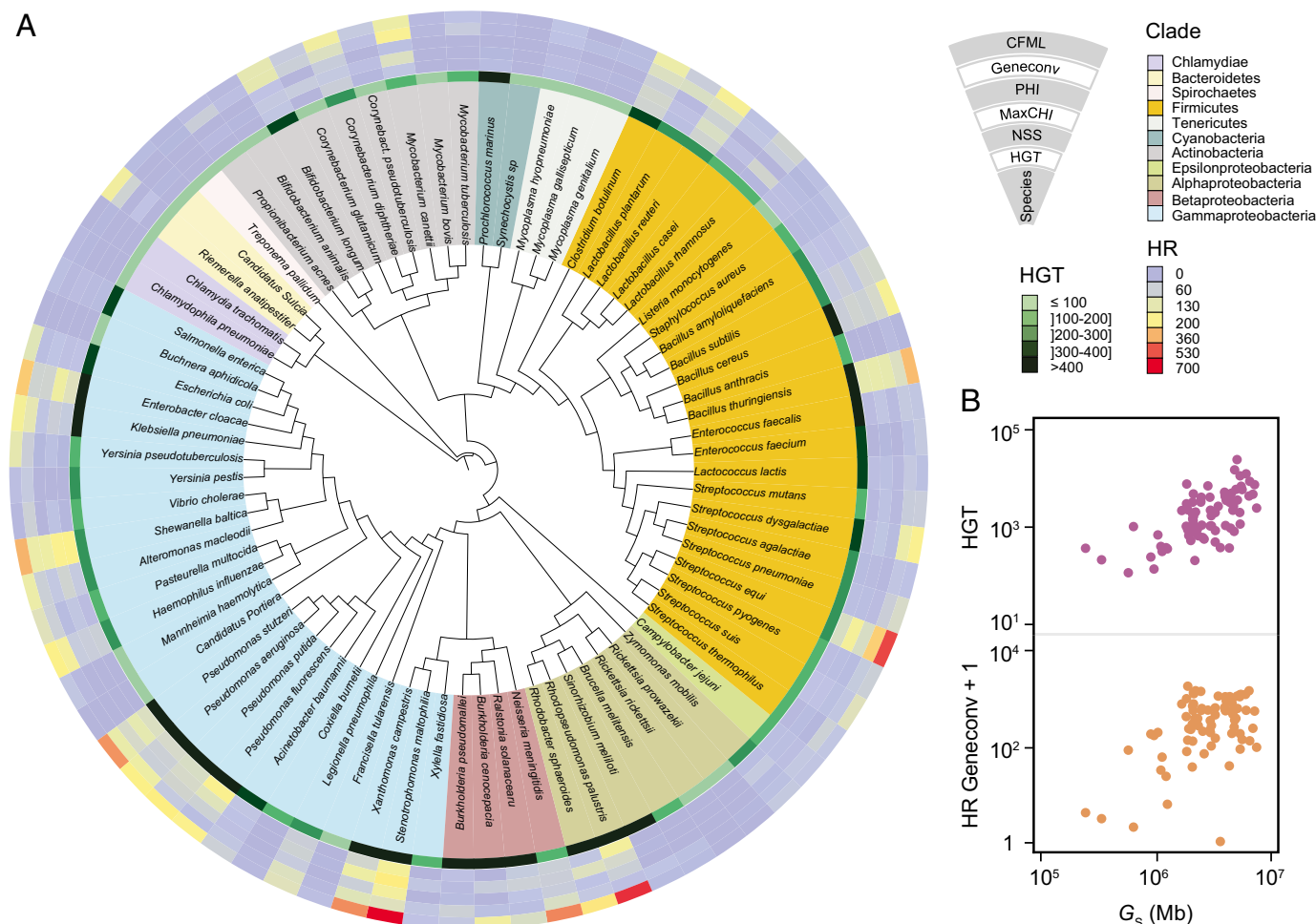


Fig. 1. Analysis of HR and HGT events. (A) 16S rRNA phylogenetic tree of the 79 bacterial species. The tree was drawn using the iTOL server (itol.embl.de/index.shtml) (40). The innermost circle layer indicates the species and associated clade. The six subsequent layers correspond (in an outwardly direction) to the average number of HGT events per genome computed using Count; the number of recombined genes per genome given by NSS, MaxChi, and PHI; and the number of recombination events per genome given by Geneconv and CFML (outermost layer), respectively. These values are given in Dataset S1. (B) Distribution of the average number of horizontal gene transfer (HGT) events and homologous recombination (HR) events (inferred by Geneconv) per clade according to genome size (G_s). Spearman's $\rho_{HGT} = 0.65$, $P_{HGT} < 10^{-4}$; Spearman's $\rho_{Geneconv} = 0.32$, $P_{Geneconv} < 10^{-2}$. Data obtained with the remaining recombination inference tools are shown in Fig. S1.

unevenly distributed among clades, from close to none in the genomes of obligatory endosymbionts to 1,538 events per genome in *Rhodospseudomonas palustris* (Fig. 1A).

The frequencies of HR and HGT were expected to depend on a number of variables, including the following: (i) genome size; (ii) phylogenetic depth (deeper lineages accumulate more events of exchange); and (iii) the number of genomes in the clade (larger samples capture more past events). We built stepwise linear models to assess the role of these variables in explaining the variance in HGT and HR (Table S1, part A). These showed that genome size had a strong direct effect on HR and HGT (Fig. 1B and Fig. S1). The remaining variables had significant, but less important, explanatory roles. HR also depended weakly on core genome size (Table S1, part B). Hence, studying the effect of R-M systems on HR and HGT requires control for phylogenetic depth, the number of genomes in the clade, and especially the genome size.

Association Between R-M Systems and Genetic Transfer. We identified 1,352 R-M systems among the 79 clades using a previously published methodology (4) (*SI Methods* and *Dataset S1*), including 233 Type II R-M systems (excluding Type IIC). The number of HGT events was higher in genomes with more R-M systems (Fig. 2A), and especially in those with Type II systems (Fig. 2B). The number of HR events increased with the number of R-M systems (Fig. 2C) and especially in the presence of Type II R-M systems (Fig. 2D). Similar results were obtained for the remaining HR inference tools (Fig. S2).

We then tested the effect of R-M systems on the number of HGT events and the rates of HR, while controlling for their covariates mentioned above. A stepwise regression showed that the numbers of Type II R-M systems were not significant predictors of HGT when the three previous variables were already introduced in the regression (the latter explaining ~76% of all variance; Table S1, part C). An analogous analysis for the frequency of HR showed that genome size and the number of Type II R-M systems were both significant predictors of HR ($R^2 = 0.42$, both variables $P < 10^{-4}$; Table S1, part C). These results show that genomes carrying more R-M systems acquire more genetic material by both HR and HGT, even if the latter association might be the result of clade-specific traits such as genome size.

Evolution of Target Motifs and Identification of Cognate R-M Systems.

To test the hypothesis that R-M systems affect the genetic flux between genomes, one needs to identify the systems recognizing the same target recognition motif. Such systems are cognates, i.e., DNA methylation by one system will protect from the other. We could not identify a method to identify cognate R-M systems in the literature. Hence, we created one based on the sequence

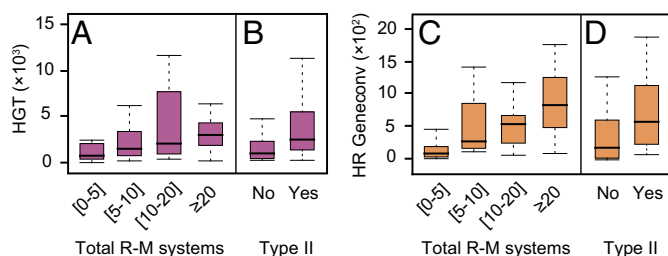


Fig. 2. Association between gene transfer and R-M systems. Distribution of the average HGT events (A) and homologous recombination (HR) events inferred by Geneconv (C) per clade according to the total number of R-M systems. Spearman's $\rho_{\text{HGT}} = 0.43$, Spearman's $\rho_{\text{Geneconv}} = 0.62$; both $P < 10^{-4}$. Distribution of the average HGT (B) and Geneconv HR events (D) per clade according to the presence (Yes)/absence (No) of Type II R-M systems (both $P < 10^{-4}$; Mann-Whitney-Wilcoxon test). We obtained similar qualitative results with the remaining recombination inference tools (Fig. S2).

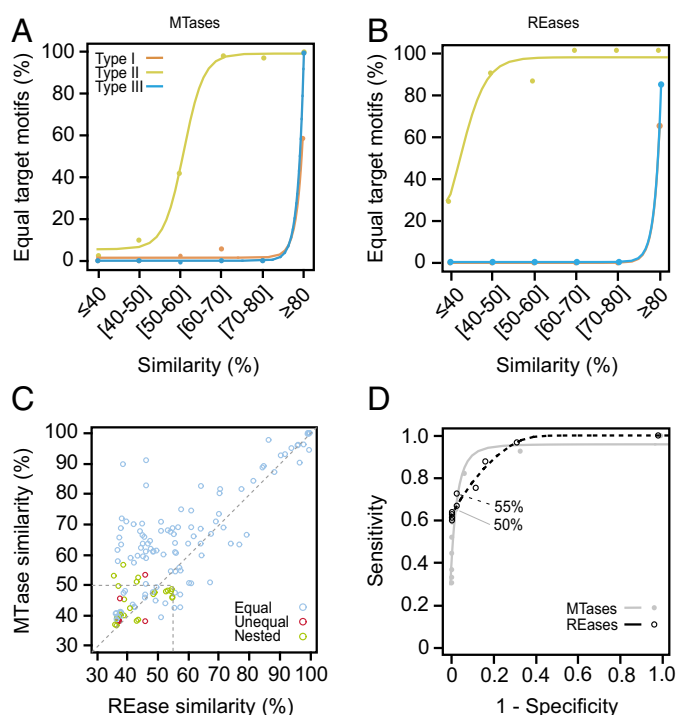


Fig. 3. Relation between target specificity and protein similarity in R-M components. Percentage of equal target motifs recognized by Types I, II, and III MTases (A) and REases (B) according to their pairwise protein sequence similarity. (C) Plot of all pairwise similarities of Type II MTases versus the cognate Type II REases of the REBASE gold standard. Blue dots correspond to equal target motifs, red dots to unequal target motifs, and green dots to nested motifs. The dashed horizontal and vertical lines indicate the threshold similarity limits for MTases and REases. (D) The same dataset was used to plot the corresponding receiver operating characteristic (ROC) curves. These curves depict the Sensitivity (true-positive rate) versus 1-Specificity (false-positive rate) for several values of percentage similarity of Type II MTases and REases. We selected the cutoff values of similarity that maximized the true-positive rate and minimized the false-positive rate. Details on the number of R-M proteins of each type can be found in Table S2. ROC data including curve-fitting equations can be found in Table S3.

conservation of MTases and REases. For this, we used the “gold-standard” component of REBASE (26) and plotted the frequency with which MTases or REases of a given type recognized the same motif (*SI Methods*) for a given bin of sequence similarity. Only nearly identical homologs of Types I and III MTases and REases recognized the same motifs (Fig. 3A and B). The analysis of the Specificity (S) and target recognition domains (TRDs) led to similar conclusions (*SI Methods* and Fig. S3A). The small number of such systems in REBASE gold standard resulted in small statistical power for this analysis, but adding more recent data from REBASE PacBio database did not change these conclusions (*SI Methods* and Fig. S3B-E). The rapid evolution of sequence target specificity precludes the identification of systems with similar restriction sites from the alignment of REases or MTases in both Type I and Type III R-M systems.

In contrast, homologs of Type II REases and MTases, which are much more numerous in the database, have different target motifs only when their sequence similarity is low (typically less than 50% for MTases and 55% for REases; Fig. 3). We used these thresholds to estimate the probability that two homologous systems recognize the same target recognition motif, and restricted our subsequent analyses to Type II systems.

R-M Systems Promote Preferential Genetic Transfer Fluxes. The observation of higher genetic fluxes in the presence of R-M systems might seem unexpected in the light of the role of the latter in

degrading exogenous DNA. To explain these results, we put forward three hypotheses.

Hypothesis 1: The relative abundance of R-M systems in a clade results from the selective pressure imposed by the abundance of MGEs in that clade. Selection for multiple R-M systems is expected to be stronger for clades enduring infections by many MGEs. R-M systems have limited efficiency and might not completely prevent MGE infection and transfer (8). This results in a weak positive association between transfer of genetic information and the abundance of R-M systems.

Hypothesis 2: R-M systems favor transfer of genetic material between cells by generating restriction breaks that stimulate recombination between homologous sequences.

Hypothesis 3: Type II R-M systems encoded in MGEs favor genetic transfer by selfishly stabilizing the element's presence in the new host (16). Genomes enduring more transfer would have more R-M systems if they were carried by MGEs. This last hypothesis is unlikely to explain our results, because we have shown that R-M systems are rare in MGEs (4). Furthermore, the association between genetic transfer and number of R-M systems remained significant when we excluded Type II R-M systems from the analysis (those more likely to act as selfish elements; Fig. S4). This fits recent findings that R-M systems occur and recombine in genomes in ways that are independent of the presence of MGEs (5).

To distinguish between the first two hypotheses, we analyzed the genetic flux between pairs of genomes with cognate Type II R-M systems. If R-M systems predominantly prevent genetic transfer (hypothesis 1), then the flux of genetic material between genomes encoding cognate R-M systems should be higher. If R-M systems predominantly stimulate genetic transfer (hypothesis 2), then pairs of genomes encoding cognate R-M systems should show lower than average genetic flux.

We tested the two hypotheses for HR and HGT separately. We selected the HR events that took place between terminal branches in the phylogenetic trees of the clades. Each terminal branch was then associated with the respective focal genome (the tip), which was labeled in terms of the target recognition motifs of the R-M systems encoded in the focal genome. We excluded HR or HGT occurring in the internal branches of the tree because of the high uncertainty in the inference of ancestral R-M systems (Fig. S5). We then computed the number of HR events between terminal branches associated with genomes encoding cognate R-M systems and compared it with the other pairs of genomes encoding R-M systems. Similar analyses were performed for HGT events that simultaneously affected pairs of terminal branches, i.e., for genes transferred to two terminal branches in two independent events. In both cases, we observed that lineages represented by genomes encoding cognate R-M systems coexchanged more genetic information (Fig. S6 A and B).

Next, we restricted our analysis to clades having at least 10 comparisons between genomes encoding cognate R-M systems and 10 comparisons between genomes lacking cognate systems (but encoding R-M systems). This avoids the confounding effect of putting together in the same analysis clades with few R-M systems or with little diversity in these systems. This restricted our dataset to eight clades: *Bacillus amyloliquefaciens*, *Bifidobacterium longum*, *Escherichia coli*, *Haemophilus influenzae*, *Listeria monocytogenes*, *N. meningitidis*, *Salmonella enterica*, and *S. pneumoniae*. Within this restricted dataset, the results were qualitatively identical: lineages associated with genomes encoding cognate R-M systems coexchanged more genetic information (Fig. 4 A–C). We confirmed that these results were insensitive to uncertainties in phylogenetic reconstruction and to the effects of HR in phylogenetic inference (SI Methods and Fig. S7). The

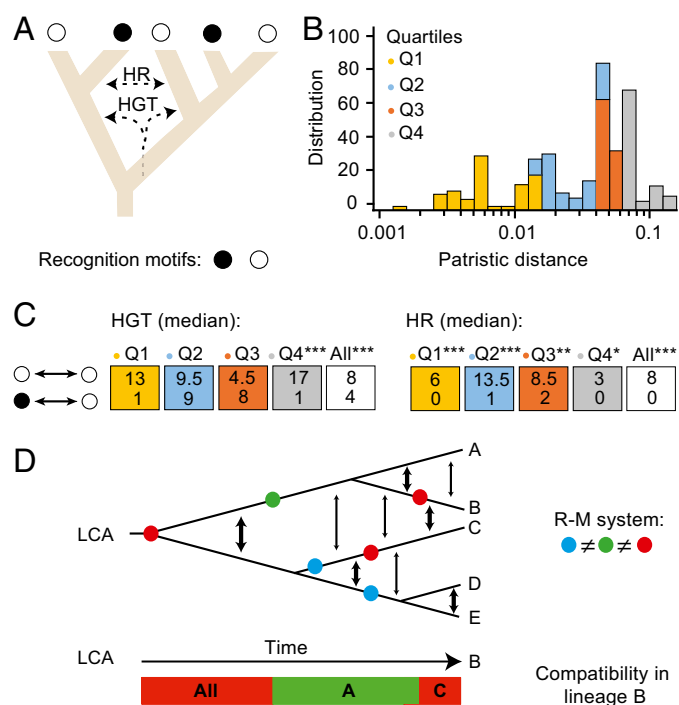


Fig. 4. Gene flux in bacteria encoding R-M systems. (A) We analyzed the patterns of HR and HGT in the tree of each clade, comparing the flux between tips ending in cognate (similar recognition motifs) or noncognate (different motifs) extant taxa. (B) Histogram of patristic distances (colored by quartiles) between bacteria with Type II R-M systems. (C) Median values of HGT and recombination events for each quartile (Q) and for the full dataset (All) between terminal branches of bacteria with Type II R-M systems recognizing (or not) the same target motif. We analyzed *Bacillus amyloliquefaciens*, *Bifidobacterium longum*, *Escherichia coli*, *Haemophilus influenzae*, *Listeria monocytogenes*, *Neisseria meningitidis*, *Salmonella enterica*, and *Streptococcus pneumoniae*. * $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$ (see Fig. S6 A and B for the data including all clades). (D) Genetic flux in function of time and the presence of R-M systems. As lineages diverge and R-M systems change (circles indicate such changes), the lineages with cognate R-M systems (same color) share more genetic material than the other lineages. For example, the lineage B changes R-M systems twice since the last common ancestor (LCA). Initially transfer is favored with all lineages, then with the sister lineage A, and finally with the distantly related lineage C.

results on HR might be strongly affected by the ability of bacteria to engage in natural transformation. We restricted our analysis to the five naturally transformable species, following ref. 27, and found similar results ($P < 10^{-3}$).

We then tested whether the clade-associated traits covarying with HR and HGT—phylogenetic depth, average genome size, and number of genomes—were affecting our conclusions by making the comparisons on each clade separately. We observed more HGT and HR among pairs of genomes encoding cognate R-M systems in six of the eight clades, which was statistically significant (each $P = 0.035$, binomial test, $P = 0.01$ for the combined test). One species (*L. monocytogenes*) was an exception to the general trend both concerning HR and HGT. This species showed very low rates of HGT and HR, and the differences in HR and HGT between R-M cognate and R-M noncognate genomes were not significant.

We mentioned in the Introduction that closely related taxa are expected to exchange more genetic information independently of the R-M systems they encode. To verify that the presence of cognate R-M systems is associated with increased genetic exchange independently of evolutionary distance, we binned the comparisons between events occurring in terminal branches in terms of the phylogenetic distance between pairs of genomes. We

then ran the same analysis in each bin separately. These analyses showed more cotransfer between genomes encoding cognate R-M systems in nearly all bins, even if this analysis had lower statistical power (fewer comparisons per bin) (Fig. 4 B and C for the eight clades and Fig. S6 A and B for all of the data). Importantly, this difference was always significant for the most distant pairs of genomes. Hence, pairs of genomes encoding cognate R-M systems were associated with more frequent HR and HGT, independently of the evolutionary distances between them.

Discussion

Genome size is the result of the balance between accretion and deletion events moderated by natural selection. Larger bacterial genomes are expected to engage in more frequent HGT because this is the dominant mechanism of genetic accretion (28). However, there are remarkably few studies demonstrating an association between HGT and genome size (29). Here, we found that larger genomes exchange DNA at higher rates, both by HGT and by HR. This association is not just caused by sexually isolated endosymbiotic bacteria with very small genomes—e.g., *Chlamydiae*, *Buchnera*, or *Spirochaetes* (Fig. 1 and Dataset S1)—because it remains significant for genomes larger than 2 Mb, which include few obligatory endosymbionts. Many reasons might explain the association between HGT, HR, and genome size: bacteria with larger genomes might have more diverse lifestyles, select for more diverse types of genes, inhabit more environments, or accommodate more MGEs. Even if the test of these different hypotheses falls outside the scope of this work, this association is important and must be accounted for when assessing the impact of R-M systems in genetic fluxes. The higher frequency of HR and HGT among larger genomes suggests that the latter are more targeted by MGEs. Accordingly, larger bacterial genomes encode more transposable elements (30), more prophages (31), and more conjugative elements (32). If MGEs targeting bacteria with larger genomes are more abundant, they might lead to strong selection for R-M systems in their bacterial hosts. This might explain why we found more R-M systems in larger genomes (4). It might also explain the positive association between the frequencies of HR and HGT and the abundance of R-M systems (Fig. 2).

R-M systems have a well-known inhibitory effect on the transfer of genetic information (9). However, whether this trait is an important driver of their evolution has remained controversial (12, 33, 34). Our results contribute to the clarification of these two issues. R-M systems can function as a barrier to MGE infection when encoded in the chromosome or other MGE. They can also stabilize the presence of MGEs in cells by preventing infections by other competing MGEs. Our previous observation that MGEs encode few R-M systems and many solitary MTases (4), suggests that R-M systems are more frequently a chromosomal-encoded barrier to MGEs than an MGE-encoded tool for cell infection. The coassociation of MGEs, bacterial genome size, and R-M systems might thus result from increased selection for R-M systems in the face of abundant MGEs in large genomes.

Contrary to the popular view that R-M systems limit the flux of genetic material (9), it has been proposed that restriction actually favors evolvability by producing DNA double-stranded ends that are recombinogenic (33, 34). This hypothesis is compatible with the observation that genomes enduring more HGT and HR have more R-M systems. However, it is not in agreement with the

observation that pairs of genomes encoding cognate R-M systems coexchange more DNA. It is also hardly reconcilable with the notorious deleterious effect of R-M systems on bacterial genetic transformation in the laboratory (35). Although R-M systems have been shown to favor intragenomic HR events (12), the overall effect of R-M systems on genetic exchange is to decrease both HR and HGT between bacteria encoding noncognate R-M systems.

Our statistical analyses could not explicitly account for the presence of the many other systems affecting genetic transfer between cells. Some of them facilitate transfer, e.g., MGEs or competence for natural transformation, and we checked that all of the clades in Fig. 4 have known phages and conjugative elements. Restricting the analysis to the five naturally transformable bacteria did not change our results. Importantly, all of these clades encoded the key enzymes involved in RecA-mediated homologous recombination, including the presynaptic pathways RecBCD/AddAB and RecOR (36). Hence, there is little ground to think that our results are strongly biased by lack of mechanisms for gene transfer. Some systems disfavor transfer between bacteria, including CRISPRs, abortive infection, or other R-M systems. It is not possible to account for all these factors in a statistical model, because of the lack of quantitative data. Nevertheless, we could verify that cognate genomes did not have fewer R-M systems than the other genomes. Even if other barriers to DNA exchange are certainly present in these species, our use of a diverse set of well-known species, numerous alternative analyses, and focus on intraspecies comparisons (in which lifestyles and other general traits are much less variable), suggests that our results are robust.

Our work shows that noncognate genomes have reduced DNA exchanges. This decreases the power of natural selection and increases the effect of drift, potentially leading to the accumulation of deleterious mutations. Importantly, R-M systems' diversification at the origin of a lineage may increase its genetic cohesion by disfavoring exchanges with the closest related ones, as previously suggested for some pathogens (20–22). Interestingly, diversification can also increase the genetic flux between distant bacteria encoding cognate R-M systems with which there were previously few genetic exchanges. Hence, R-M systems might shape population structure in complex ways depending on the repertoire of R-M systems in the other lineages.

The study of the flux of genetic information among bacteria using network-based approaches is rising in importance (37–39). Our work shows that R-M systems may carve preferential routes of DNA exchange between certain bacterial subpopulations. Their rapid diversification constantly changes these preferences, thereby producing complex patterns of genetic exchange with time.

Methods

Details on the data used, identification of core genomes and pangenomes, phylogenetic analyses, inference of HR, reconstruction of the evolution of gene families, identification of R-M systems, robustness of the target motif predictor, and robustness of the Count analysis to phylogenetic reconstruction can be found in *SI Methods*.

ACKNOWLEDGMENTS. We thank Vincent Daubin (Université de Lyon) for the suggestion to use Count. We also thank Florent Lassale (University College London) and the anonymous reviewers for critically reviewing the manuscript. This work was supported by European Research Council Grant EVOMOBILOME 281605.

1. Frost LS, Leplae R, Summers AO, Toussaint A (2005) Mobile genetic elements: The agents of open source evolution. *Nat Rev Microbiol* 3(9):722–732.
2. Vulić M, Dionisio F, Taddei F, Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA* 94(18):9763–9767.
3. Didelot X, Wilson DJ (2015) ClonalFrameML: Efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 11(2):e1004041.
4. Oliveira PH, Touchon M, Rocha EP (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res* 42(16):10618–10631.
5. Mruk I, Kobayashi I (2014) To be or not to be: Regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res* 42(1):70–86.
6. Pingoud A, Wilson GG, Wende W (2014) Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Res* 42(12):7489–7527.
7. Vasu K, Nagamalleswari E, Nagaraja V (2012) Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria. *Proc Natl Acad Sci USA* 109(20):E1287–E1293.
8. Korona R, Korona B, Levin BR (1993) Sensitivity of naturally occurring coliphages to type I and type II restriction and modification. *J Gen Microbiol* 139(Pt 6): 1283–1290.



9. Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat Rev Microbiol* 3(9):711–721.
10. Labrie SJ, Samson JE, Moineau S (2010) Bacteriophage resistance mechanisms. *Nat Rev Microbiol* 8(5):317–327.
11. Korona R, Levin BR (1993) Phage-mediated selection for restriction-modification. *Evolution* 47(2):565–575.
12. Kobayashi I (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res* 29(18):3742–3756.
13. Xu Q, Morgan RD, Roberts RJ, Blaser MJ (2000) Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proc Natl Acad Sci USA* 97(17):9671–9676.
14. Jeltsch A, Pingoud A (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J Mol Evol* 42(2):91–96.
15. Seshasayee AS, Singh P, Krishna S (2012) Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res* 40(15):7066–7073.
16. Kusano K, Naito T, Handa N, Kobayashi I (1995) Restriction-modification systems as genomic parasites in competition for specific sequences. *Proc Natl Acad Sci USA* 92(24):11095–11099.
17. Didelot X, et al. (2011) Recombination and population structure in *Salmonella enterica*. *PLoS Genet* 7(7):e1002191.
18. Doroghazi JR, Buckley DH (2010) Widespread homologous recombination within and between *Streptomyces* species. *ISME J* 4(9):1136–1143.
19. Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* 315(5811):476–480.
20. Budroni S, et al. (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc Natl Acad Sci USA* 108(11):4494–4499.
21. Croucher NJ, et al. (2014) Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat Commun* 5:5471.
22. Nandi T, et al. (2015) *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res* 25(1):129–141, and erratum (2015) 25(4):608.
23. Roberts GA, et al. (2013) Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant *Staphylococcus aureus* (MRSA) populations. *Nucleic Acids Res* 41(15):7472–7484.
24. Chan CX, Beiko RG, Ragan MA (2006) Detecting recombination in evolving nucleotide sequences. *BMC Bioinformatics* 7:412.
25. Csurös M (2010) Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* 26(15):1910–1912.
26. Roberts RJ, Vincze T, Posfai J, Macelis D (2010) REBASE—a database for DNA restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res* 38(Database issue):D234–D236.
27. Johnston C, Martin B, Fichant G, Polard P, Claverys JP (2014) Bacterial transformation: Distribution, shared mechanisms and divergent control. *Nat Rev Microbiol* 12(3):181–196.
28. Treangen TJ, Rocha EP (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7(1):e1001284.
29. Cordero OX, Hogeweg P (2009) The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proc Natl Acad Sci USA* 106(51):21748–21753.
30. Touchon M, Rocha EP (2007) Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol* 24(4):969–981.
31. Bobay LM, Rocha EP, Touchon M (2013) The adaptation of temperate bacteriophages to their host genomes. *Mol Biol Evol* 30(4):737–751.
32. Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EP (2011) The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet* 7(8):e1002222.
33. Arber W (2000) Genetic variation: Molecular mechanisms and impact on microbial evolution. *FEMS Microbiol Rev* 24(1):1–7.
34. Vasu K, Nagaraja V (2013) Diverse functions of restriction-modification systems in addition to cellular defense. *Microbiol Mol Biol Rev* 77(1):53–72.
35. Corvaglia AR, et al. (2010) A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical *Staphylococcus aureus* strains. *Proc Natl Acad Sci USA* 107(26):11954–11958.
36. Rocha EP, Cornet E, Michel B (2005) Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* 1(2):e15.
37. Halary S, Leigh JW, Cheaib B, Lopez P, Baptiste E (2010) Network analyses structure genetic diversity in independent genetic worlds. *Proc Natl Acad Sci USA* 107(1):127–132.
38. Skippington E, Ragan MA (2011) Lateral genetic transfer and the construction of genetic exchange communities. *FEMS Microbiol Rev* 35(5):707–735.
39. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T (2011) Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. *Genome Res* 21(4):599–609.
40. Letunic I, Bork P (2011) Interactive Tree of Life v2: Online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* 39(Web Server issue):W475–W478.
41. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61–D65.
42. Touchon M, et al. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
43. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* 12:116.
44. Guindon S, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* 59(3):307–321.
45. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
46. Schliep KP (2011) phangorn: Phylogenetic analysis in R. *Bioinformatics* 27(4):592–593.
47. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
48. Jakobsen IB, Easteal S (1996) A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Comput Appl Biosci* 12(4):291–295.
49. Smith JM (1992) Analyzing the mosaic structure of genes. *J Mol Evol* 34(2):126–129.
50. Bruen T, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–2681.
51. Sawyer S (1989) Statistical tests for detecting gene conversion. *Mol Biol Evol* 6(5):526–538.
52. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289–290.
53. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ (2014) PopGenome: An efficient Swiss army knife for population genomic analyses in R. *Mol Biol Evol* 31(7):1929–1936.
54. Wolf YI, Makarova KS, Yutin N, Koonin EV (2012) Updated clusters of orthologous genes for Archaea: A complex ancestor of the Archaea and the byways of horizontal gene transfer. *Biol Direct* 7:46.
55. Cohen O, Pupko T (2010) Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol Biol Evol* 27(3):703–713.
56. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30(7):1575–1584.
57. Katoh K, Standley DM (2014) MAFFT: Iterative refinement and additional methods. *Methods Mol Biol* 1079:131–146.
58. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 27(2):221–224.
59. Finn RD, Clements J, Eddy SR (2011) HMMER Web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29–W37.
60. Furuta Y, Kobayashi I (2012) Mobility of DNA sequence recognition domains in DNA methyltransferases suggests epigenetics-driven adaptive evolution. *Mob Genet Elements* 2(6):292–296.
61. Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* 175(3):1251–1266.

ARTICLE

DOI: [10.1038/s41467-017-00808-w](https://doi.org/10.1038/s41467-017-00808-w)

OPEN

The chromosomal organization of horizontal gene transfer in bacteria

Pedro H. Oliveira ^{1,2}, Marie Touchon^{1,2}, Jean Cury ^{1,2} & Eduardo P.C. Rocha^{1,2}

Bacterial adaptation is accelerated by the acquisition of novel traits through horizontal gene transfer, but the integration of these genes affects genome organization. We found that transferred genes are concentrated in only ~1% of the chromosomal regions (hotspots) in 80 bacterial species. This concentration increases with genome size and with the rate of transfer. Hotspots diversify by rapid gene turnover; their chromosomal distribution depends on local contexts (neighboring core genes), and content in mobile genetic elements. Hotspots concentrate most changes in gene repertoires, reduce the trade-off between genome diversification and organization, and should be treasure troves of strain-specific adaptive genes. Most mobile genetic elements and antibiotic resistance genes are in hotspots, but many hotspots lack recognizable mobile genetic elements and exhibit frequent homologous recombination at flanking core genes. Overrepresentation of hotspots with fewer mobile genetic elements in naturally transformable bacteria suggests that homologous recombination and horizontal gene transfer are tightly linked in genome evolution.

¹Microbial Evolutionary Genomics, Institut Pasteur, 25–28 rue du Docteur Roux, Paris 75015, France. ²CNRS, UMR3525, 25–28 rue du Docteur Roux, Paris 75015, France. Pedro H. Oliveira and Marie Touchon contributed equally to this work. Correspondence and requests for materials should be addressed to P.H.O.(email: pcphco@gmail.com) or to M.T.(email: mtouchon@pasteur.fr)

The gene repertoires of bacterial species are often very diverse, which is central to bacterial adaptation to changing environments, new ecological niches, and co-evolving eukaryotic hosts¹. Novel genes arise in bacterial genomes mostly by horizontal gene transfer (HGT)², a pervasive evolutionary process that spreads genes between, eventually very distant, bacterial lineages³. It is commonly thought that the majority of genes acquired by HGT are neutral or deleterious and thus rapidly lost⁴. Yet, HGT is also responsible for the acquisition of many adaptive traits, including antibiotic resistance in nosocomials⁵. Hence, genome diversification is shaped by the balancing processes of gene acquisition and loss⁶, moderated by positive selection on some genes, and purifying selection on many others⁷.

Chromosomes are organized to favor the interactions of DNA with the cellular machinery⁸. For example, most bacterial genes are co-transcribed in operons, leading to strong and highly conserved genetic linkage between neighboring genes⁹. At a more global level, early-replicating regions are enriched in highly expressed genes in fast-growing bacteria to enjoy replication-associated gene dosage, creating a negative gradient of expression along the axis from the origin (*ori*) to the terminus (*ter*) of replication (*ori*->*ter*)^{10,11}. These organizational traits can be disrupted by the integration of novel genetic information. At a local level, new genes rarely integrate within an operon and, instead, they tend to be incorporated at its edges, where they are less likely to affect gene expression¹². At the genome level, the frequency of integration of prophages in the genome of *Escherichia coli* increases along the *ori*->*ter* axis¹³. The results of these studies suggest that the fitness effects of HGT in terms of chromosome organization depend on the specific site of integration.

In prokaryotes, HGT takes place by three main mechanisms: natural transformation, conjugation, and transduction. Mobile genetic elements (MGEs) play a key role in HGT because they are responsible for the latter two processes, respectively by the activity of conjugative elements and phages¹⁴. Integrative conjugative elements (ICEs) and prophages are large genetic elements that may account for a significant fraction of the bacterial genome^{15,16}, and bring to the chromosome many genes in a single event of integration. For example, some strains of *E. coli* have up to 18 prophages¹⁷, and *Mesorhizobium loti*

encodes one ~500 kb ICE¹⁸. The integration of these large MGEs changes the chromosome size and may split adaptive genetic structures such as operons. This might contribute to explain why most integrative MGEs use site-specific recombinases (integrases) that target very specific sites in the chromosome¹⁹. Integrases and MGEs have co-evolved with the host genome to decrease the fitness cost of their integration¹³.

MGEs carrying similar integrases tend to integrate at the same sites in the chromosome, leading to regions with unexpectedly high frequency of MGEs at homologous regions. This concentration of MGEs in few sites has been frequently described^{20,21}, especially in relation to the presence of neighboring tRNA and tmRNA genes²². Yet, a previous work described the existence of regions with high rates of diversification in *E. coli* (hotspots), some of which lacked recognizable integrases²³. In particular, the genes flanking two hotspots were associated with high rates of homologous recombination (*rfb* and *leuX*). In *Streptococcus pneumoniae*, the chromosomal genes flanking MGEs also showed higher rates of homologous recombination^{24,25}. In this species, it was suggested that integration of MGEs close to core genes under selection for diversification could be adaptive by facilitating the transfer and subsequent recombination of the latter²⁶.

Here, we define and identify hotspots in a large and diverse panel of bacterial species and show how they reflect the mechanisms driving genome diversification by HGT.

Results

Quantification of HGT and definition of hotspots. To study the distribution of gene families in bacterial chromosomes, we analyzed 932 complete genomes of 80 bacterial species (Supplementary Data set 1). We inferred the core genome, the pan-genome, the accessory genome (genes from the pan-genome absent from the core), and the phylogeny of each species, as before²⁷ (Methods, Supplementary Figs. 1 and 2). We partitioned the genomes into an array of core genes and intervals (Fig. 1, Table 1). The latter were defined as the positions between consecutive core genes. We defined a spot as the set of intervals delimited by members of the same two families of core genes in the genomes of the clade (see Methods for rigorous definitions, Supplementary Fig. 2a, b). We observed that 99.4% of the intervals were part of the species' spots and only 0.6% were in

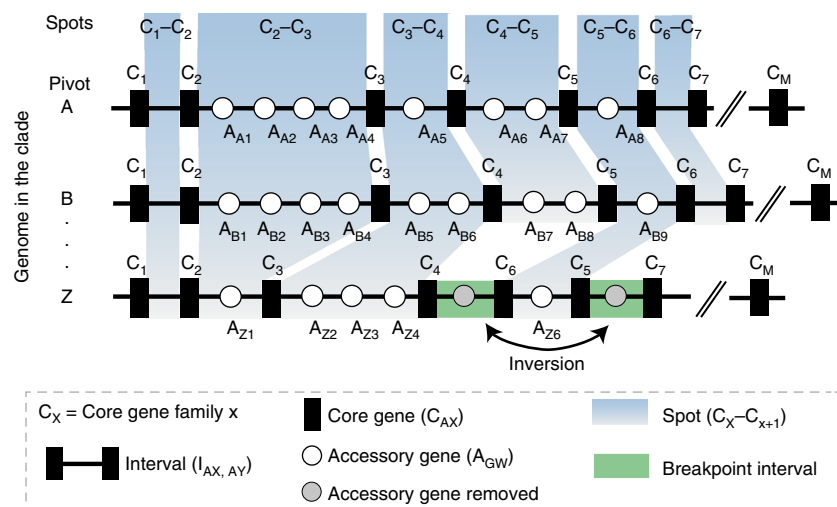


Fig. 1 Scheme depicting key concepts used in this study. Intervals flanked by the same core gene families (C_x, C_y) as those from pivot genome A were defined as syntenic intervals (i.e., the members of the core gene families X and Y were also contiguous in the pivot). The intervals that do not satisfy this constraint were classed as breakpoint intervals (green-shaded regions) and excluded from our analysis. For every interval in the pivot genome, we defined spot as the set of intervals flanked by members of the same core gene families (blue-shaded regions)

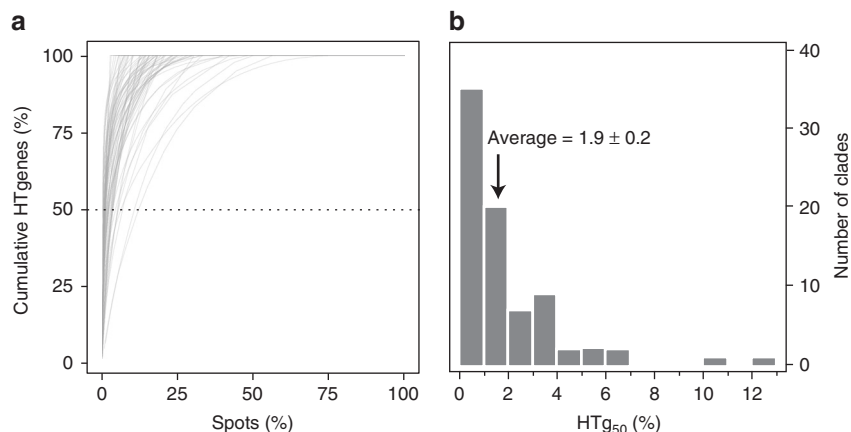


Fig. 2 Cumulative frequency of HTgenes. **a** Cumulative distribution of horizontally transferred genes (HTgenes, %) in spots for the 80 bacterial clades. **b** Histogram of the minimum number of spots needed to attain 50% of the total number of HTgenes (HTg₅₀ index). The average HTg₅₀ was only 1.9% (± 0.2 ; standard deviation)

Table 1 Acronyms used in this study

MGE	Mobile genetic element (i.e., prophage, ICE, IME and integron)
ICE	Integrative conjugative element
IME	Integrative mobilizable element
MAP	Mobility-associated protein (i.e., integrase and transposase (IS))
ARG	Antibiotic resistance gene
HGT	Horizontal gene transfer
HTgenes	Genes having been horizontally transferred
HTg ₅₀	Number of spots required to include 50% of HTgenes
T _{95%}	Minimal number of HTgenes required to define a hotspot

breakpoint intervals. Since 99.8% of spots are flanked by the same two families of core genes in at least half of the genomes of each clade (and 99% in all genomes), it is most parsimonious to consider that the two core genes were already contiguous in the last common ancestor of the clade. Hence, we split the pan-genomes in spot pan-genomes, i.e., sets of gene families located in each spot (Methods, Supplementary Fig. 2). The genes outside spots, i.e., in intervals that were split by events of rearrangement, accounted for < 2% of the total number of genes and were discarded from further analysis.

We used birth-and-death models to identify HGT events in the clade's phylogenetic trees from the patterns of presence/absence of each gene family (Methods). Note that HGT events are defined gene per gene (which will be called HTgenes for Horizontally Transferred Genes), not as blocks, because there are no tools available for the latter and because the goal of our work was to study the clustering of genes acquired by HGT without using a priori models. Spots contained 170,041 HTgenes (15.5% of the total number of accessory genes). We quantified the clustering of these genes by counting the minimal number of spots required to accumulate at least 50% of the HTgenes (HTg₅₀) (Fig. 2a). The distribution of these values was skewed toward small values (Fig. 2b). Hence, < 2% of the largest hotspots accumulate > 50% of all HTgenes. Conversely, 72.6% of the spots were on average empty, i.e., had no accessory gene in any genome. Similar qualitative conclusions were obtained in the analysis of the distribution of all accessory genes, despite the latter being slightly less clustered (Supplementary Fig. 3). These results show that most HTgenes are integrated in a very small number of sites in the genome.

We used simulations to infer the statistical thresholds for the degree of clustering of HTgenes in each clade (Methods, Supplementary Fig. 4). We made the null hypothesis that these genes are organized in operons like the other genes, and are uniformly distributed among spots. We identified the spot with the highest number of HTgenes in each simulation ($\text{Max}_{\text{HTg},i}$), and computed the 95th percentile of the distribution of these maximal values ($T_{95\%}$, Supplementary Data set 1). Simulations disregarding the existence of operons produced lower values of $T_{95\%}$ showing the importance of incorporating information about genetic organization in the model (Supplementary Fig. 5). Spots with more than $T_{95\%}$ HTgenes were called hotspots, spots lacking accessory genes were called empty, and the others were called coldspots. We found a total of 1841 hotspots in the 80 clades (Supplementary Data set 1). They represent only 1.2% of the spots, but they concentrate 47% of the accessory gene families and 60% of the HTgenes.

The number of hotspots differed widely among clades, from none or very few in *Acetobacter pasteurianus*, *Bacillus anthracis*, and the obligatory endosymbionts, to more than 60 in *Bacillus thuringiensis*, *E. coli*, and *Pseudomonas putida* (Fig. 3a). This variance was partly a function of chromosome size (Fig. 3b), but was especially associated with the number of HTgenes (Fig. 3c). Increases in the latter resulted in a less-than-linearly increase in the number of hotspots and in a linear increase in hotspot density per Mb (Supplementary Fig. 6). Hence, a few hotspots aggregate most of the genes acquired by horizontal transfer and this trend is more pronounced when the rates of transfer are high.

Functional and genetic characterization of hotspots. We investigated the function of the genes in the spots, using the eggNOG categories, to assess if hotspots were enriched in particular traits (Methods, Fig. 4a). Genes classified as poorly characterized or as having an unknown function were not considered in the subsequent functional analyses (they were 13.1% of the total). We then compared the distribution of the functions of all accessory genes and that of HTgenes in hotspots relative to coldspots. Both analyses showed an underrepresentation of translation and post-translational modification genes in hotspots. These genes tend to be essential and are less frequently transferred horizontally²⁸. In contrast, hotspots overrepresented genes associated with cell motility, defense mechanisms, transcription, and replication and repair. Moreover, around 9% of the hotspots encoded antibiotic

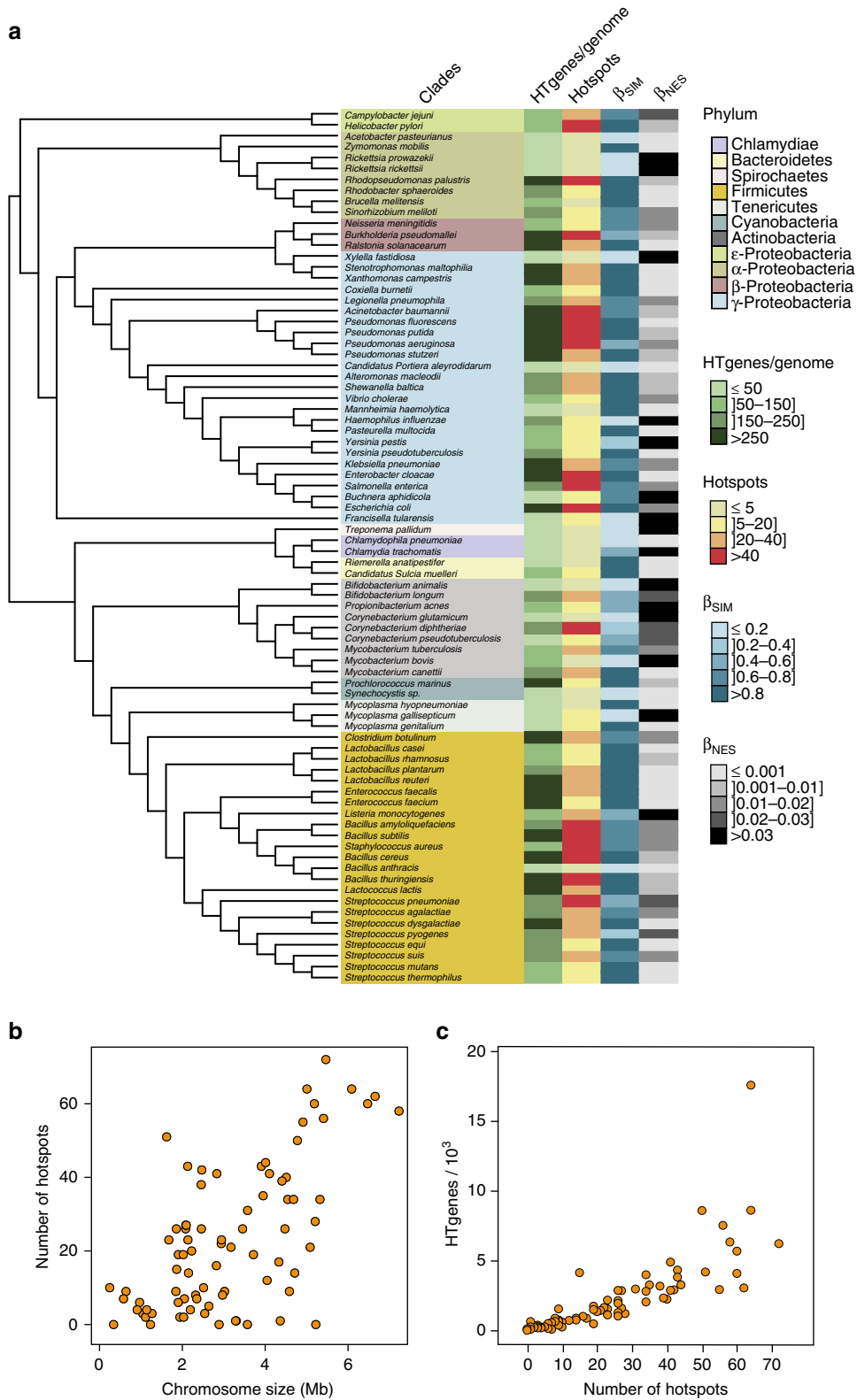


Fig. 3 Analysis of HTgenes and the abundance and distribution of hotspots. **a** 16S rRNA phylogenetic tree of the 80 bacterial clades. The tree was drawn using the iTOL server (itol.embl.de/index.shtml)⁷⁰. The first column indicates the clade and is colored by phylum. The four subsequent columns correspond respectively to: the average number of HTgenes per genome computed using Count, the number of hotspots, the average Simpson dissimilarity index (β_{SIM} , accounting for turnover), and the average multiple-site dissimilarity index accounting only for nestedness (β_{NES}). These values are given in Supplementary Data set 1. **b** Distribution of the average number of hotspots per clade according to the average genome size (G_S). **c** Association between the number of hotspots and the number of HTgenes in the clade

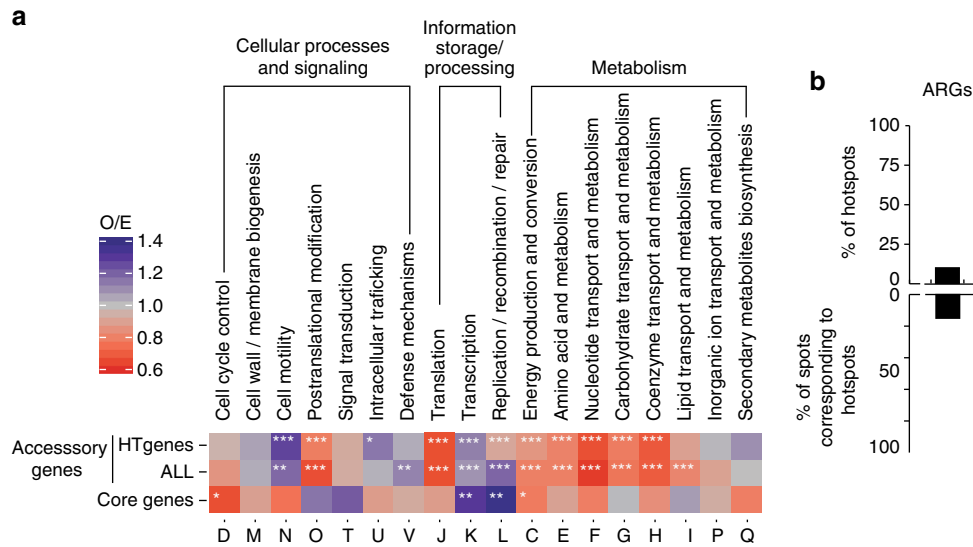


Fig. 4 Functional characterization of hotspots. **a** Observed/expected (O/E) ratios of non-supervised orthologous groups (NOGs, shown as *capitalized letters*). The first two *lines* represent the values of HTgenes and accessory genes observed in hotspots when the null model was computed from the distribution of the same type of genes in coldspots. The last *line* shows the same type of analysis for the core genes flanking hotspots when the null model is computed using the core genes not flanking hotspots. Expected values were obtained by multiplying the number of HTgenes, accessory, or core genes in hotspots by the fraction of genes assigned to each NOG. * $P < 0.05$; ** $P < 10^{-2}$; *** $P < 10^{-3}$, χ^2 -test. **b** Percentage of hotspots with antibiotic resistance genes (ARGs, *top*), and percentage of spots with ARGs that are hotspots (*bottom*). Note that hotspots are only 1.2% of all the spots

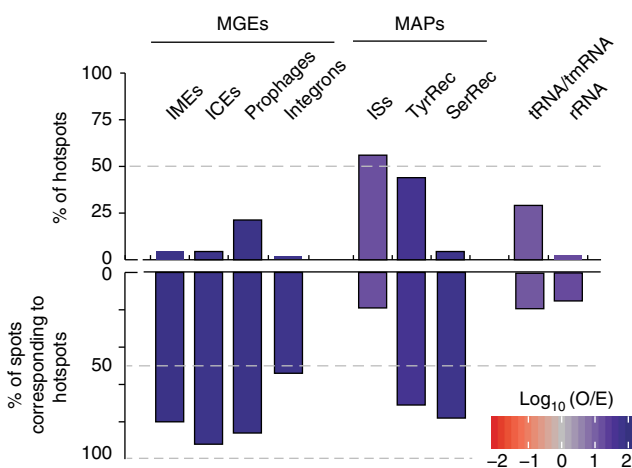


Fig. 5 Genetic mobility of hotspots. We represent the percentage of hotspots containing the different genetic elements (*top*) and the percentage of spots containing such elements that are hotspots (*bottom*). Note that hotspots are only 1.2% of all the spots. The analysis includes MGEs (IMEs, ICes, prophages, integrations), mobility-associated proteins (MAPs) (ISs, TyrRec, SerRec), and tRNA/tmRNA, rRNA. Also, shown in *colored bins* are the observed/expected ($\log_{10}O/E$) number of hotspots that contain the abovementioned elements, when the null model was computed from the distribution of coldspots containing the same type of elements. Expected values were obtained by multiplying the number of hotspots by the fraction of spots containing each type of element

resistance genes (ARGs), which is much more than expected by chance (0.8%) (Fig. 4b).

Some of the functions overrepresented in hotspots—defense, replication, repair—are typically found in MGEs, which concentrate in specific loci targeted by integrases (often at tRNAs). Accordingly, the vast majority of self-mobilizable MGEs—89% of the prophages and 90% of ICes—were identified in hotspots (Supplementary Data set 3 and Supplementary Fig. 7). On the

other hand, only around 9% of the hotspots encoded ICEs or integrative mobilizable elements (IMEs), and only 23% encoded prophages (Fig. 5). Integrations were even rarer (present in 1% of the hotspots). Non-self-transferable MGEs lack conjugation or virion structural genes, but usually encode integrases. The vast majority of integrases was identified in hotspots, but less than half (45%) of the hotspots encoded an integrase and only 29% encoded tRNA or tmRNA genes (Fig. 5). Hence, although most self-mobilizable MGEs are in hotspots, most hotspots lack them (Supplementary Fig. 8).

Insertion sequences (ISs) encoding DDE recombinases (transposases) are frequent within MGEs, and we found them in many hotspots (56%). The integration of these elements has low-sequence specificity, which explains why hotspots accounted for a small fraction of the locations with ISs (19%), unlike what we observed for self-mobilizable MGEs and integrases. Altogether, half of the hotspots lacked evidence for the presence of MGEs and 27% lacked any of the mobility-associated proteins (MAPs, integrases and transposases) that we searched for. These results confirm that hotspots concentrate most MGEs and integrases, but not the majority of ISs. They also show that regions with high concentration of HTgenes often lack recognizable MGEs, suggesting that other mechanisms are implicated in their genesis and turnover.

The chromosomal context of hotspots. We then searched to identify the preferential genetic contexts of hotspots, since they might illuminate constraints associated with the chromosomal organization of HGT. We analyzed whether the distribution of hotspots was random relative to the function of the neighboring core genes. Interestingly, these core genes showed an overrepresentation of several functions, notably replication, recombination/repair, and transcription (Fig. 4a). In contrast, cell cycle control genes were underrepresented. Hence, hotspots are preferentially associated with specific functions of neighboring core genes.

We then tested whether hotspots were randomly distributed in genomes. Since replication drives much of the large-scale

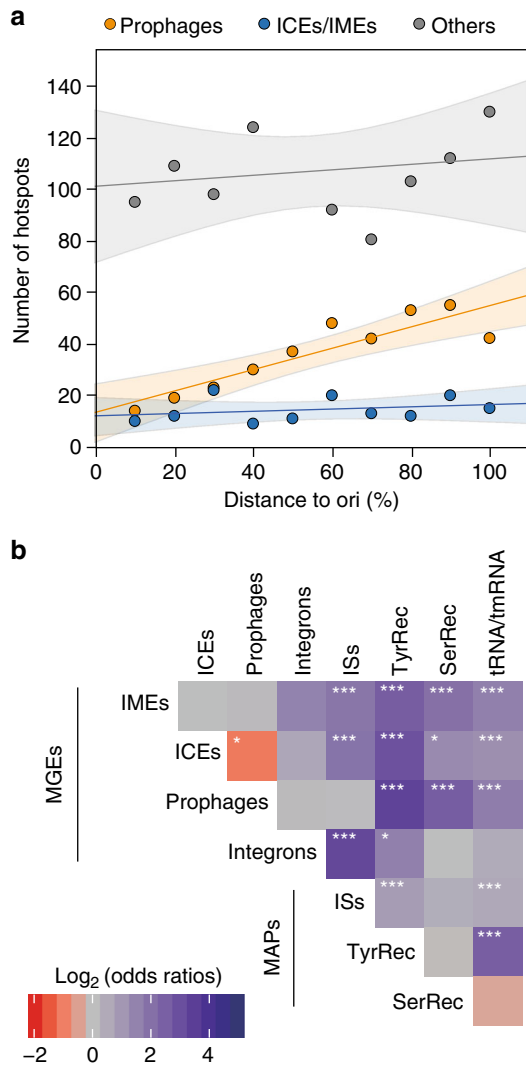


Fig. 6 Chromosomal context of hotspots. **a** Number of hotspots containing prophages, ICEs/IMEs, and none of the above along the origin-terminus axis of replication. Linear regression and the confidence limits (shaded area) for the expected value (mean) were indicated for each category. The number of hotspots including prophages increases linearly with the distance to the origin of replication (Spearman's $\rho = 0.87$, $P < 10^{-3}$), but this is not the case for the other two categories (both $P > 0.05$). **b** Heatmap of odds ratios of co-localizations in hotspots of MGEs, mobility-associated proteins (MAPs) and RNAs. *** $P < 10^{-3}$; ** $P < 10^{-2}$; * $P < 0.05$; Fisher's exact test

organization of bacterial genomes⁸, we analyzed the position of hotspots relative to the distance to the origin of replication along the replicore. These results showed that the frequency of hotspots including prophages, as previously shown in *E. coli*¹³, increases linearly along the ori->ter replication axis (Fig. 6a). Interestingly, this does not seem to be the case for ICEs and IMEs, nor for the very large category of hotspots that lack ICEs, IMEs, and prophages.

As these results show that prophages and ICEs have different distribution patterns, we quantified the frequency of co-occurrence of different MGEs and MAPs in the same hotspots (but not necessarily in the same intervals, Fig. 6b). In line with expectations, most MGEs significantly co-occurred with integrases, integrons, ISs, and tRNAs. The most notable exceptions concerned the prophages, that did not significantly

co-occurred with ISs, presumably because ISs are rare in phages²⁹, or integrons, and they were found less frequently than expected in spots with conjugative elements. This is in line with the analysis showing that they have specifically different distributions along the chromosome replication axis.

Genetic diversity of hotspots. The integration of a MGE in the chromosome adds a large number of genes in one single location, potentially creating a hotspot on itself. Such events result in a concentration of HTgenes in a genome (strain-specific integration), or in several genomes (when the integration took place at the last common ancestor of several strains). The distribution of the number of genomes with orthologous HTgenes in hotspots suggests that these cases are relatively rare (Supplementary Fig. 9a). Only 8% of the hotspots had all accessory gene families represented in one genome (Supplementary Fig. 9b, c). Hence, few of these regions seem to have been created by the integration of a single MGE.

To assess whether genetic diversity in hotspots was compatible with one single ancient integration event, we introduced measures derived from the analysis of beta diversity in Ecology, where it is used to measure the differences in species composition between different locations³⁰ (Methods). Here we used it to measure the difference in gene repertoires among a set of intervals from the same spot. We measured the Sørensen index (β_{SOR}) for hotspots and coldspots of each species using the binary matrix of gene presence/absence. Diversity results from a mixture of independent gene acquisitions and replacements (turnover) and differential gene loss (nestedness), and β_{SOR} can be partitioned into the two related additive terms: turnover (β_{SIM}) and nestedness (β_{NES}) ($\beta_{SOR} = \beta_{NES} + \beta_{SIM}$, Fig. 7a).

Beta diversity of accessory genes was higher in hotspots than in coldspots (Fig. 7b). This difference was caused by turnover, since only β_{SIM} was significantly higher in hotspots than in coldspots (Fig. 7c). The values of β_{NES} were very low in both cases; confirming that most hotspots are not caused by singular events of integration of MGEs. We obtained similar results when the analysis of diversity was restricted to HTgenes (Supplementary Fig. 10). While genetic diversity is high in hotspots and coldspots, these results show faster diversification in hotspots because they endure higher genetic turnover.

Finally, we wished to test whether hotspots lacking MAPs had such a high genetic turnover that MGEs would be rapidly removed. We split the hotspots into two categories: hotspots containing and lacking MAPs. Both categories showed values of genetic diversity close to one that were caused by high turnover. Nevertheless, hotspots lacking MAPs showed slightly lower values for these variables (Supplementary Fig. 11). Hence, the absence of MAPs in these hotspots is not due to an excess of genetic turnover.

Hotspots of homologous recombination. Many hotspots lack identifiable MGEs or even integrases. Yet, they show high genetic diversity, suggesting that other mechanisms may drive their evolution. We tested the possibility that these regions could integrate HTgenes by homologous recombination at the flanking core genes, as suggested for certain hotspots of *E. coli*²³ and *S. pneumoniae*²⁴ (Fig. 8a). Our hypothesis predicts higher levels of homologous recombination in core genes flanking hotspots than in the rest of the core genome. We tested this prediction in two complementary ways. Firstly, we detected homologous recombination events in the core genes using ClonalFrameML (Methods). We found 50% more recombination events in core genes flanking hotspots than in the other core genes (Fig. 8b). Secondly, we searched for evidence of phylogenetic incongruence between each

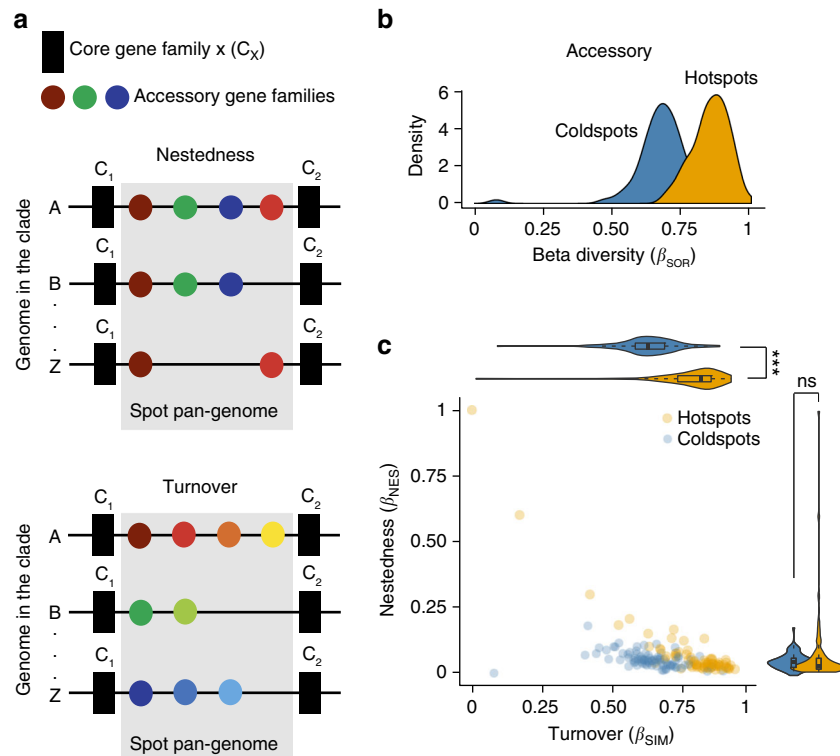


Fig. 7 Genetic diversity of the accessory genes present in hotspots and coldspots. **a** Examples of gene nestedness and turnover in a spot. Turnover measures the segregation between intervals in terms of gene families, i.e., it accounts for the replacement of some genes by others. Nestedness accounts for differential gene loss and measures how the gene repertoires of some intervals are a subset of the repertoires of the others. It typically reflects a non-random process of gene loss. **b** Distributions of β diversity (β_{SOR}) in hotspots and coldspots. **c** Partition of β_{SOR} in its components of nestedness (β_{NES}) and turnover (β_{SIM}) for hotspots and coldspots ($\beta_{SOR} = \beta_{NES} + \beta_{SIM}$). *** $P < 10^{-3}$; Mann–Whitney–Wilcoxon test; ns: not significant

core gene family and the whole core genome tree of the clade using the Shimodaira–Hasegawa (SH) test (Methods). The number of genes with significant phylogenetic incongruence was 30% higher among core genes flanking hotspots than among the others (Fig. 8b). In line with these observations, core genes flanking hotspots also had higher nucleotide diversity (Fig. 8c). We found qualitatively similar results when the analysis was performed on a per species or per genus basis (Supplementary Data set 5). Hence, core genes flanking hotspots are more targeted by recombination processes than the others.

Naturally transformable bacteria have the ability to acquire genetic material independently of MGEs. In these species, transfer of chromosomal material mediated by homologous recombination at the flanking core genes might be particularly frequent. To test this hypothesis, we put apart the 19 bacterial species that are known to be naturally transformable in our dataset³¹ (Supplementary Data set 1). We observed that these species had more hotspots than the others ($P < 0.05$, Mann–Whitney–Wilcoxon test). We searched for MAPs in these hotspots and observed that they also had fewer hotspots with MAPs ($P < 0.05$, Mann–Whitney–Wilcoxon test). Finally, recombination was 20% more frequent in core genes flanking hotspots in naturally transformable than in the remaining bacteria ($P < 10^{-4}$; χ^2 -test). These results suggest that recombination at core genes flanking hotspots might be particularly important in driving genetic diversification of naturally transformable bacteria.

Discussion

Our study showed high concentration of HTgenes in a small number of locations in the chromosomes of many bacterial species. These hotspots include most MGE-related genes, fitting

previous observations that the latter co-evolved with the host to use integrases targeting specific locations in the chromosome that minimize the fitness cost of chromosomal integration. For example, many temperate phages integrate tRNA genes without disrupting their function³². The concentration of most self-mobilizable MGEs at few loci might be thought sufficient to justify the existence of hotspots, but we found that few hotspots had identifiable prophages or conjugative elements and that most lacked integrases. These puzzling results could be caused by failure to identify MGEs, but our methods were shown to be highly accurate at identifying conjugative elements and prophages^{13, 33}, or by the presence of many radically novel integrase-lacking MGEs in these model microbial species, which would be very surprising. Hotspots also contain degenerate MGEs that we have failed to identify. Yet, inactivated elements are not expected to drive the observed rapid genetic turnover of these regions.

Our results suggest that an MGE-independent mechanism, double homologous recombination at the flanking core genes, contributes to hotspot diversification. The mechanism only requires housekeeping recombination functions and exogenous DNA with homology to the flanking core genes. This last condition is easy to fulfill, because these genes are present in all genomes of the species (and usually in closely related species). In agreement with our hypothesis, we showed that naturally transformable species had more hotspots, and fewer MAPs in hotspots, than the others. There are other mechanisms of transfer that can bring homologous sequences without MAPs in non-transformable bacteria, including generalized transduction, gene transfer agents, or DNA-carrying vesicles³⁴. Their role in hotspot diversification remains to be explored.

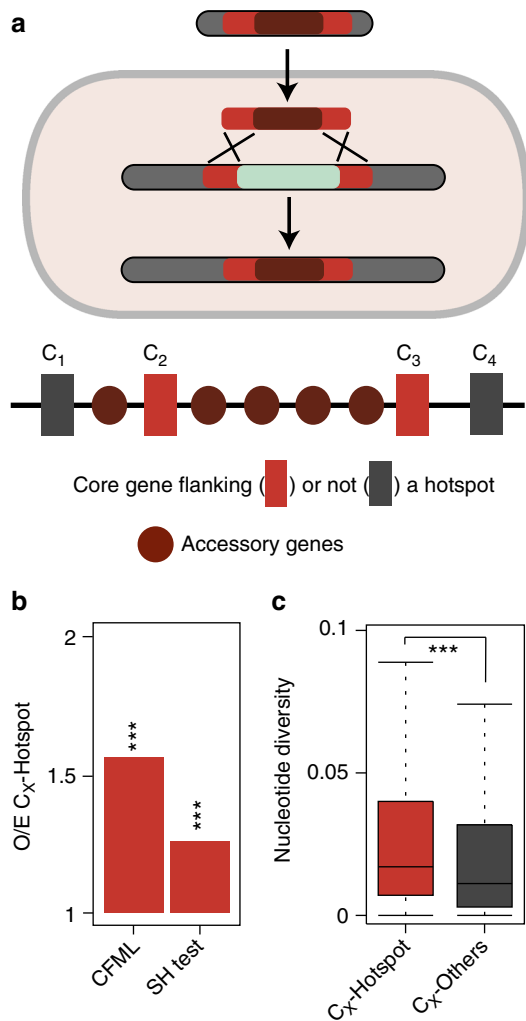


Fig. 8 Evidence for more frequent homologous recombination in core genes flanking hotspots than in the other core genes. **a** Model for the creation and evolution of hotspots by homologous recombination at the flanking core genes. **b** We detected homologous recombination events in the core genes using ClonalFrameML, and searched for evidence of phylogenetic incongruence ($P < 0.05$) between each core gene family and the whole core genome tree of the clade using the Shimodaira-Hasegawa (SH) test. The observed-expected ratios (O/E) for these two analyses are significantly higher than one. $***P < 10^{-3}$; χ^2 -test. **c** Differences in nucleotide diversity between core genes flanking hotspots and the others. Nucleotide diversity was calculated using the R package “PopGenome” v2.1.6⁴⁹ by implementation of the diversity.stats() command. $***P < 10^{-3}$; Mann-Whitney-Wilcoxon test

Many HTgenes are not adaptive (or even deleterious) and are rapidly lost by genetic drift (or purifying selection)^{6, 7, 35}. Nevertheless, regions of high concentration of HTgenes must also include adaptive genes, as shown here for ARGs. In these circumstances, the high genetic turnover at hotspots might seem paradoxical, because it may lead to their loss. Actually, even adaptive genes can be lost with little fitness cost under certain circumstances. Genes under diversifying selection, such as defense systems, may be adaptive for short periods of time and subsequently lost (or replaced by analogous genes)³⁶. Some costly genes may be adaptive in only very specific conditions, such as ARGs³⁷, and become deleterious for the cell fitness upon environmental change. Finally, some genes under

frequency-dependent selection, such as toxins³⁸, may stop being adaptive when their frequency changes in the population. Genetic drift, purifying, diversifying, and frequency-dependent selection can thus contribute to the rapid turnover of HTgenes. As a consequence of their high genetic turnover, hotspots are expected to be enriched in genes of specific adaptive value.

Hotspots may affect bacterial fitness not only by the genes they contain, but also by the way they drive genome diversification. According to the chromosome-curing model³⁹, hotspots may facilitate the elimination of elements with deleterious fitness effects, such as certain MGEs, by double recombination at the flanking core genes. This fits our observation that core genes flanking hotspots endure higher rates of homologous recombination. As a response to chromosome curing, natural selection is expected to favor MGEs that inactivate genes encoding recombination and repair proteins³⁹. Interestingly, we also found that hotspots tend to be flanked by recombination and repair core genes. Although these genes seem intact, at least they respect the constraints that we imposed for their classification as core genes, their expression may be affected by HTgenes in the neighboring hotspot. For example, excision of a MGE in *Vibrio splendidus* 12B01 from a *mutS* gene downregulates the expression of the latter leading to a hypermutator phenotype⁴⁰.

Several selective effects can contribute to explain the very different number of hotspots per species, which were strongly correlated with the number of HTgenes and weakly with genome size (itself also correlated with the rate of HGT²⁷). The first association may explain why species with little genetic diversity, such as *B. anthracis* and mycobacteria, have few hotspots in spite of their large genome size. It is also possible that our statistical tests lack power when species have few HTgenes. Some ecological determinants also affect the number of HTgenes, and their concentration in the genome. For example, sexually isolated species with few MGEs, such as obligatory endosymbionts, are expected to have few hotspots. Many of these species may also inefficiently select for hotspots because they have low effective population sizes. Conversely, the highest number of hotspots was found in facultative pathogens with very diverse gene repertoires, including *E. coli*, *Pseudomonas spp.*, and *Bacillus cereus*. A rigorous statistical assessment of the ecological traits affecting the organization of HTgenes will require the analysis of a larger panel of species representative of the different prokaryotic lifestyles.

Overall, our results suggest that hotspots are the result of the interplay of several recombination mechanisms and natural selection, presumably because they minimize disruption of genome organization by circumscribing gene flux to a small number of permissive chromosomal locations. For example, the increase in prophage-containing hotspots along the ori-→ ter axis suggests co-evolution between these elements and the host to remove prophages from early replicating regions that are also rich in highly expressed genes in fast growing bacteria¹³. Interestingly, the spatial distribution of the remaining hotspots does not show similar patterns, which can be due to the lower fitness costs associated with their excision. Further work is needed to understand if there are other organizational traits that constrain the distribution of hotspots in the chromosome, and in particular in those devoid of recognizable MGEs. Knowing these traits might facilitate large-scale genetic engineering and should lead to a better understanding of the evolutionary interactions between horizontal gene transfer and genome organization.

Finally, our study focused on the dynamics of hotspots and how they contribute to genome diversification, but left unanswered the questions related to their origin and fate. Previous studies identified common prophage hotspots between *E. coli* and *Salmonella enterica*¹³. Hence, we will have to study

taxonomical units broader than the species level to unravel their origin. As for their fate, long-term adaptive HTgenes may become fixed in the population, explaining the patterns of nestedness of certain hotspots, and leading eventually to the split of the hotspot into two new (eventually hot) spots.

Methods

Data. The sequences and annotations of 932 bacterial genomes from 80 bacterial species were retrieved from GenBank RefSeq (<ftp://ftp.ncbi.nih.gov/genomes>, last accessed in February 2014)⁴¹. We made no selection on the species that were to be analyzed, except that we required a minimum of four complete genomes per species. We have made no attempt to re-define species: we used the information presented in GenBank. Their list is available in Supplementary Data set 1. We excluded CDS annotated as partial genes, as well as those lacking a stop codon or having stop codons within the reading frame. Core genomes and phylogenetic reconstructions were obtained from our previous work²⁷ (Supplementary Data set 2). Our data set includes several species from the same genera. It also includes species with diverse numbers of genomes and HTgenes. To minimize the effects of these unavoidable biases most of our analyses are non-parametric and each species has the same weight. When they were done on the data cumulated from all species, we made a control where each species is analyzed separately. We also made complementary analyses where we aggregated the results per genus. The references for these supplementary controls are indicated in the main text, and the data are in the Supplementary Material.

Identification of core genomes. We used 80 core genomes previously published²⁷. These core genomes were built for clades with at least four complete genomes available in GenBank RefSeq (Supplementary Data set 1, Supplementary Fig. 1). Briefly, a preliminary list of orthologs was identified as reciprocal best hits using end-gap-free global alignment, between the proteome of a reference genome (pivot, typically the first completely sequenced isolate) and each of the other strain's proteomes. Hits with <80% similarity in amino-acid sequence or >20% difference in protein length were discarded. This list of orthologs was then refined for every pairwise comparison using information on the conservation of gene neighborhood. Thus, positional orthologs were defined as bi-directional best hits adjacent to at least four other pairs of bi-directional best hits within a neighborhood of 10 genes (five upstream and five downstream). These parameters (four genes being less than half of the diameter of the neighborhood) allow retrieving orthologs at the edge of rearrangement breakpoints (positions where intervals were split by events of chromosome rearrangement) and therefore render the analysis robust to the presence of a few rearrangements. The core genome of each clade was defined as the intersection of pairwise lists of positional orthologs.

Definitions of interval and spot. The core genome is the collection of all gene families present in one and only one copy in each genome of a clade (Supplementary Fig. 2). Let C_X and C_Y be two families of core genes in a clade with N taxa where one of the taxa is a pivot (reference genome, see above). We call C_{AX} and C_{AY} contiguous core genes in a given chromosome A if they are adjacent in the list of core genes sorted in terms of the position in the chromosome. We defined an interval $(I_{AX, AY})$ as the location between the pair of contiguous core genes C_{AX} and C_{AY} in chromosome A . The content of an interval is the set of accessory genes in the interval. The HTgenes content of an interval is the number of genes that were acquired by HGT in the interval. Multiple chromosomes, when present, were treated independently.

Intervals flanked by the same core gene families (C_X, C_Y) as the pivot genome were defined as syntenic intervals (i.e., the members of the core gene families X and Y were also contiguous in the pivot). The intervals that do not satisfy this constraint were classed as breakpoint intervals and excluded from our analysis. They contain <2% of all genes. For every interval in the pivot genome, we defined spot as the set of syntenic intervals flanked by members of the same pair of core gene families (Supplementary Fig. 2).

Identification of spot pan-genomes. The pan-genome is the full complement of homologous gene families in a clade. We built a pan-genome for each species using the gene repertoire of each genome. Initially, we determined a preliminary list of putative homologous proteins between pairs of genomes (excluding plasmids) by searching for sequence similarity between each pair of proteins with BLASTP v.2.2.28+ (default parameters). We then used the e -values ($<10^{-4}$) of the BLASTP output to cluster them using SILIX (v1.2.8, <http://lbb.e.univ-lyon1.fr/SiLiX>)⁴². We set the parameters of SILIX such that two proteins were clustered in the same family if the alignment had at least 80% identity and covered >80% of the smallest protein (options -l 0.8 and -r 0.8). We computed the diversity of gene families observed in each spot. The spot pan-genome is the set of gene families present in the intervals associated with the spot (Supplementary Fig. 2).

Reconstruction of the evolution of gene repertoires. We assessed the evolutionary dynamics of gene repertoires of each clade using Count⁴³ (downloaded in

April 2015). This program uses birth-death models to identify the rates of gene deletion, duplication, and loss in each branch of a phylogenetic tree. We used the spots' pan-genomes matrices, and the phylogenetic birth-and-death model of Count, to evaluate the most likely scenario for the evolution of a given gene family on the clade's tree. Rates were computed with default parameters, assuming a Poisson distribution for the family size at the tree root, and uniform gain, loss, and duplication rates. One hundred rounds of rate optimization were computed with a convergence threshold of 10^{-3} . After optimization of the branch-specific parameters of the model, we performed ancestral reconstructions by computing the branch-specific posterior probabilities of evolutionary events, and inferred the gains in the terminal branches of the tree. The posterior probability matrix was converted into a binary matrix of presence/absence of HTgenes using a threshold probability of gain higher than 0.95 at the terminal branches and excluding gains occurring in the last common ancestor with a probability higher than 0.5.

Identification of hotspots. We made simulations to obtain the expected distribution of HTgenes in the spots given the numbers of HTgenes and spots (Supplementary Fig. 4). We made the null hypothesis that the distribution of these genes was constrained by the frequency of genes in operons, and followed a uniform distribution in all other respects. Previous works have shown that two-third of the genes are in operons and one-third are in mono-cistronic units⁴⁴, with little inter-species variation for the average length of poly-cistronic units (3.15 ± 0.06)⁴⁵. Hence, given N HTgenes per clade we created two groups of elements: $N/3$ isolated genes and $2N/3$ in operons with three genes. These elements were then randomly placed among the spots following a uniform distribution. For each of the 1000 simulations (per species), we recorded the maximal value of genes within a single spot ($\text{Max}_{\text{HTg},i}$), which was used to identify the value of the 95th percentile ($T_{95\%}$) of the distribution of $\text{Max}_{\text{HTg},i}$. Hence, 95% of the simulations have no spot with more than $T_{95\%}$ genes (Supplementary Data set 1). Spots (in the real genomes) with more than $T_{95\%}$ HTgenes were regarded as hotspots. Spots lacking accessory genes were called empty spots. The other spots were called coldspots.

As a control, we also made simulations considering that HTgenes were acquired independently of the structure in operons (i.e., considering N isolated genes). The values of $T_{95\%}$ of the two analyses were highly correlated (Spearman's $\rho = 0.89$, $P < 10^{-4}$, Supplementary Data set 1), but those of the latter were smaller (linear regression: $T_{95\%} \text{ isolated} = -0.62 + 0.66 T_{95\%} \text{ operons}$, $R^2 = 0.87$). This is expected because the operon structure should increase the variance of the genes per spot, and thus increase $T_{95\%}$.

Measures of gene repertoire diversity. Since most spots have few or no genes, and most gene families have few (or no gene) per genome, we computed the genetic diversity of spots using matrices of presence/absence of gene families (computed from the pan-genome).

We computed beta diversity per clade, using a multiple-site version (each interval is the equivalent of a site)⁴⁶ of the widely used Sørensen dissimilarity index (β_{SOR}):

$$\beta_{\text{SOR}} = \frac{\sum_{i < j} \min(b_{ij}, b_{ji}) + \sum_{i < j} \max(b_{ij}, b_{ji})}{2 \cdot (\sum_i S_i - S_T) + \sum_{i < j} \min(b_{ij}, b_{ji}) + \sum_{i < j} \max(b_{ij}, b_{ji})}, \quad (1)$$

where S_i is the total number of accessory genes in genome i , S_T is the total number of accessory genes in all genomes considered together, and b_{ij}, b_{ji} are the numbers of accessory genes present in genome i but not in j ($b_{i,j}$) and vice-versa ($b_{j,i}$).

We then used a partitioned version of the ecological concept of beta diversity to characterize the gene diversity of spots⁴⁶. β_{SOR} can be partitioned into two additive terms: turnover (β_{SIM}) and nestedness (β_{NES}) ($\beta_{\text{SOR}} = \beta_{\text{NES}} + \beta_{\text{SIM}}$, Fig. 7a).

To compute the turnover we used the multiple-site version⁴⁶ of the Simpson dissimilarity index (β_{SIM}):

$$\beta_{\text{SIM}} = \frac{\sum_{i < j} \min(b_{ij}, b_{ji})}{(\sum_i S_i - S_T) + \sum_{i < j} \min(b_{ij}, b_{ji})}, \quad (2)$$

This index is a measure of the evenness with which families of genes are distributed across intervals of a spot (it is a measure of segregation). Turnover implies the replacement of some gene families by others.

By definition, the multiple-site dissimilarity term accounting only for nestedness (β_{NES}) results from the subtraction⁴⁶:

$$\beta_{\text{NES}} = \beta_{\text{SOR}} - \beta_{\text{SIM}}. \quad (3)$$

Nestedness occurs when intervals with fewer genes are subsets of intervals with larger gene repertoires. It reflects a non-random process of gene loss.

The above formulae were computed as follows: first, we plotted the distribution of the number of accessory genes from the hotspots of all clades analyzed. We took the minimum of this distribution (min_d) and used it to select coldspots with a number of accessory genes equal or higher than min_d . By doing this, we eliminated coldspots with very few accessory genes, and likely to introduce a bias while computing diversity (leading to extreme situations where $\sum_i S_i \approx S_T$; $\beta_{\text{SIM}} \approx 1$, and as consequence $\beta_{\text{NES}} \approx 0$). After this filtering step, we put together all hotspots and all coldspots of each genome in two separate concatenates to avoid statistical artifacts

associated with poorly populated spots. The diversity was computed per clade for each of the concatenates.

Inference of homologous recombination. We inferred homologous recombination on the multiple alignments of the core genes of each clade using ClonalFrameML (CFML) v10.7.5⁴⁷ with a predefined tree (i.e., the clade's tree), default priors $R/\theta = 10^{-1}$, $1/\delta = 10^{-3}$, and $\nu = 10^{-1}$, and 100 pseudo-bootstrap replicates, as previously suggested⁴⁷. Mean patristic branch lengths were computed with the R package "ape" v3.3⁴⁸, and transition/transversion ratios were computed with the R package "PopGenome" v2.1.6⁴⁹. The priors estimated by this mode were used as initialization values to rerun CFML under the "per-branch model" mode with a branch dispersion parameter of 0.1.

Functional assignment. Gene functional assignment was performed by searching for protein similarity with HMMer (hmmsearch) on the bactNOG subset of the eggNOG v4.5 database⁵⁰ (downloaded in March 2016). We have considered the pivot (reference) genomes as good representatives of each clade, and limited our analysis to these. We have kept hits with an *e*-value lower than 10^{-5} , a minimum alignment coverage of 50%, and when the majority (>50%) of non-supervised orthologous groups (NOGs) attributed to a given gene pertained to the same functional group. Hits corresponding to poorly characterized or unknown functional groups were discarded.

Identification of MGEs and proteins associated to mobility. Temperate phages integrated in the bacterial chromosome (prophages) were identified using Phage Finder v4.6⁵¹ (stringent option). Prophages with > 25% of the predicted genes belonging to ISs, and partially degraded prophages (shorter than 30 kb) were removed⁵². Integrons were identified using IntegronFinder v.1.4 with the -local_max option⁵³. Integrative conjugative elements (ICEs) and integrative mobilizable elements (IMEs) were identified using MacSyFinder v.1.0.2⁵⁴ with TXSScan profiles⁵⁵. Elements with a full conjugative apparatus were classed as ICE, the others as IME (see ref. ³³ for criteria). Integrases were identified using the PFAM profiles PF00589 for tyrosine recombinases, and the pair of profiles PF00239 and PF07508 for serine recombinases (<http://pfam.xfam.org/>)⁵⁶. All the protein profiles were searched using hmmsearch from the HMMer suite v.3.1b1 (default parameters). Hits were regarded as significant when their *e*-value was smaller than 10^{-3} and their alignment covered at least 50% of the protein profile. Insertion sequences (ISs) were detected combining two approaches (i) using hmmsearch from the HMMer with IS HMM profiles (as previously proposed)⁵⁷ and (ii) by a BLAST-based method using the ISFinder database⁵⁸. Integrases and transposases were defined as mobility-associated proteins (MAPs). tRNA genes were identified using tRNAscan SE v.1.21⁵⁹, tmRNA genes were identified using Aragorn v.1.2.37⁶⁰, and the location of the rRNA genes was taken from the Genbank annotation file. Antibiotic resistance genes were detected using HMMer against the curated database of antibiotic resistance protein families ResFams (Core v.1.2, <http://www.dantaslab.org/resfams>)⁶¹ using the '-cut_ga' option. A hotspot was considered to encode a peculiar MGE or MAP when at least one genome of the clade contained such element (Supplementary Data set 3).

Identification of origin and terminus of replication. The ori and ter of replication were predicted using Ori-Finder in the pivot genome of each clade⁶². When the ratio of the predicted replicore length was greater than 1.2, the clade was removed from the analysis (Supplementary Data set 4). Then, we divided each replicore in 10 equally sized regions from the ori to the ter of replication.

Phylogenetic analyses. We retrieved the 16S rRNA sequences of the sequenced type strains (also used as reference genomes, see above) of the 80 bacterial clades (Fig. 3a). We made a multiple alignment of them with MAFFT v7.305b⁶³ using default settings, and removed poorly aligned regions with BMGE v1.12⁶⁴ using default settings. The tree was computed by maximum likelihood with PHYML v3.0⁶⁵ under the general time reversible (GTR) + $\Gamma(4)$ + I model (Supplementary Data set 2a). This tree is never used in the calculations; it is only used in Fig. 3a to display the relative position of each clade in the phylogeny of bacteria.

We built core genome trees for each clade using a concatenate of the multiple alignments of the core genes (see main text). Each clade's tree was computed with RAxML v8.00⁶⁶ under the GTR model and a gamma correction (GAMMA) for variable evolutionary rates. All trees are shown in Supplementary Data set 2b. We performed 100 bootstrap experiments on the concatenated alignments to assess the robustness of the topology of each clade's tree. The vast majority of nodes were supported with bootstrap values higher than 90% (Supplementary Data set 2b). We inferred the root of each phylogenetic clade's tree using the midpoint-rooting approach of the R package "phangorn" v1.99.14⁶⁷. The alignment and the tree for each individual core gene were used for topology testing against the clade's tree (i.e., the concatenate tree of all the core genes of the clade) using the Shimodaira-Hasegawa (SH) congruence test⁶⁸ (1000 replicates) implemented in IQ-Tree v1.4.3⁶⁹.

Data availability. The authors declare that all data supporting the findings of this study are available within the article and its Supplementary Information files and from the corresponding author upon reasonable request.

Received: 6 March 2017 Accepted: 31 July 2017

Published online: 10 October 2017

References

- Wilmes, P., Simmons, S. L., Denef, V. J. & Banfield, J. F. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol. Rev.* **33**, 109–132 (2009).
- Treangen, T. J. & Rocha, E. P. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).
- Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).
- van Passel, M. W., Marri, P. R. & Ochman, H. The emergence and fate of horizontally acquired genes in *Escherichia coli*. *PLoS Comput. Biol.* **4**, e1000059 (2008).
- Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
- Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17**, 589–596 (2001).
- Koskineniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-driven gene loss in bacteria. *PLoS Genet.* **8**, e1002787 (2012).
- Rocha, E. P. The organization of the bacterial genome. *Annu. Rev. Genet.* **42**, 211–233 (2008).
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96**, 2896–2901 (1999).
- Vieira-Silva, S. & Rocha, E. P. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* **6**, e1000808 (2010).
- Sharp, P. M., Shields, D. C., Wolfe, K. H. & Li, W. H. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **246**, 808–810 (1989).
- Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Bobay, L. M., Rocha, E. P. & Touchon, M. The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol. Evol.* **30**, 737–751 (2013).
- Frost, L. S., Leplae, R., Summers, A. O. & Toussaint, A. Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).
- Burrus, V., Pavlovic, G., Decaris, B. & Guedon, G. Conjugative transposons: the tip of the iceberg. *Mol. Microbiol.* **46**, 601–610 (2002).
- Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**, 238–276 (2003).
- Asadulghani, M. et al. The defective prophage pool of *Escherichia coli* O157: prophage-prophage interactions potentiate horizontal transfer of virulence determinants. *PLoS Pathog.* **5**, e1000408 (2009).
- Sullivan, J. T. & Ronson, C. W. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. *Proc. Natl Acad. Sci. USA* **95**, 5145–5149 (1998).
- Murphy, K. C. Phage recombinases and their applications. *Adv. Virus Res.* **83**, 367–414 (2012).
- Balbontin, R., Figueroa-Bossi, N., Casadesu, J. & Bossi, L. Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in *Salmonella enterica*. *J. Bacteriol.* **190**, 4075–4078 (2008).
- Boyd, E. F., Almagro-Moreno, S. & Parent, M. A. Genomic islands are dynamic, ancient integrative elements in bacterial evolution. *Trends Microbiol.* **17**, 47–53 (2009).
- Williams, K. P. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* **30**, 866–875 (2002).
- Touchon, M. et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).
- Croucher, N. J. et al. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
- Chancey, S. T. et al. Composite mobile genetic elements disseminating macrolide resistance in *Streptococcus pneumoniae*. *Front. Microbiol.* **6**, 26 (2015).
- Everitt, R. G. et al. Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. *Nat. Commun.* **5**, 3956 (2014).
- Oliveira, P. H., Touchon, M. & Rocha, E. P. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. USA* **113**, 5658–5663 (2016).
- Homma, K., Fukuchi, S., Nakamura, Y., Gojobori, T. & Nishikawa, K. Gene cluster analysis method identifies horizontally transferred genes with high

- reliability and indicates that they provide the main mechanism of operon gain in 8 species of gamma-Proteobacteria. *Mol. Biol. Evol.* **24**, 805–813 (2007).
29. Leclercq, S. & Cordaux, R. Do phages efficiently shuttle transposable elements among prokaryotes? *Evolution* **65**, 3327–3331 (2011).
 30. Koleff, P. Measuring beta diversity for presence–absence data. *J. Anim. Ecol.* **72**, 367–382 (2003).
 31. Johnston, C., Martin, B., Fichant, G., Polard, P. & Claverys, J. P. Bacterial transformation: distribution, shared mechanisms and divergent control. *Nat. Rev. Microbiol.* **12**, 181–196 (2014).
 32. Campbell, A. Prophage insertion sites. *Res. Microbiol.* **154**, 277–282 (2003).
 33. Guglielmini, J., Quintais, L., Garcillan-Barcia, M. P., de la Cruz, F. & Rocha, E. P. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation. *PLoS Genet.* **7**, e1002222 (2011).
 34. Garcia-Aljaro, C., Balleste, E. & Muniesa, M. Beyond the canonical strategies of horizontal gene transfer in prokaryotes. *Curr. Opin. Microbiol.* **38**, 95–105 (2017).
 35. Kuo, C. H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
 36. Oliveira, P. H., Touchon, M. & Rocha, E. P. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
 37. Andersson, D. I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8**, 260–271 (2010).
 38. Levin, B. R. Frequency-dependent selection in bacterial populations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **319**, 459–472 (1988).
 39. Croucher, N. J. et al. Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. *PLoS Biol.* **14**, e1002394 (2016).
 40. Chu, N. D. et al. A Mobile element in *mutS* drives hypermutation in a marine *Vibrio*. *MBio.* **8**, e02045–16 (2017).
 41. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
 42. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).
 43. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
 44. Touchon, M. & Rocha, E. P. Coevolution of the organization and structure of prokaryotic genomes. *Cold Spring Harb. Perspect. Biol.* **8**, a018168 (2016).
 45. Zheng, Y., Szustakowski, J. D., Fortnow, L., Roberts, R. J. & Kasif, S. Computational identification of operons in microbial genomes. *Genome Res.* **12**, 1221–1230 (2002).
 46. Baselga, A. Partitioning the turnover and nestedness components of beta diversity. *Global Ecol. Biogeogr.* **19**, 134–143 (2010).
 47. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
 48. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
 49. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
 50. Powell, S. et al. eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**, D231–D239 (2014).
 51. Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
 52. Bobay, L. M., Touchon, M. & Rocha, E. P. Manipulating or superseding host recombination functions: a dilemma that shapes phage evolvability. *PLoS Genet.* **9**, e1003825 (2013).
 53. Cury, J., Jove, T., Touchon, M., Neron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
 54. Abby, S. S., Neron, B., Menager, H., Touchon, M. & Rocha, E. P. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS ONE* **9**, e110726 (2014).
 55. Abby, S. S. et al. Identification of protein secretion systems in bacterial genomes. *Sci. Rep.* **6**, 23080 (2016).
 56. Finn, R. D. et al. The Pfam protein families database. *Nucleic Acids Res.* **36**, D281–D288 (2008).
 57. Kamoun, C., Payen, T., Hua-Van, A. & Filee, J. Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics* **14**, 700 (2013).
 58. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
 59. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
 60. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**, 11–16 (2004).
 61. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).
 62. Gao, F. & Zhang, C. T. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinformatics* **9**, 79 (2008).
 63. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
 64. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC. Evol. Biol.* **10**, 210 (2010).
 65. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
 66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
 67. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
 68. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1116 (1999).
 69. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
 70. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* **39**, W475–W478 (2011).

Acknowledgements

This work was supported by an European Research Council grant (EVOMOBILOME no. 281605). J.C. is a member of the 'Ecole Doctorale Frontière du Vivant (FdV)—Programme Bettencourt'. We thank the members of the Microbial Evolutionary Genomics group for comments and suggestions on the manuscript.

Author contributions

P.H.O., M.T. and E.P.C.R. designed the research; P.H.O., M.T. and E.P.C.R. analyzed the data; J.C. provided data and tools; P.H.O., M.T. and E.P.C.R. wrote the paper.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00808-w.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

Epigenomic characterization of *Clostridioides difficile* finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis

Pedro H. Oliveira¹, John W. Ribis², Elizabeth M. Garrett³, Dominika Trzilova³, Alex Kim¹, Ognjen Sekulovic², Edward A. Mead¹, Theodore Pak¹, Shijia Zhu¹, Gintaras Deikus¹, Marie Touchon^{4,5}, Martha Lewis-Sandari¹, Colleen Beckford¹, Nathalie E. Zeitouni¹, Deena R. Altman^{1,6}, Elizabeth Webster¹, Irina Oussenko¹, Supinda Bunyavanich¹, Aneel K. Aggarwal⁷, Ali Bashir¹, Gopi Patel⁶, Frances Wallach⁶, Camille Hamula⁶, Shirish Huprikar⁶, Eric E. Schadt^{1,8}, Robert Sebra¹, Harm van Bakel¹, Andrew Kasarskis¹, Rita Tamayo³, Aimee Shen^{1,2*} and Gang Fang^{1*}

***Clostridioides* (formerly *Clostridium*) *difficile* is a leading cause of healthcare-associated infections. Although considerable progress has been made in the understanding of its genome, the epigenome of *C. difficile* and its functional impact has not been systematically explored. Here, we perform a comprehensive DNA methylome analysis of *C. difficile* using 36 human isolates and observe a high level of epigenomic diversity. We discovered an orphan DNA methyltransferase with a well-defined specificity, the corresponding gene of which is highly conserved across our dataset and in all of the approximately 300 global *C. difficile* genomes examined. Inactivation of the methyltransferase gene negatively impacts sporulation, a key step in *C. difficile* disease transmission, and these results are consistently supported by multiomics data, genetic experiments and a mouse colonization model. Further experimental and transcriptomic analyses suggest that epigenetic regulation is associated with cell length, biofilm formation and host colonization. These findings provide a unique epigenetic dimension to characterize medically relevant biological processes in this important pathogen. This study also provides a set of methods for comparative epigenomics and integrative analysis, which we expect to be broadly applicable to bacterial epigenomic studies.**

Clostridioides difficile is a spore-forming Gram-positive obligate anaerobe and the leading cause of nosocomial antibiotic-associated disease in the developed world¹ (Supplementary Notes). Despite the substantial progress that has been achieved in the understanding of *C. difficile* physiology, genetics and genomic evolution^{2,3}, the roles played by epigenetic factors—namely DNA methylation—have not been systematically studied^{4–6}. In the bacterial kingdom, there are three major forms of DNA methylation: *N*⁶-methyladenine (6mA; the most prevalent form representing ~80%), *N*⁴-methylcytosine (4mC) and 5-methylcytosine (5mC). Increasing evidence suggests that DNA methylation regulates a number of biological processes, including DNA replication and repair, cell cycle, chromosome segregation and gene expression^{7–13}. Efficient high-resolution mapping of bacterial DNA-methylation events has only recently become possible with the advent of single-molecule real-time sequencing (SMRT-seq)^{14,15}. This technique enabled the characterization of the first bacterial methylomes^{16,17} and, since then, more than 2,200 (as of September 2019) have been mapped, heralding a new era of bacterial epigenomics¹⁸.

Here we mapped and characterized the DNA methylomes of 36 human *C. difficile* isolates using SMRT-seq and comparative

epigenomics. We observed substantial epigenomic diversity across *C. difficile* isolates, as well as the presence of a highly conserved methyltransferase (MTase). Inactivation of this MTase had a functional impact on sporulation, a key step in the transmission of *C. difficile*. Further experimental and integrative transcriptomic analysis suggested that epigenetic regulation by DNA methylation also modulates the cell length, host colonization and biofilm formation of *C. difficile*. These discoveries are expected to stimulate future investigations along a new epigenetic dimension to characterize and potentially repress medically relevant biological processes in this important pathogen.

Results

Methylome analysis reveals great epigenomic diversity in *C. difficile*. From an ongoing Pathogen Surveillance Program at Mount Sinai Medical Center, 36 *C. difficile* isolates were collected from faecal samples of infected patients (Supplementary Table 1). A total of 15 different multilocus sequence types (STs) belonging to clades 1 (human and animal, HA1) and 2 (hypervirulent or epidemic)¹⁹ are represented in our dataset (Fig. 1a). Using SMRT-seq with long-library-size selection, de novo genome assembly was achieved at

¹Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, NY, USA. ²Department of Molecular Biology and Microbiology, Tufts University School of Medicine, Boston, MA, USA. ³Department of Microbiology and Immunology, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, NC, USA. ⁴Microbial Evolutionary Genomics, Institut Pasteur, Paris, France. ⁵CNRS, UMR3525, Paris, France. ⁶Department of Medicine, Division of Infectious Diseases, Mount Sinai School of Medicine, New York, NY, USA. ⁷Department of Pharmacological Sciences and Department of Oncological Sciences, Mount Sinai School of Medicine, New York, NY, USA. ⁸Sema4, Stamford, CT, USA. *e-mail: aimee.shen@tufts.edu; gang.fang@mssm.edu

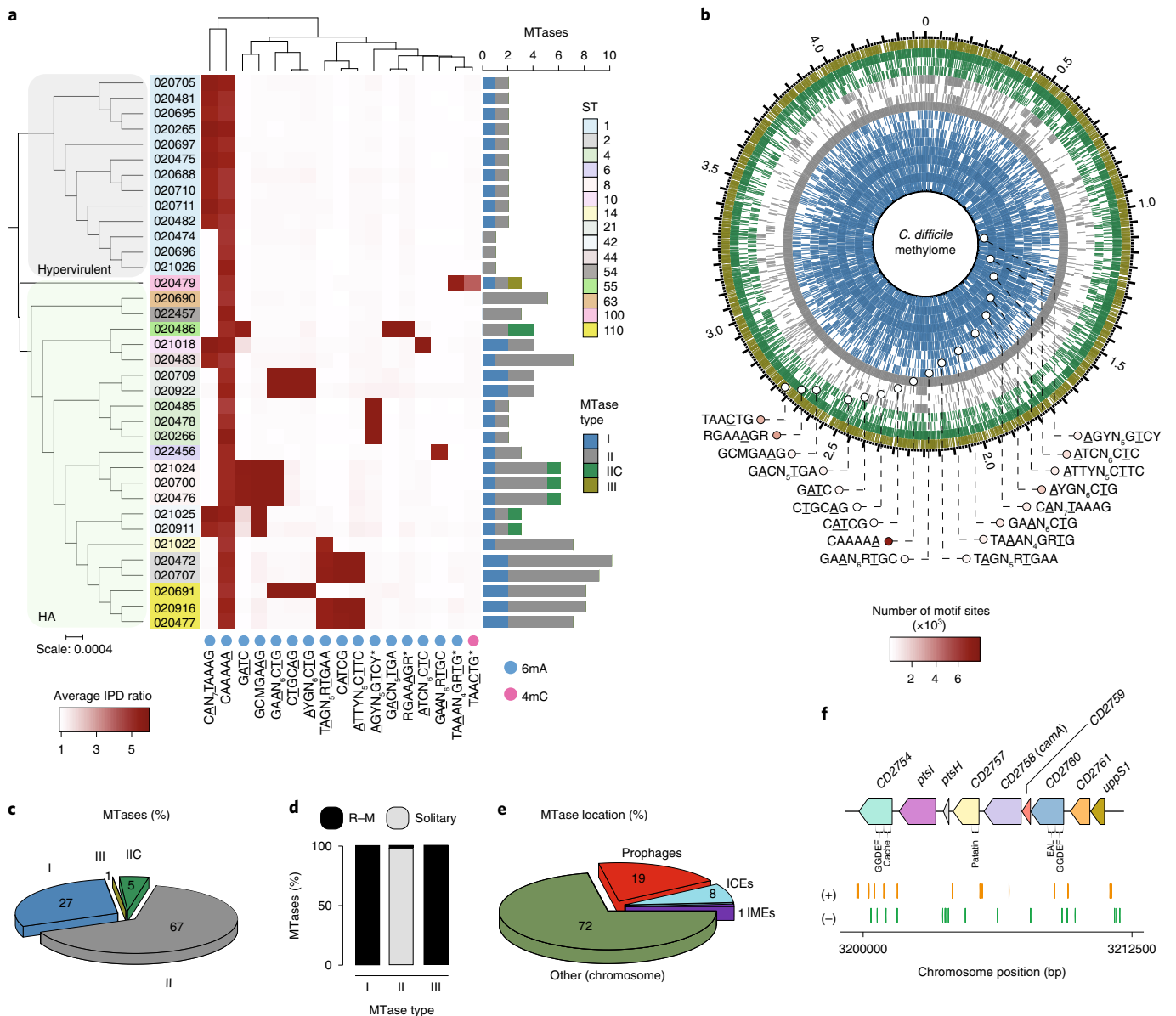


Fig. 1 | The methylomes of the 36 strains of *C. difficile*. **a**, A phylogenetic tree of the 36 *C. difficile* strains coloured by clade (hypervirulent, human and animal (HA) associated) and ST. A heat map of the landscape of methylated motifs per genome, and their average interpulse duration (IPD) ratio is also shown. The asterisks indicate new motifs that were not previously listed in the reference database REBASE. Methylated bases are underlined. The CAAAAA motif was consistently methylated across isolates. The bar plot indicates the number and types of active MTases detected per genome. In type IIC systems, MTase and restriction endonuclease (REase) are encoded in the same polypeptide. **b**, The *C. difficile* methylome. The positions of all of the methylation motif sites in the reference genome of *C. difficile* 630 are indicated, coloured according to MTase type. The average motif occurrences per genome (across the 36 isolates) are also indicated. **c**, The percentage of MTases detected according to type. **d**, The percentage of MTases pertaining to complete R-M systems or without cognate REase (solitary). **e**, Breakdown of MTases by location—integrative mobile elements (IMEs), integrative conjugative elements (ICEs), prophages and other (within the chromosome). No hits were obtained in plasmids. **f**, Immediate genomic context of *camA*. The example shown (including coordinates) refers to the reference genome of *C. difficile* 630. The plus and minus signs indicate the sense and antisense strands, respectively. The vertical bars indicate the distribution of the CAAAAA motif. CD2754, a phosphodiesterase with a GGDEF domain (PFAM PF00990) and a cache domain (PF02743); *ptsI* and *ptsH* belong to a phosphotransferase (PTS) system; CD2757, patatin-like phospholipase (PF01734); CD2758 (*camA*), type II MTase; CD2759, Rrf2-type transcriptional regulator; CD2760, phosphodiesterase with a GGDEF domain and a conserved EAL domain (PF00563); CD2761, *N*-acetylmuramoyl-L-alanine amidase; *uppS1*, undecaprenyl diphosphate synthase. The genomic context of *camA* is largely conserved across strains, located approximately 25 kb upstream of the *S*-layer biogenesis locus (Extended Data Fig. 4c,d). Several of the genes that flank *camA* (in addition to *camA* itself) are part of the *C. difficile* core genome (see below), suggesting that they may have biological functions that are fundamental to *C. difficile*.

a high quality (Supplementary Table 1). Methylation motifs were found using the SMRTportal protocol. We found a total of 17 unique high-quality methylation motifs in the 36 genomes (an average of 2.6 motifs per genome; Fig. 1a, Supplementary Table 2a). The large majority of target motifs were of the 6mA type, one motif

(TAACTG; methylated bases are underlined) belonged to the 4mC type and no 5mC motifs were detected with confidence (Supplementary Notes). As with most bacterial methylomes, more than 95% of the 6mA and 4mC motif sites were methylated (Fig. 1b, Supplementary Table 2a).

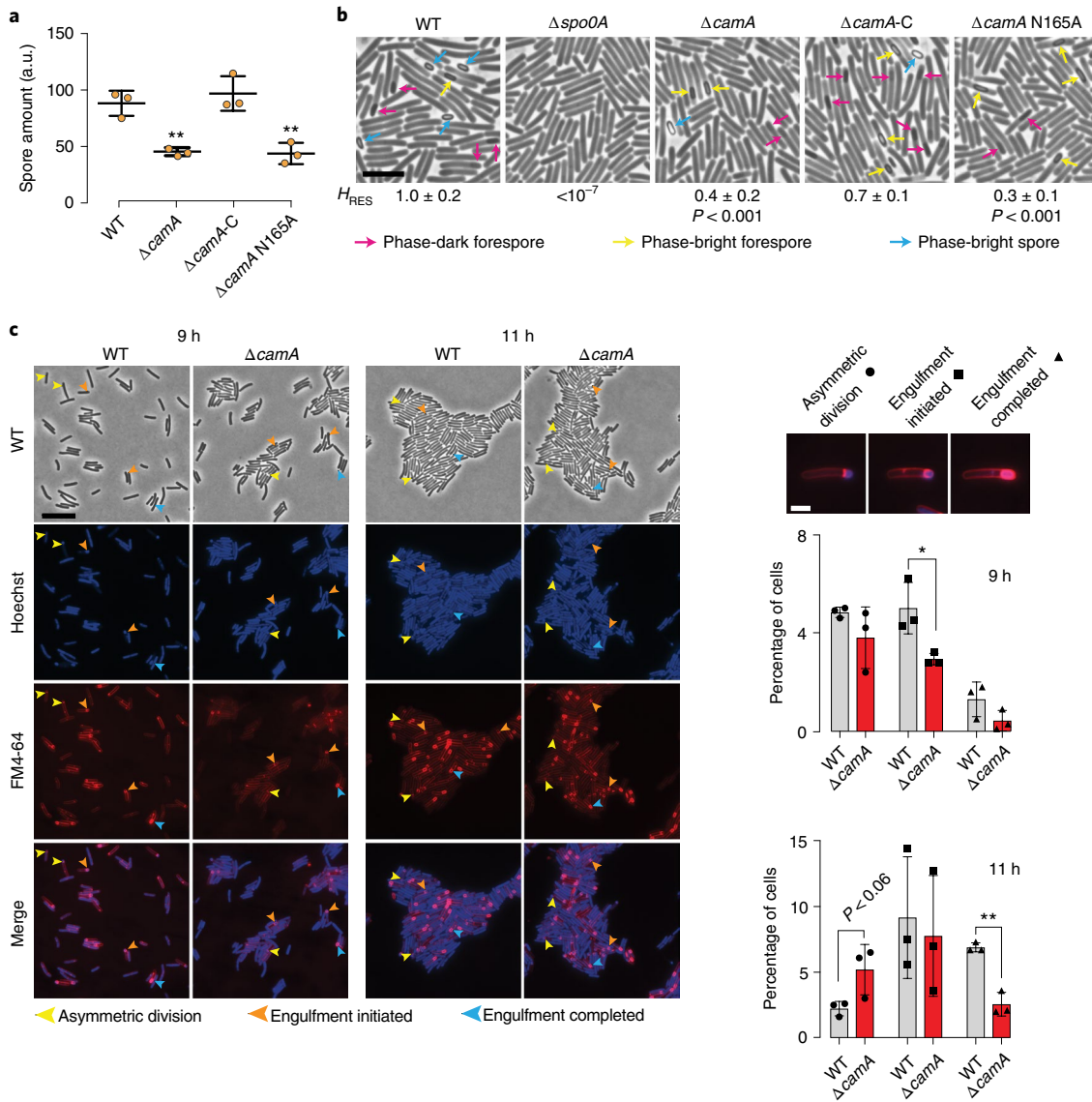


Fig. 2 | CamA modulates sporulation levels in *C. difficile*. **a**, Spore purification efficiencies obtained from sporulating cells; $n = 3$ independent spore preparations. ****** $P < 10^{-2}$; the statistical analysis was performed using one-way ANOVA and Tukey's test. The spore yield was determined by measuring the optical density at 600 nm (OD_{600}) of the resulting spore preparations and correcting for the volume of resuspension water. Data are mean \pm s.d. **b**, Phase-contrast microscopy after 20 h of sporulation induction. The $\Delta spo0A$ strain was used as a negative control because it does not initiate sporulation⁴⁴. Immature phase-dark forespores are indicated by pink arrows, and mature phase-bright forespores and free spores are indicated by yellow and blue arrows, respectively. Scale bar, 5 μ m. Heat-resistance (H_{RES}) efficiency data are mean \pm s.d. of $n = 3$ independent replicates. **c**, Morphological analysis of WT and $\Delta camA$ cells using fluorescent stains comparing 9 h and 11 h after sporulation induction. The polar septum formed during asymmetric division is visible using FM4-64 membrane staining, whereas the chromosome that is pumped into the forespore after polar-septum formation can be seen as a bright focus using Hoechst DNA staining. FM4-64 staining enables the visualization of engulfing membranes. As the mother-cell-derived membrane fully encircles the forespore-derived membrane, the FM4-64 signal becomes more intense around the forespore. When these membranes undergo fission, the forespore becomes fully suspended in the cytosol of the mother cell, and both stains are excluded. The yellow arrows indicate cells that are undergoing asymmetric division (indicated by a flat polar septum); the orange arrows indicate cells that are in the process of engulfment (indicated by a curved polar septum); and blue arrows indicate cells that have completed engulfment (indicated by bright membrane staining fully surrounding the forespore). Scale bar, 10 μ m. The bar plots indicate the percentage of sporulating cells at different stages of spore assembly in both WT and $\Delta camA$ cells. Data are mean \pm s.d. of $n = 3$ independent replicates. Images above bar plots show examples of each spore assembly stage; scale bar, 2 μ m. A total of 3,747 (WT, 9 h), 3,879 ($\Delta camA$, 9 h), 4,960 (WT, 11 h) and 4,650 ($\Delta camA$, 11 h) cells were screened. ***** $P \leq 0.05$; the statistical analysis was performed using two-tailed unpaired Student's *t*-tests.

Genomes pertaining to the same ST tend to have more similar sets of methylation motifs relative to those from different STs. Those genomes belonging to ST-2, ST-8, ST-21 and ST-110 showed the highest motif diversities. One 6mA motif, CAAAAA, was present across all of the genomes; we therefore hypothesized that 6mA

methylation events at this motif and its corresponding MTase have an important and conserved function in *C. difficile*.

A DNA methyltransferase and its target motif are ubiquitous in *C. difficile*. Motivated by the consistent presence of the methylation

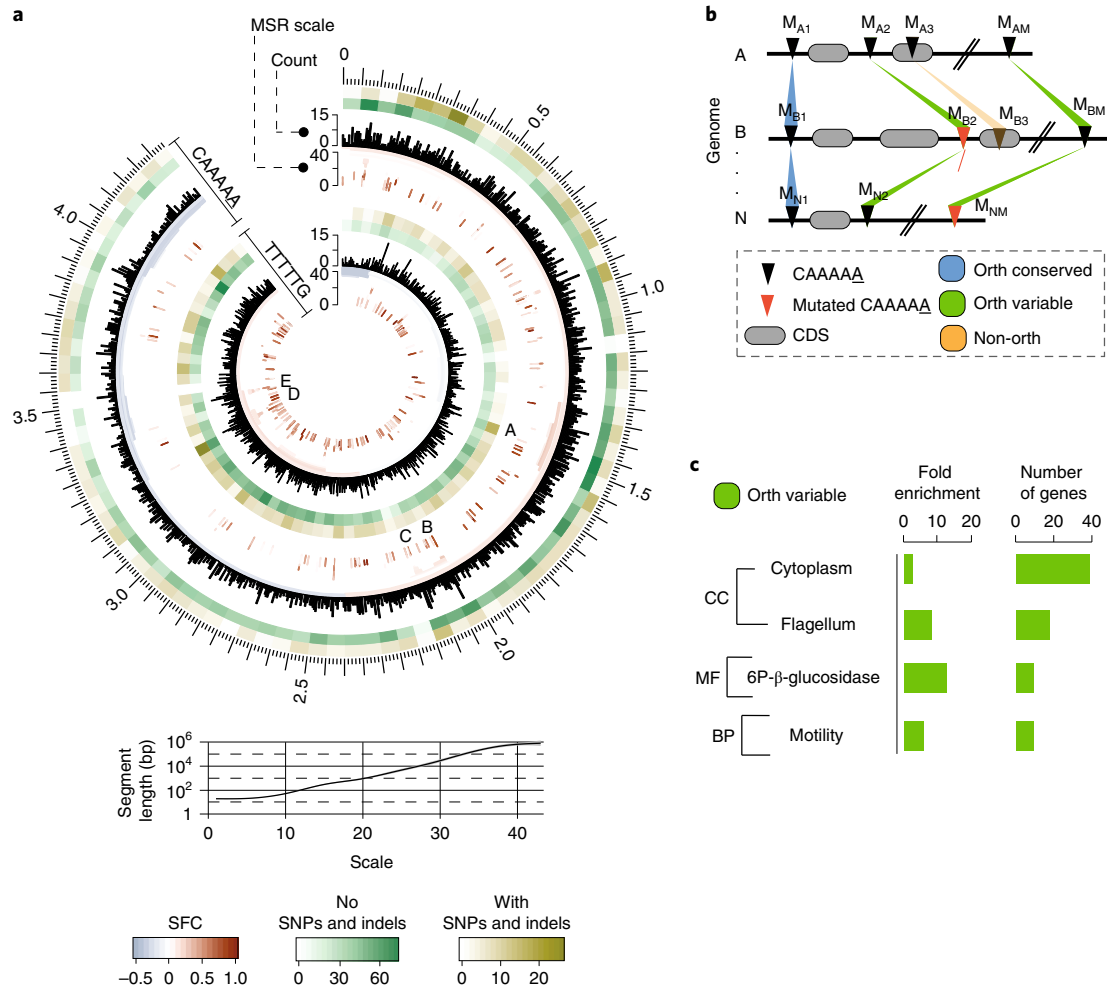


Fig. 3 | Abundance, distribution and conservation of CAAAAA motif sites. a, The distribution of CAAAAA sites in both strands of the reference *C. difficile* 630 genome and corresponding genomic signal obtained by MSR. In brief, MSR uses wavelet transformation to examine the chromosome at a succession of increasing length scales by testing for enrichment or depletion of a given genomic signal. Whereas scale values of less than 10 are typically associated with regions of less than 100 bp, genomic regions enriched for CAAAAA sites at scale values of more than 20 correspond to segments larger than 1 kb (that is, gene and operon scale). The letters (A–E) represent regions with a particularly high abundance of CAAAAA motif sites, including genes related to sporulation (such as *spo0A*, *spolIIAA–AH*, *spolVB* and *sigK*), membrane transport (PTS and ABC-type systems), transcriptional regulation (such as *iscR* and *fur*) and coding for multiple cell wall proteins (Supplementary Table 6d). The relationship between MSR scale and segment length is also shown. The significant fold change (SFC) corresponds to the fold change (\log_2 -transformed ratio) between observed and randomly expected overlap, statistically significant at $P=10^{-6}$ on the basis of the Z-test. Heat-map layers correspond to the number of orthologous conserved (no SNPs or indels; green) and orthologous variable (with SNPs and/or indels) CAAAAA motif positions. **b**, Whole-genome alignment of 37 *C. difficile* genomes (36 isolates and *C. difficile* 630 as reference) was performed using Mauve. We defined an orthologous (Orth) occurrence of the CAAAAA motif (black triangles) if an exact match to the motif was present in each of the 37 genomes (conserved; blue) or if at least one motif (and a maximum of $n-1$, being n the number of genomes) contained positional polymorphisms (maximum of two SNPs or indels per motif; variable; green). Non-orthologous CAAAAA positions are indicated as orange regions. The results are shown in **a** as heat maps. The numbering scheme is based on mapping location. **c**, DAVID enrichment analysis of genes containing intragenic and regulatory (100 bp upstream the start codon) orthologous variable CAAAAA motif sites. Genes that were found to over-represent orthologous variable CAAAAA positions include cytoplasm-related genes (such as *pheA*, *fdhD*, *ogt1* and *spoIVA*) and motility-related genes (such as *fliZ*, *fliN*, *fliM* and *flgL*). CC, cellular component; MF, molecular function; BP, biological process. Single categories were considered to be significantly enriched at $P < 0.05$ and correspond to 73 out of a total of 617 genes analysed; FDR-corrected P values were calculated using one-tailed Fisher's exact tests.

motif CAAAAA across all of the *C. difficile* isolates, we proceeded to examine the encoded MTases. We identified a total of 139 MTase genes (an average of 3.9 per genome; Fig. 1a, Supplementary Table 2b) that represent all of the four major types²⁰ and appear either in a solitary context or within restriction–modification (R–M) systems (Fig. 1c–e, Supplementary Table 2b–d). We also found multiple additional defence systems (such as abortive infection systems, CRISPR–Cas and toxin–antitoxin) and performed an integrative analysis with R–M systems in relation to host defence and gene

flux (Extended Data Figs. 1 and 2, Supplementary Table 3a–g), such as that resulting from phage infection (Extended Data Fig. 3, Supplementary Notes).

Consistent with the presence of a highly conserved CAAAAA motif, we identified a type II 6mA solitary DNA MTase (577 amino acids) that is present across isolates (Fig. 1f, Supplementary Table 2b, Supplementary Notes) and is responsible for methylation of the CAAAAA motif. This MTase is encoded by *CD2758* in the reference strain *C. difficile* 630 (refs. ^{2,21}). Here we named CD2758 CamA

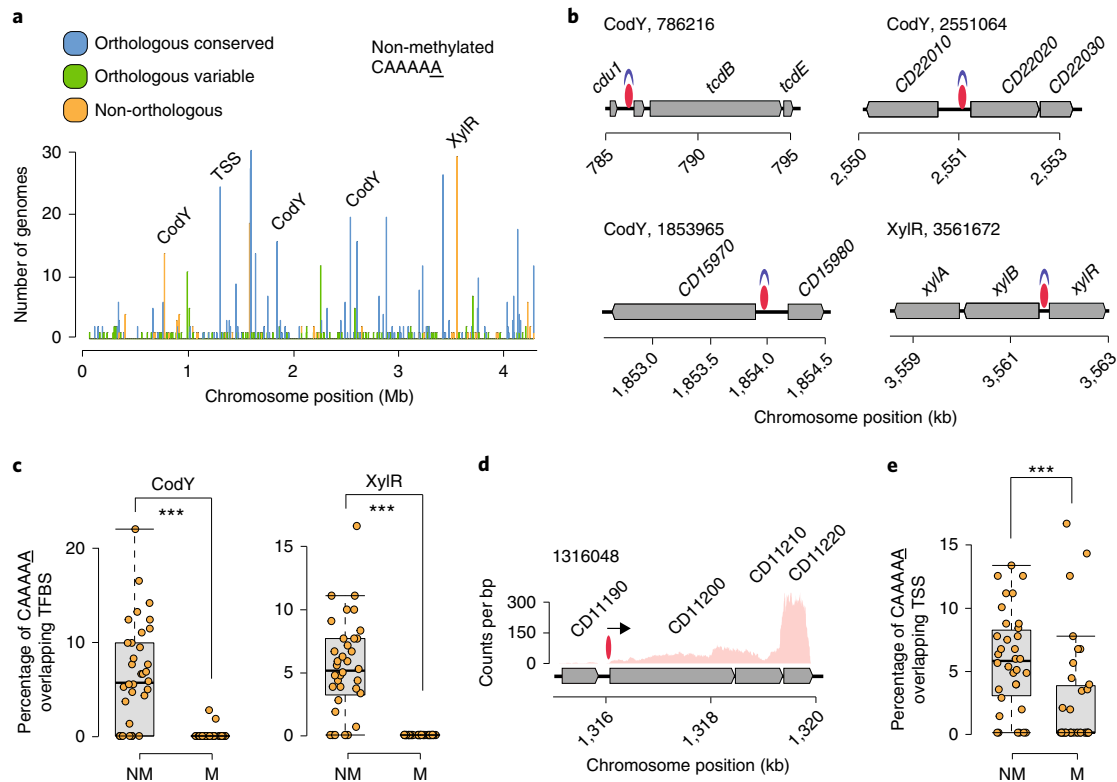


Fig. 4 | The distribution of non-methylated CAAAAA motif sites and their overlap with TFBSs and TSSs. **a**, The number of *C. difficile* isolates in which non-methylated CAAAAA motif sites were detected at a given chromosome position (coordinates are relative to the reference genome of *C. difficile* 630). Peak colours correspond to orthologous (conserved and variant) and non-orthologous CAAAAA positions. We found that some of the major peaks of non-methylated CAAAAA positions overlapped with TFBSs (such as CodY and XylR) and TSSs. **b**, Genetic regions for which overlap was observed between highly conserved non-methylated CAAAAA motif sites (red ovals) and TFs (CodY and XylR, shown in blue). Other examples of conserved non-methylated CAAAAA motif sites are provided in Extended Data Fig. 7b. **c**, The percentage of CAAAAA motif sites (non-methylated (NM) and methylated (M)) that overlap CodY and XylR binding sites for each of the $n = 36$ *C. difficile* isolates. $***P < 10^{-3}$. **d**, An example of a chromosomal region in which non-methylated CAAAAA motifs overlap a TSS (arrow). **e**, The percentage of CAAAAA motifs (non-methylated and methylated) that overlap TSSs for each of the $n = 36$ *C. difficile* isolates. For the box plots, the centre line indicates the median value, the boxes indicate the 25th and 75th quartiles, and the whiskers indicate 1.5x the interquartile range. For **c** and **e**, the statistical analysis was performed using one-sided Mann-Whitney-Wilcoxon rank-sum tests with continuity correction.

(*C. difficile* adenine methyltransferase A). Its ubiquity was not restricted to the 36 isolates, as we were able to retrieve orthologues in a list of around 300 global *C. difficile* isolates from GenBank (Supplementary Table 4). REBASE also showed functional orthologues of *camA* in only a very small number of other *Clostridiales* and *Fusobacteriales* (Extended Data Fig. 4), suggesting that this MTase is fairly unique to *C. difficile*.

Inactivation of *camA* reduces sporulation levels in vitro. Given the critical role of sporulation in the persistence and dissemination of *C. difficile* in humans and hospital settings²², we decided to test whether *camA* inactivation could reduce spore purification efficiencies in the 630 strain as previously suggested for its homologue in the 027 isolate R20291 (ref. ²³). We constructed an in-frame deletion in this gene ($\Delta camA$) and complemented it with either wild-type (WT) *camA* ($\Delta camA-C$) or a variant encoding a catalytic site mutation (N165A) of the MTase ($\Delta camA$ N165A) (Extended Data Fig. 5a, Supplementary Table 5a,b). We observed that spore purification efficiencies decreased by around 50% in the mutant relative to the WT (Fig. 2a). Complementation of $\Delta camA$ with the WT, but not the catalytic mutant, restored spore purification efficiencies to values that are similar to those observed in the WT cells (Fig. 2a, Supplementary Table 5c). No differences in growth were observed between the WT and mutant strains (Extended Data Fig. 5b). Thus,

this complementation experiment suggests that the loss of methylation events by CamA, rather than the loss of non-catalytic roles of this protein, leads to the decrease in spore yield.

The diminished efficiencies in spore purification observed in the $\Delta camA$ mutants could be due to a reduction in the number of cells that induce sporulation or defects in spore assembly²⁴. Visual inspection of samples before and after spore purification on a density gradient revealed qualitatively lower levels of mature phase-bright spores (Extended Data Fig. 5c). As purified WT and $\Delta camA$ spores had similar levels of chloroform resistance and germinated with similar efficiency (Extended Data Fig. 5d,e), the reduced spore purification efficiencies of the MTase mutants probably reflect a defect in sporulation initiation rather than the sporulation process itself. Accordingly, we observed that fewer $\Delta camA$ cells were sporulating in phase-contrast microscopy analyses relative to WT cells (Fig. 2b).

To gain insights into the sporulation stage that is affected by the loss of CamA, we quantified the number of sporulating cells at different stages of spore assembly (Fig. 2c). Although similar numbers of WT and $\Delta camA$ cells were observed at asymmetric division (the first morphological stage of sporulation) 9 h after sporulation induction, 50% fewer $\Delta camA$ cells had initiated engulfment. Furthermore, around twofold more $\Delta camA$ cells were at asymmetric division relative to WT cells 11 h after sporulation induction,

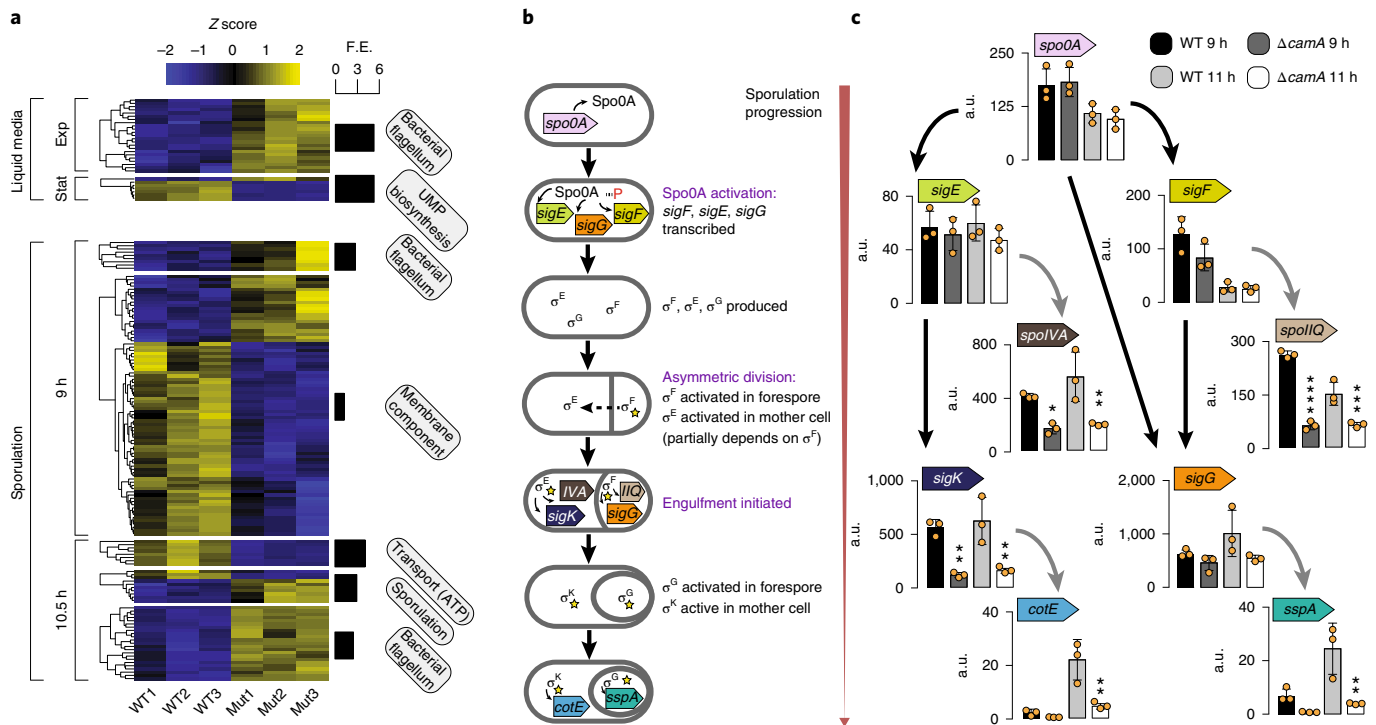


Fig. 5 | Gene-expression analysis. **a**, A heat map of 161 genes from three replicates of *C. difficile* 630 compared with an equal number of replicates of *C. difficile* 630 $\Delta camA$ that are enriched for the Gene Ontology (GO) terms shown in the boxes and described in Supplementary Table 8c. The Z score reflects the degree of downregulation (Z score < 0) or upregulation (Z score > 0), computed by subtracting the mean of the log-transformed expression values and dividing by the s.d. for each gene over all of the samples scored. FE, fold enrichment; Exp, exponential growth stage; Stat, stationary growth stage. **b**, Schematic of the sequence of sporulation sigma factor gene transcription and protein activation coupled to morphological changes during sporulation. Activated SpoOA induces the expression of genes encoding σ^F , σ^E and σ^G as well as factors required for asymmetric division and the post-translational activation of the early stage sporulation sigma factors, σ^F and σ^E . σ^F is the first sporulation-specific sigma factor to be fully activated and it becomes active in the forespore only after asymmetric division is completed¹⁰⁹. Activated σ^F subsequently induces the transcription of genes of which the products mediate σ^G activation in the forespore and partially mediate σ^E activation in the mother cell³⁵. Activated σ^E induces the transcription of *sigK*³³ and factors required for the excision of a prophage-like element from the *sigK* gene¹¹⁰. *C. difficile* sporulation is therefore controlled by a transcriptional hierarchy that is coupled to morphological events such that downstream sigma factors (σ^G and σ^K) depend on the activation of upstream sigma factors (σ^F and σ^E). **c**, A comparison of relative transcript levels in WT and $\Delta camA$ cells as determined by RT-qPCR for sporulation sigma factor genes and representative genes in the regulons of sporulation-specific sigma factors at 9 h and 11 h after sporulation induction (a separate set of $n = 3$ RNA sample replicates was used). Note that the primers for *sigK* amplify a region before the *sigK* excision site¹¹⁰. Data are mean \pm s.d. * $P < 0.05$, ** $P < 10^{-2}$, *** $P < 10^{-3}$, **** $P < 10^{-4}$; statistical significance was determined using one-way ANOVA and Tukey's test for multiple comparisons.

whereas 50% fewer $\Delta camA$ cells had completed engulfment compared with the WT cells. As similar numbers of sporulating cells were observed between WT and $\Delta camA$ at 11 h, the sporulation defect of $\Delta camA$ cells seems to arise because fewer cells progress beyond asymmetric division, rather than due to a defect in sporulation induction.

To confirm that the loss of CamA leads to a decrease in the number of cells that produce functional spores, we compared the ability of $\Delta camA$ to form heat-resistant spores that are capable of germinating and outgrowing using a heat-resistance assay²⁵. The $\Delta camA$ mutant and the catalytic-mutant complementation strain produced approximately 50% fewer heat-resistant spores than the WT and WT-complementation strains (Extended Data Fig. 5e). Taken together, these findings suggest that CAAAAA methylation enhances sporulation in vitro. This functional difference prompted us to perform a comprehensive methylome and transcriptome analysis of WT and $\Delta camA$ strains.

Comparative analysis of CAAAAA sites across *C. difficile* genomes. The *C. difficile* genome has an average of 7,721 CAAAAA motif sites (Supplementary Table 6a). Adjusting for the *k*-mer frequency of the AT-rich *C. difficile* genome (70.9%) using Markov

models²⁶, CAAAAA motif sites are significantly under-represented in intragenic regions (Extended Data Fig. 6a, Supplementary Table 6a,b). To evaluate whether specific chromosomal regions are enriched or depleted for this motif, we used a multiscale signal representation (MSR) approach²⁷. We observed strong enrichment for CAAAAA sites within genes related to sporulation, membrane transport, transcriptional regulation and coding for multiple cell wall proteins (Fig. 3a, Supplementary Table 6c,d).

To further characterize CAAAAA motif sites, we categorized them on the basis of their positional conservation across genomes. We performed whole-genome alignment of the isolates and classified each motif position in the alignment as follows: (1) conserved orthologous (devoid of single-nucleotide polymorphisms (SNPs) or indels); (2) variable orthologous (in which at least one genome contains a SNP or indel); and (3) non-orthologous (Fig. 3b, Supplementary Data 1). We found a total of 5,828 conserved orthologous motif positions, 1,050 variable orthologous positions and an average of 843 non-orthologous positions per genome (Supplementary Table 6e). Among orthologous positions, the variable positions contribute to variations at CAAAAA sites across genomes with subsequent methylation abrogation (Supplementary Table 6f). Such across-genome variation seems to be at least partially

fuelled by events of homologous recombination (Extended Data Fig. 6b–f, Supplementary Table 6g). Finally, DAVID gene enrichment analysis²⁸ found that cytoplasm- and motility-related genes over-represent orthologous variable CAAAAA positions (Fig. 3c). The very large number and dispersion of conserved orthologous positions precluded a similar functional analysis. Collectively, genome-wide distribution and across-genome comparative analyses suggest that CAAAAA sites are enriched in regions that harbour genes related to sporulation and colonization and that orthologous variable CAAAAA positions are enriched in regions that harbour cytoplasm- and motility-related genes.

Non-methylated CAAAAA motif sites are enriched in regulatory elements. The on/off switch of DNA methylation in a bacterial cell can contribute to epigenetic regulation as a result of competitive binding between DNA MTases and other DNA-binding proteins (such as transcription factors (TFs)) as previously described^{12,29–31}. Previous bacterial methylome studies that analysed one or few genomes had insufficient statistical power to perform a systematic interrogation of non-methylated motifs sites¹⁶. Building on our collection of *C. difficile* methylomes, we performed a systematic detection and analysis of non-methylated CAAAAA sites and found an average of 21.5 of such sites per genome (Extended Data Fig. 7a, Supplementary Table 7a). We found that non-methylated motif sites were dispersed throughout the full length of the *C. difficile* chromosome, yet were over-represented in orthologous variable and non-orthologous CAAAAA positions (observed/expected, 1.51 and 1.49, respectively) and under-represented in orthologous conserved CAAAAA positions (observed/expected, 0.84; all $P < 10^{-4}$; χ^2 test). This is consistent with the idea that variable positions are more likely to be non-methylated to provide breadth of expression variation. Most of the non-methylated positions (85.4% of 245) failed to conserve such status in more than three genomes at orthologous positions, whereas a small percentage of positions (5.5%) remained non-methylated in at least one-third of the isolates, suggesting that competitive protein binding is expected to be more active in certain genomic regions (Fig. 4a).

The non-methylated CAAAAA positions detected across the *C. difficile* genomes enabled us to perform a systematic search for evidence of overlap between the CAAAAA motifs and TF binding sites (TFBSs) and transcription start sites (TSSs). First, we found overlaps between prominent peaks of non-methylated CAAAAA positions and the TFBSs of CodY and XylR (Fig. 4a,b, Extended Data Fig. 7b, Supplementary Table 7b,c). Performing the analysis at the genome level, both CodY and XylR binding sites showed

significant enrichment ($P < 10^{-3}$, Mann–Whitney–Wilcoxon test) for non-methylated CAAAAA (Fig. 4c, Extended Data Fig. 7c). Second, using TSSs reconstructed from RNA-sequencing (RNA-seq) data coverage, we found a similar genome-level enrichment for non-methylated CAAAAA sites (Fig. 4d,e, Extended Data Fig. 7d,e, Supplementary Table 7d; $P < 10^{-3}$, Mann–Whitney–Wilcoxon test). Thus, these results demonstrate the occurrence of an on/off epigenetic switch of CAAAAA sites that preferentially overlaps with putative TFBSs and TSSs.

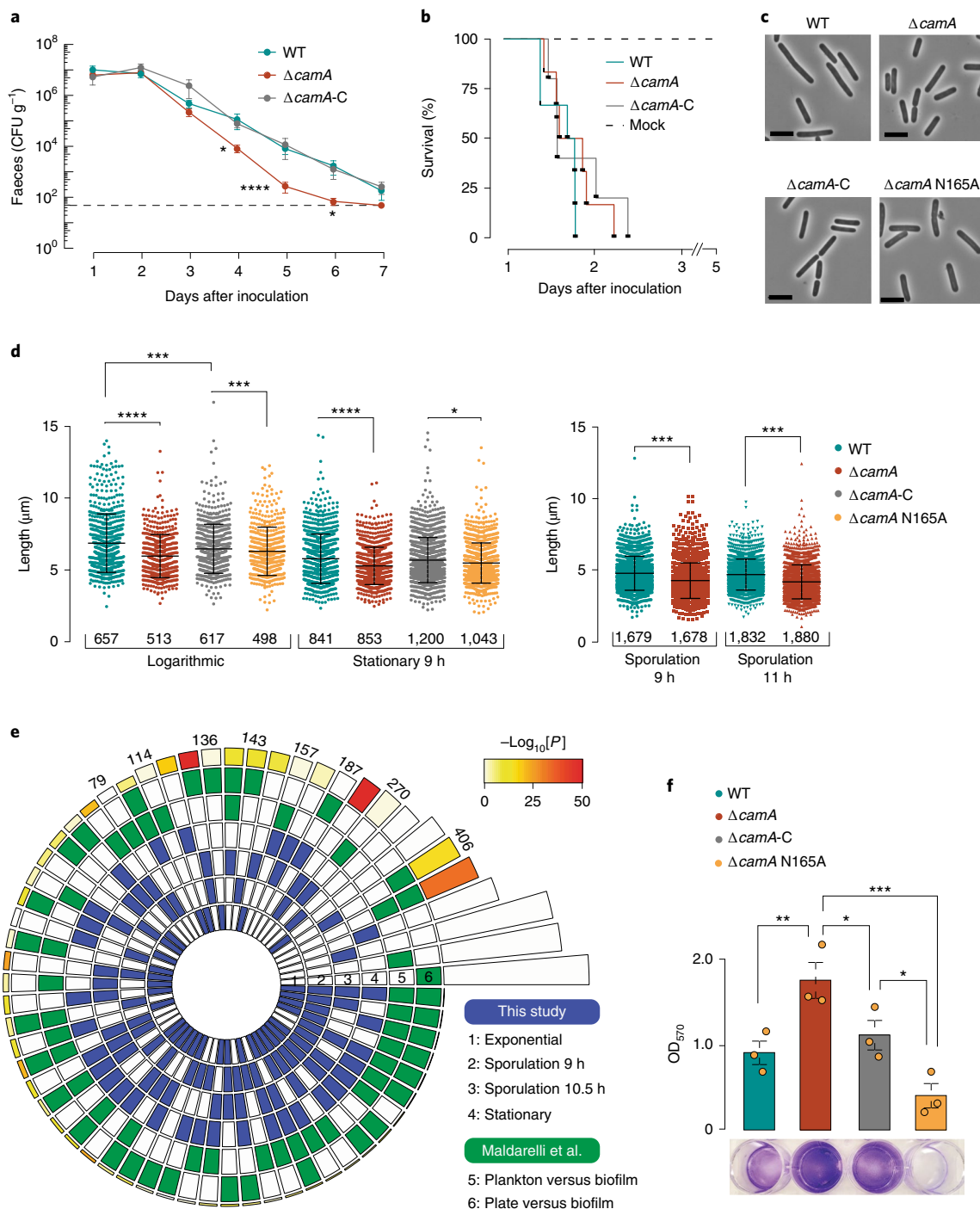
Loss of CAAAAA methylation impacts the transcription of multiple gene categories, including sporulation. To study the functional importance of methylation at CAAAAA sites, we used RNA-seq to compare the transcriptome of WT *C. difficile* 630 with that of $\Delta camA$ both in liquid medium (exponential and stationary growth stage) and after sporulation induction (9 h and 10.5 h; Extended Data Fig. 8, Supplementary Table 8a, Supplementary Data 2). Of the 3,896 genes annotated in *C. difficile* 630, 36–361 (0.9–9.3%, depending on the time point) were differentially expressed (DE) at a 5% false discovery rate (FDR) and $|\log_2[FC]| > 1$ (twofold change in gene expression; Fig. 5a, Supplementary Table 8b–d). DE genes in $\Delta camA$ cells relative to the WT showed significant enrichment in CAAAAA motif sites compared with non-DE genes ($P < 10^{-2}$, Mann–Whitney–Wilcoxon test) in broth culture, and a qualitatively similar trend was also observed during sporulation (Extended Data Fig. 9a). Consistent with our finding that the loss of CamA reduces spore formation, the transcriptome analyses revealed that 118 and 120 genes that have previously identified to be induced during sporulation^{32,33} were expressed at $\geq 50\%$ reduced levels in $\Delta camA$ cells relative to WT cells at 9 h and 10.5 h, respectively (Supplementary Table 8b).

The transcriptional program that mediates sporulation in *C. difficile* is controlled by a master transcriptional activator, Spo0A, and four sporulation-specific sigma factors, σ^F , σ^E , σ^G and σ^K . These factors activate distinct regulons that ultimately lead to the assembly of functional spores^{34,35} (Fig. 5b); the early acting sigma factors, σ^F and σ^E , are required for the activity of the later-acting sigma factors, σ^G and σ^K , respectively. A transcriptional hierarchy therefore governs sporulation in *C. difficile* with downstream factors depending on the activation of upstream sigma factors. As genes in the regulons of all four sporulation-specific sigma factors were underexpressed in $\Delta camA$ cells relative to the WT, whereas a relatively small subset of Spo0A regulon genes exhibited this pattern of regulation (Extended Data Fig. 9b, Supplementary Table 8e), loss of CamA probably affects early events during sporulation.

Fig. 6 | In vivo and additional functional impacts of the $\Delta camA$ mutation. **a**, Kinetics of infection in antibiotic-treated mice ($n = 12$) after treatment with a sub-lethal inoculation (10^5 spores) of WT *C. difficile* 630 Δerm , MTase mutant $\Delta camA$ and complement $\Delta camA$ -C. When inoculated with spontaneous erythromycin sensitive derivative (630 Δerm) strains, antibiotic-treated mice do not typically develop fulminant disease and instead serve as a model of intestinal colonization and persistence by *C. difficile*^{105,106,111}. The dotted line indicates the limit of detection. Data are mean \pm s.e.m.; \log_{10} -transformed data from each time point were analysed by ANOVA for each time point. **b**, Kaplan–Meier survival curves for clindamycin-treated golden Syrian hamsters ($n = 6$) infected with 10^3 spores of WT *C. difficile* 630 Δerm , $\Delta camA$ or complement $\Delta camA$ -C. **c**, Representative phase-contrast images ($n = 3$ independent biological replicates) of vegetative WT, $\Delta camA$, $\Delta camA$ -C and $\Delta camA$ N165A cells. Scale bars, 5 μ m. **d**, Comparison of cell length. Data are mean \pm s.d. of $n = 3$ independent biological replicates (the exact numbers of cells measured are indicated in the figure). The statistical analysis was performed using one-way ANOVA and Tukey's test for multiple comparisons. **e**, Significance of overlap between multiple datasets of DE genes. Comparisons were performed between DE genes called in this study for each time point (blue, $n = 1,537$) and those from Maldarelli et al.⁴⁰ (green, $n = 1,735$). The latter corresponds to *C. difficile* DE genes in conditions that favour biofilm formation compared with growth on a plate or planktonic form. Colour intensities of the outermost layer represent the P value significance of the intersections (3,896 genes used as background). The height of the corresponding bars is proportional to the number of common genes in the intersection (shown for pairwise comparisons across different studies). DE genes in the $\Delta camA$ mutant (sporulation phases) were found to have a significant overlap with DE genes in conditions that favour the production of biofilm; $P < 10^{-9}$; the statistical analysis was performed using one-tailed hypergeometric tests implemented in SuperExactTest, Bonferroni adjusted. **f**, Biofilm production, measured by crystal-violet staining absorbance at 570 nm. The differences in biofilm production between $\Delta camA$ and $\Delta camA$ N165A could be explained if the $\Delta camA$ N165A retained some DNA-binding ability and was able to alter the transcription of some genes even in the absence of methylation. Data are mean \pm s.d. of $n = 3$ independent biological replicates, with each strain assayed in quadruplicate in each experiment. * $P < 0.05$, ** $P < 10^{-2}$, *** $P < 10^{-3}$. The statistical analysis was performed using two-way ANOVA with Dunnett's post hoc test.

To identify the regulatory stage of sporulation that CamA-mediated DNA methylation specifically impacts, we used quantitative PCR with reverse transcription (RT-qPCR) to analyse the expression of genes encoding Spo0A, the sporulation-specific sigma factors^{32,36} and genes in their individual regulons^{32,36,37}. Consistent with our RNA-seq analyses, Spo0A regulon genes—*spo0A*, *sigF* and *sigE*^{32,37}—were expressed at similar levels between WT and Δ *camA* cells at both 9 h and 11 h, implying that the Δ *camA* mutant activates Spo0A at levels similar to the WT. By contrast, the σ^F and σ^E regulon genes, *spoIIQ* and *spoIVA*^{32,38}, respectively, were underexpressed in Δ *camA* compared with the WT cells (Fig. 5c). Reduced levels of SpoIIQ and SpoIVA were observed in Δ *camA* cells by western blot, confirming the transcriptional

analyses (Extended Data Fig. 9c). On the basis of the hierarchical organization of the sporulation regulatory cascade, σ^F activation seems to be the earliest sporulation stage that is affected by CamA. This conclusion is supported by our morphological analyses, because fewer Δ *camA* cells progress to engulfment (a process that requires both σ^F and σ^E activation³⁹) than WT cells (Fig. 2c), whereas similar numbers of Δ *camA* and WT cells initiate sporulation. Indeed, similar levels of Spo0A activation are observed in WT and Δ *camA* (Fig. 5c), and the small subset of Spo0A regulon genes that are underexpressed in Δ *camA* cells could be dually regulated by Spo0A and σ^F . For example, *spoIIR*³⁶—which encodes a signalling protein required for σ^E activation—is activated by both σ^F and Spo0A^{32,35}.



In vivo effects of the *camA* mutation. To test whether the sporulation defect of $\Delta camA$ impacts the infection or transmission of *C. difficile*, we analysed the effect of the $\Delta camA$ mutation in an established mouse model of infection. Groups of mice (6 males and 6 females) were inoculated by oral gavage with spores of the three genotypes: WT, $\Delta camA$ and $\Delta camA$ -C. No mortality was observed at the given doses of *C. difficile* spores, as expected. Faecal samples were collected every 24 h for 7 d. All three *C. difficile* strains reached comparable levels in faeces at days 1 and 2 after inoculation, indicating that they germinate and establish colonization with equal efficiency (Fig. 6a). As expected, colony-forming unit (CFU) levels decreased steadily from day 2 after inoculation to day 7. However, the $\Delta camA$ mutant showed CFU levels that were 10–100 times lower than those observed in the WT and complemented strains throughout this time frame. The level of bacteria declined to near the limit of detection in the faeces 6 d after inoculation for the MTase mutant, whereas the WT and complemented strains remained detectable at days 6 and 7.

To test whether the loss of CamA leads to defects in virulence, we compared *C. difficile* $\Delta camA$ and WT in a hamster model of infection. Clindamycin-treated golden Syrian hamsters are highly susceptible to the effects of the *C. difficile* toxins and, therefore, represent a model of acute disease. Groups of 6 hamsters were inoculated by oral gavage with spores of the WT, $\Delta camA$ and $\Delta camA$ -C strains. These strains of *C. difficile* elicited diarrhoeal symptoms and weight loss in hamsters, and we observed no difference in the survival times of hamsters after inoculation (Fig. 6b). This result is consistent with the observation that the WT, $\Delta camA$ and $\Delta camA$ -C strains exhibit no differences in toxin gene expression (Supplementary Table 8a) and produce comparable levels of TcdA in vitro (Extended Data Fig. 9d). Together, these data indicate that CAAAAA methylation by CamA does not influence toxin-mediated aspects of *C. difficile* pathogenesis but, instead, impacts the ability of *C. difficile* to persist within the host intestinal tract.

Additional functional effects of the *camA* mutation. Considering the high conservation of *camA* across *C. difficile* genomes, we examined whether some additional phenotypes could be effected by the inactivation of *camA*. While analysing images of sporulating *C. difficile*, we noticed that $\Delta camA$ -mutant cells appeared to be shorter on average than the WT. To test this possibility, we measured the lengths of WT and $\Delta camA$ cells during broth culture and sporulation and found that $\Delta camA$ cells were around 15% shorter than the WT cells (Fig. 6c,d) even though no difference in growth was observed (Extended Data Fig. 5b). Interestingly, genes that encode putative cell-wall remodelling enzymes were overexpressed in the $\Delta camA$ mutant cells relative to the WT during growth in broth culture (Extended Data Fig. 9e).

We next performed an overlap analysis between the list of DE genes from our RNA-seq data (WT versus $\Delta camA$ mutant; four different time points) and those from published studies focusing on the colonization and infection by this pathogen (Supplementary Table 8f). First, DE genes in the $\Delta camA$ mutant (sporulation phases) had a significant overlap with DE genes in conditions that favour the production of biofilm on a solid substrate⁴⁰ (Fig. 6e). Motivated by this significant overlap, we performed crystal-violet staining assays of the biomass of adherent biofilm, and consistently observed that the $\Delta camA$ mutant produced more biofilm than the WT cells (Fig. 6f). These results suggest that methylation inhibits the expression of genes that promote biofilm formation. Second, significant overlaps were found when comparing with genes that are DE during infection in different mice gut microbiome compositions⁴¹ (Extended Data Fig. 10a, Supplementary Table 8f). Finally, significant overlaps were found when comparing with DE genes obtained from mice gut isolates at increasing time points after infection⁴² (Extended Data Fig. 10b, Supplementary Table 8f). Collectively,

these integrative analyses provide further evidence that DNA methylation events by CamA may directly and/or indirectly affect the expression of multiple genes involved in the in vivo colonization and biofilm formation of *C. difficile* and inspire future studies to elucidate the mechanisms that underlie the functional roles of CAAAAA methylation in the pathogenicity of *C. difficile*.

Discussion

C. difficile is responsible for one of the most common hospital-acquired infections and is classified by the US Centers for Disease Control and Prevention as an urgent healthcare risk associated with substantial morbidity and mortality⁴³. As *C. difficile* infection is spread by bacterial spores that are found within faeces, extensive research has been devoted to better understand the genome of this important pathogen and its sporulation machinery. To address these common goals, we performed a comprehensive characterization of the DNA methylation landscape across a diverse collection of clinical isolates. During our analysis, we identified a 6mA MTase (*camA*) that is conserved across all of the isolates (and in another ~300 published *C. difficile* genomes) that share a common methylation motif (CAAAAA). Inactivation of the gene encoding this MTase resulted in a sporulation defect in vitro (Fig. 2). Infection studies using a mouse model indicate a role for CamA in the persistence of *C. difficile* in the intestinal tract. As enumeration of *C. difficile* recovered from faeces of the infected animals reflects the number of *C. difficile* spores in the gut, the reduced burden of $\Delta camA$ in mice might be due to the mutant's defect in sporulation (Fig. 6a), as the ability to form spores was previously shown to be important for persistence⁴⁴. The comparable virulence between $\Delta camA$ cells and the WT in the hamster model suggests that DNA methylation does not impact toxin-mediated disease. However, owing to the pleiotropic nature of the MTase it remains possible that multiple factors contribute to the more pronounced effect that is observed in the mouse model.

The highly conserved nature of *camA* and its flanking genes across *C. difficile* genomes suggests that additional phenotypes may be regulated by CamA beyond sporulation. Consistent with this, we found CAAAAA sites are over-represented in a set of regions enriched in genes with functions linked to sporulation, motility and membrane transport. Further supporting a broader regulatory network of CamA, the loss of CamA reduces cell length and results in a statistically significant overlap between the transcriptional signatures identified in our study (WT versus $\Delta camA$ mutant cells) and those of others observed during the in vivo colonization and biofilm formation (Fig. 6e, Extended Data Fig. 10a,b).

The fact that *camA* is a solitary MTase gene without a cognate restriction gene further supports the view that widespread methylation in bacteria has functional importance beyond the role that is attributed to R–M systems. Previously, the most extensively characterized 6mA MTase was Dam-targeting GATC in *Escherichia coli*. Dam has multiple important functions and is essential in some pathogens¹². However, because Dam is conserved in the large diversity of Enterobacteria, it was not considered to be promising drug target. By contrast, the uniqueness of *camA* in all of the *C. difficile* genomes and in just a few *Clostridiales* makes it a promising drug target that may inhibit *C. difficile* in a much more specific manner, which is particularly relevant because gut dysbiosis potentiates *C. difficile* infection^{45,46}. Furthermore, as this MTase does not seem to impact the general fitness of *C. difficile*²³, a drug that specifically targets it may be developed with a lower chance for resistance.

Considering the large number of genes that were DE in the $\Delta camA$ mutant, the functional impact of CAAAAA methylation is probably mediated by multiple genes that are either directly regulated by DNA methylation or indirectly regulated by a transcriptional cascade. Mechanistically, DNA methylation can either activate or repress a gene depending on other DNA-binding proteins that compete with DNA MTases^{7,8,12,47}; therefore, the

competition between TFs and MTases may form an epigenetic switch to turn a gene on and off.

With more than 2,200 bacterial methylomes published to date, it is becoming increasingly evident that epigenetic regulation of gene expression is highly prevalent across bacterial species. Despite the exciting prospects for studying epigenetic regulation, our ability to comprehensively analyse bacterial epigenomes is limited by a bottleneck in integratively characterizing methylation events, methylation motifs, transcriptomic data and functional genomic data. In this regard, this study provides a comprehensive comparative analysis of a large collection of a single bacterial species, as well as a detailed roadmap that can be used by the scientific community to leverage the current status quo of epigenetic analyses.

Methods

***C. difficile* isolates and culture.** We obtained 36 clonal *C. difficile* isolates from infected faecal samples using protocols that were developed in the ongoing Pathogen Surveillance Program at Mount Sinai Hospital (Supplementary Table 1). Furthermore, 9 fully sequenced and assembled *C. difficile* genomes were retrieved from GenBank RefSeq (<ftp://ftp.ncbi.nih.gov/genomes>, accessed November 2016; Supplementary Table 1). Raw sequencing data from global and UK collections comprising 291 *C. difficile* 027/BI/NAPI genomes were used³ (Supplementary Table 4). *C. difficile*-positive stool samples were frozen at -80°C before analysis. All of the stool samples underwent culture for *C. difficile* using an ethanol-shock culture method⁴⁸. In brief, approximately 80 mg of solid stool (50 μl liquid stool samples) was added to 0.5 ml of 70% ethanol wash and the sample was mixed using a vortex and incubated at room temperature for 20 min. A loopful was then cultured onto *C. difficile* selective agar (CDSA, Becton Dickinson) and the plates were incubated anaerobically at 37°C for up to 72 h. A single colony was subcultured onto a trypticase soy agar plate with 5% defibrinated sheep blood (TSA II, Becton Dickinson) and incubated anaerobically at 37°C for 48 h. Colonies that had the *C. difficile* odour and showed fluorescence under illumination with ultraviolet light were then obtained and confirmed by matrix-assisted laser desorption/ionization on a Bruker biotyper. For long-term storage, individual colonies were emulsified in tryptic soy broth containing 15% glycerol and stored at -80°C .

SMRT-seq. Primers were annealed to size-selected (>8 kb) SMRTbell templates with the full-length libraries (80°C for 2 min 30 s followed by decreasing the temperature to 25°C by $3^{\circ}\text{C min}^{-1}$). The polymerase-*template* complex was then bound to P6 enzyme using a ratio of 10:1 polymerase to SMRTbell at 0.5 nM for 4 h at 30°C and then held at 4°C until ready for magnetic-bead loading, before sequencing. The magnetic-bead-loading step was performed at 4°C for 60 min according to the manufacturer's guidelines. The magnetic-bead-loaded polymerase-bound SMRTbell libraries were placed onto the RSII machine at a sequencing concentration of 125–175 pM and configured for a 240 min continuous sequencing run.

De novo genome assembly and motif discovery. The RS_HGAP3 protocol was used for de novo genome assembly, followed by the use of custom scripts for genome finishing and annotation (<https://github.com/powerpak/pathogendb-pipeline>). RS_Modification_and_Motif_Analysis.1 was used for de novo methylation motif discovery. A custom script was used to examine each motif to ensure its reliable methylation states. In brief, variations in a putative motif were examined by comparing the distribution of the interpulse duration ratio of each variation with non-methylated motifs.

The presence and conservation of *camA* in *C. difficile* isolates. To investigate the pervasive role and conservation of *camA*, we searched for its presence in a global and UK collection of *C. difficile* 027/BI/NAPI ($n = 291$)³ genomes (Supplementary Table 4). For this, SRA Illumina reads were converted to FASTQ files using fastq-dump v.2.8.0 and subsequently mapped to the *C. difficile* 630 reference genome using Bowtie2 v.2.2.9 (ref. ⁴⁹) in paired-end mode. The resulting SAM files were converted to BAM format (unmapped reads and PCR duplicates were removed) and sorted using SAMTOOLS v.1.9 (ref. ⁵⁰). To assess coverage, sequence depths were computed using the genomeCov function of BEDTOOLS v.2.26.0 (ref. ⁵¹) for each strand separately. Variant sites were called from the aligned reads using the mpileup and bcftools tools in SAMTOOLS.

Identification of defence systems. Identification of R–M systems was performed as previously described⁵². In brief, curated reference protein sequences of types I, II, IIC and III R–M systems, and type IV REases were downloaded from the dataset 'gold standards' of REBASE⁵³ (accessed November 2016). All-against-all searches were performed for REase and MTase standard protein sequences retrieved from REBASE using BLASTP v.2.5.0+ (default settings, $e < 10^{-7}$). The resulting *e* values were log-transformed and used for clustering into protein families by Markov Clustering v.14-137 (ref. ⁵⁴). Each protein family was aligned with MAFFT v.7.305b

(ref. ⁵⁵) using the E-INS-i option, 1,000 cycles of iterative refinement and offset 0. The alignments were visualized in SEAVIEW v.4.6.1 (ref. ⁵⁶) and manually trimmed to remove poorly aligned regions at the extremities. Hidden Markov model (HMM) profiles were then built from each multiple sequence alignment (available at <https://github.com/pedrocas81>) using the hmmbuild program from the HMMER v.3.0 suite⁵⁷ (default parameters). Types I, II and III R–M systems were identified by searching genes encoding the MTase and REase components at less than five genes apart. CRISPR repeats were identified using the CRISPR Recognition Tool (CRT) v.1.2 (ref. ⁵⁸) with default parameters. For the CRISPR spacer homology search, hits with at least 80% identity were considered to be positive. For *cas* gene identification, we obtained Cas protein family HMMs from the TIGRFAM database⁵⁹ v.15.0 and PFAM families annotated as Cas families (https://ftp.ncbi.nih.gov/pub/wolf/_suppl/CRISPRclass/crisprPro.html). In total we collected 129 known Cas protein families (98 TIGRFAMS and 31 PFAMs), which were used for similarity searching. Genes pertaining to abortive infection systems were searched with the PFAM profiles PF07751, PF08843 and PF14253 (accessed January 2018). Bacteriophage Exclusion (BREX) systems were searched using PFAM profiles for the core genes *pglZ* (PF08665) and *brxC* or *pglY* (PF10923) and specific PFAM profiles for each BREX type as indicated previously⁶⁰. DISARM systems were identified using the PFAM signature domains (PF09369, PF00271 and PF13091) that belong to the core gene triplet characteristic of this system⁶¹. To search for prokaryotic Argonaute (pAgo) genes, we built a dedicated HMM profile on the basis of a list of 90 Ago-PIWI proteins⁶². Searches for the ensemble of newly found antiphage systems were performed using a previously published list of PFAM profiles⁶³. Type II toxin–antitoxin systems were detected using the TAFinder tool⁶⁴ with default parameters. Matches of CRISPR spacers were performed against well-known *C. difficile* phages as follows: five siphophages (ϕCD111 (NC_028905.1), ϕCD146 (NC_028958.1), $\phi\text{CD38-2}$ (NC_015568.1), ϕCD6356 (NC_015262.1) and ϕCD211 (NC_029048.2)); five small-tail myophages (ϕMMP04 (NC_019422.1), ϕCD506 (NC_028838.1), ϕCDHM11 (NC_029001.1), $\phi\text{CD481-1}$ (NC_028951.1) and ϕCDHM13 (NC_029116.1)); five medium-tail myophages (ϕMMP03 (NC_028959.1), ϕCDMH1 (NC_024144.1), ϕC2 (NC_009231.1), ϕCD119 (NC_007917.1) and ϕCDHM19 (NC_028996.1)); and four long-tail myophages (ϕCD27 (NC_011398.1), ϕMMP02 (NC_019421.1), ϕCD505 (NC_028764.1) and ϕMMP01 (NC_028883.1)).

Identification and classification of prophages, conjugative/mobilizable elements and integrons. Prophages were detected with Phage Finder v.2.1 (ref. ⁶⁵) using strict mode and PHASTER⁶⁶ using the default settings. We took the common hits obtained by both programs, as well as those very few cases (~10% of the hit list) that corresponded to complete prophages predicted by just one of the programs. All elements that were either smaller than 18 kb or lacking matches to core phage proteins (such as terminase, capsid, head and tail proteins) were removed. Integrons were searched using IntegronFinder⁶⁷ with the default settings. The identification of genes encoding the functions related to conjugation in ICES was performed as previously described⁶⁸. In brief, an element was considered to be conjugative when it contained the following components of the conjugative system: a VirB4/TraU ATPase, a relaxase, a coupling ATPase (T4CP) and a minimum number of mating pair formation (MPF) type-specific genes—two for types MPF_{FA} and MPF_{FATAP}, or three for the others (types F, T and G). IMEs were identified by the fact that they encode relaxases but lack a complete conjugative transfer system, which is encoded in *trans* by another mobile element. Delimitation of ICES and IMEs was performed considering flanking core genes as upper bounds for their extremities.

Phylogenetic analyses. The reference phylogenetic tree of *C. difficile* was built from the concatenated alignment of protein families of the core-genome using MUSCLE⁶⁹ v.3.8.31 (default parameters). As the DNA sequences provide more phylogenetic signal than protein sequences at this evolutionary distance, we back-translated the alignments to DNA. Poorly aligned regions were removed using BMGE⁷⁰ v.1.12. The tree was computed using RAXML⁷¹ v.8.0.0 under the GTR model and a gamma correction (GAMMA) for variable evolutionary rates. We performed 100 bootstraps on the concatenated alignment to assess the robustness of the topology of the tree.

Identification of the core and pan-genome. The *C. difficile* core genome was built using a previously published methodology⁷². In brief, a preliminary list of orthologues was identified as reciprocal best hits using end-gap-free global alignment between the proteome of a pivot (*C. difficile* 630) and each of the other strain's proteomes. Hits with less than 80% similarity in amino-acid sequence or more than 20% difference in protein length were discarded. This list of orthologues was then refined for every pairwise comparison using information on the conservation of the gene neighbourhood. Positional orthologues were defined as bi-directional best hits adjacent to at least four other pairs of bi-directional best hits within a neighbourhood of ten genes (five upstream and five downstream). The core genome of each clade was defined as the intersection of pairwise lists of positional orthologues. The pan-genome was built using the complete gene repertoire of *C. difficile*. We determined a preliminary list of putative homologous proteins between pairs of genomes by searching for sequence similarity between

each pair of proteins using BLASTP (default parameters). We then used the e values ($<10^{-4}$) of the BLASTP output to cluster these proteins using SILIX⁷³ v.1.2.11. We set the parameters of SILIX such that two proteins were clustered in the same family if the alignment had at least 80% identity and covered more than 80% of the smallest protein (options $-1.0.8$ and $-r.0.8$). Core- and pan-genome accumulation curves were built using a dedicated R script. Regression analysis for the pan-genome was performed as described previously⁷⁴ by the Heap's power law $n = k \cdot N^{\gamma}$ where n is the pan-genome family size, N is the number of genomes and k, γ ($\alpha = 1 - \delta$) are specific fitting constants. For $\alpha > 1$ ($\delta < 0$) the pan-genome is considered to be closed, that is, sampling more genomes will not affect its size. For $\alpha < 1$ ($0 < \delta < 1$) the pan-genome remains open and the addition of more genomes will increase its size.

Inference of homologous recombination. We inferred homologous recombination on the multiple alignments of the core-genome of *C. difficile* (ordered locally collinear blocks (LCBs) obtained by progressiveMauve were used) using ClonalFrameML⁷⁵ v.10.7.5 and Geneconv⁷⁶ v.1.81a. The first used a predefined tree (that is, the species tree), default priors $R/\theta = 10^{-1}$ (ratio of recombination and mutation rates), $1/\delta = 10^{-3}$ (inverse of the mean length of recombination events) and $\nu = 10^{-1}$ (average distance between events), and 100 pseudo-bootstrap replicates, as previously suggested⁷⁵. Mean patristic branch lengths were computed with the R package ape⁷⁷ v.3.3, and transition/transversion ratios were computed using the R package PopGenome⁷⁸ v.2.1.6. The priors estimated by this mode were used as initialization values to rerun ClonalFrameML under the 'per-branch model' mode with a branch dispersion parameter of 0.1. The relative effect of recombination to mutation (r/m) was calculated as $r/m = R/\theta \times \delta \times \nu$. Geneconv was used with options /w123 to initialize the program's internal random number generator and -Skip_indels, which ignores all of the sites with missing data.

Reconstruction of the evolution of gene repertoires. We assessed the dynamics of gene family repertoires using Count⁷⁹ (downloaded in January 2018). This program uses birth-death models to identify the rates of gene deletion, duplication and loss in each branch of a phylogenetic tree. We used presence/absence pan-genome matrix and the phylogenetic birth-and-death model of Count to evaluate the most likely scenario for the evolution of a given gene family on the clade's tree. Rates were computed using default parameters, assuming a Poisson distribution for the family size at the tree root and uniform duplication rates. We computed 100 rounds of rate optimization with a convergence threshold of 10^{-3} . After optimization of the branch-specific parameters of the model, we performed ancestral reconstructions by computing the branch-specific posterior probabilities of evolutionary events, and inferred the gains in the terminal branches of the tree. The posterior probability matrix was converted into a binary matrix of presence/absence of horizontal gene transfer genes using a threshold probability of gain higher than 0.2 at the terminal branches. To control for the effects of the choices made in the definition of our model, we computed the gain/loss scenarios using the Wagner parsimony (same parameters, relative penalty of gain with respect to loss of 1). The horizontal gene-transfer events inferred by maximum likelihood and those obtained under Wagner's parsimony were highly correlated (Spearman's $\rho = 0.96$, $P < 10^{-4}$).

Strain construction and growth conditions. The 630 Δ erm Δ pyrE parental strain was used for pyrE-based allele-coupled exchange (ACE⁸⁰). A list of *C. difficile* and *E. coli* strains are provided in Supplementary Table 5a. *C. difficile* strains were grown from frozen stocks on brain-heart-infusion-supplemented (BHIS)⁸¹ medium plates supplemented with taurocholate (TA, 0.1% w/v; 1.9 mM), kanamycin (50 μ g ml⁻¹) and cefoxitin (8 μ g ml⁻¹) as needed. For ACE, *C. difficile* defined medium (CDDM)⁸² was supplemented with 5-fluoroorotic acid at 2 mg ml⁻¹ and uracil at 5 μ g ml⁻¹. Cultures were grown at 37 °C under anaerobic conditions in a gas mixture containing 85% N₂, 5% CO₂ and 10% H₂. The growth curves were performed in BHIS media with gentle shaking. *E. coli* strains were grown at 37 °C with shaking at 225 r.p.m. in Luria-Bertani broth. The medium was supplemented with chloramphenicol (20 μ g ml⁻¹) and ampicillin (50 μ g ml⁻¹) as needed.

E. coli strain construction. The primers used in this manuscript are provided in Supplementary Table 5b. *C. difficile* 630 genomic DNA was used as the template. To clone the pMTL-YN3- Δ camA construct, primer pairs 2332/2334 and 2333/2335 were used to amplify the region 662 bp upstream and 226 bp downstream of CD630_27580, respectively. The resulting PCR products were cloned into pMTL-YN3 using Gibson assembly⁸³. This construct encodes a CD630_27580 deletion in which the first 14 codons are linked to the last 139 codons with an intervening stop codon between the 5' and 3' end of the gene to avoid production of the last 139 amino acids of CamA. To clone the camA complementation constructs, primer pair 2286/2287 was used to amplify camA and 163 bp of its upstream region. The resulting PCR product was recombined into pMTL-YN1C by Gibson assembly. The N165A complementation construct was cloned in a similar manner, except that the primer pairs consisted of 2286/2532 and 2531/2287. The plasmids were transformed into *E. coli* DH5 α , and the resulting plasmids were confirmed by sequencing and then transformed into HB101/pRK24 for conjugations.

C. difficile strain construction. ACE was used to construct 630 Δ erm Δ pyrE Δ camA using uracil and 5-fluoroorotic acid to select for plasmid excision as previously described⁸⁴. The flanking primer pair 2274/2279 was used to screen for the camA deletion as shown in Extended Data Fig. 5a (the primers are provided in Supplementary Table 5b). Colonies that appeared to harbour gene deletions were validated by performing an internal PCR using a primer (2288) that binds within the region deleted and a primer (2279) that binds to the region flanking the deletion. Two independent clones from the allelic exchange were phenotypically characterized. The camA complementation strains were constructed as previously described using CDDM plates to select for restoration of the pyrE locus by recombination⁸⁴. Two independent clones from each complementation strain were phenotypically characterized.

Cell length measurements. Cells were grown to mid-log and stationary phase in BHIS broth or sporulation was induced as described below for three biological replicates. Cells were imaged using phase-contrast microscopy on a Zeiss Axioskop with a $\times 100/1.3$ NA Zeiss Plan Neofluar objective at each time point. Cell length was calculated using the MicrobeJ plugin for Fiji/ImageJ⁸⁵. Image thresholding was performed using the local default method in MicrobeJ/Fiji to account for variations in background. Cell detection parameters were optimized (area, 0–20 μ m²; length, 1 μ m maximum; width, 0.5–1 μ m) and contours were generated using an interpolated rod-shaped method. Cell length data were exported from MicrobeJ and analysed using Prism 8 (GraphPad).

Sporulation. *C. difficile* strains were inoculated from glycerol stocks overnight onto BHIS-TA plates. Liquid BHIS cultures were inoculated from colonies arising on these plates. The cultures were grown to early stationary phase, back-diluted 1:50 into BHIS, grown until they reached an OD₆₀₀ of between 0.35 and 0.75, and then 120 μ l of this culture was spread onto 70:30 (70% SMC media and 30% BHIS media) plates (40 ml). Sporulating cultures were collected into phosphate-buffered saline (PBS), the sample was pelleted and sporulation levels were visualized by phase-contrast microscopy as previously described²⁵.

Fluorescence microscopy. Fluorescence microscopy was performed on sporulating cultures using Hoechst 33342 (Molecular Probes; 15 μ g ml⁻¹) and FM4-64 (Invitrogen; 1 μ g ml⁻¹) to stain nucleoid and membrane, respectively. Cells were mounted on a 1% agarose in PBS pad. The images were acquired using a Nikon 80i upright epifluorescence microscope with a Nikon $\times 60/1.4$ NA plan apochromat phase-contrast objective in 12-bit format using Nikon NIS elements software. The images were processed using Adobe Photoshop CC for adjustment of brightness, contrast levels and pseudocolouring.

Spore purification. Sporulation was induced on four 70:30 plates for 48–65 h for each strain tested as described above, and spores were purified as previously described⁸⁶. In brief, sporulating cultures were scraped up, washed repeatedly in ice-cold water, incubated overnight in water on ice, treated with DNase I (New England Biolabs) at 37 °C for 45–60 min and then purified on a density gradient (Histodenz, Sigma Aldrich). Spores were resuspended in 600 μ l water for final storage at 4 °C. Spore purity was assessed using phase-contrast microscopy (>95% pure) and the OD₆₀₀ was measured. Spore purification yields were determined from three independent spore preparations. Statistical significance was determined using one-way ANOVA with Tukey's test.

Heat-resistance assay. Heat-resistant spore formation was measured in sporulating *C. difficile* cultures after 20–24 h as previously described²⁵. The H_{RES} efficiency represents the average ratio of heat-resistant CFUs to total CFUs for a given strain relative to the average ratio determined for the WT. H_{RES} was determined on the basis of the average H_{RES} values for a given strain in three biological replicates. Statistical significance was determined using one-way ANOVA with Tukey's test.

Germination assay. Germination assays were performed as previously described⁸⁴. Spores (OD₆₀₀ of 0.35, corresponding to $\sim 1 \times 10^7$) were resuspended in 100 μ l of water, and 10 μ l of this mixture was removed for tenfold serial dilutions in PBS. The dilutions were plated on BHIS-TA, and colonies arising from germinated spores were enumerated after 18–21 h. Germination efficiencies were calculated by averaging the CFUs produced by spores for a given strain relative to the number produced by the WT spores for three biological replicates. Statistical significance was determined by performing one-way ANOVA on natural log-transformed data with Tukey's test. The data were transformed because the use of independent spore preparations resulted in a non-normal distribution. Regardless, no statistical significance in germination efficiency was observed for the mutant and its complements.

Spore chloroform resistance. Spores (OD₆₀₀ of 0.75, corresponding to around 2×10^7 spores) were resuspended in 190 μ l water. Then, 90 μ l of the resuspension was added to tubes containing either 10 μ l of water or chloroform for 15 min, after which 10 μ l of the sample was serially diluted in PBS and plated on BHIS-TA as described previously^{86,87}.

CAAAAA motif abundance and exceptionality. We evaluated the exceptionality of the CAAAAA motif using R'MES³⁰ v.3.1.0. This tool computes scores of exceptionality for k -mers of length l , by comparing observed and expected counts under Markov models that take sequence composition into consideration. R'MES outputs scores of exceptionality, which are—by definition—obtained from P values through the standard one-to-one probit transformation. Analysis of motif abundance was performed using a previously developed framework²⁷ involving an MSR of genomic signals. We created a binary genomic signal for motif content, which was 1 at motif positions and 0 otherwise. We used 50 length scales. Pruning parameter values were set to default and the P -value threshold was set to 10^{-6} .

Whole-genome multiple alignment and classification of CAAAAA positions.

Whole-genome multiple alignment of 37 genomes (36 *C. difficile* isolates and *C. difficile* 630) was produced using the progressiveMauve program⁸⁸ v.2.4.0 with default parameters. As progressiveMauve does not rely on annotations to guide the alignment, we first used the Mauve Contig Mover⁸⁹ to reorder and reorient draft genome contigs according to the reference genome of *C. difficile* 630. A core alignment was built after filtering and concatenating LCBs of at least 50 bp using the stripSubsetLCBs script (<http://darlinglab.org/mauve/snapshots/2015-01-09/linux-x64/>). The lower value chosen for LCB size accounts for the specific aim of maximizing the number of orthologous motifs detected. The XMFA output format of Mauve was converted to VCF format using dedicated scripts, and VCFtools⁹⁰ was used to parse positional variants (SNPs and indels). Orthologous occurrences of the CAAAAA motif were defined if an exact match to the motif was present in each of the 37 genomes (conserved orthologous positions) or if at least one motif (and a maximum of $n - 1$, with n being the number of genomes) contained positional polymorphisms (maximum of two SNPs or indels per motif; variable orthologous positions). Non-orthologous occurrences of CAAAAA were obtained from the whole-genome alignment before the extraction of LCBs and correspond to those situations in which the CAAAAA motif was absent in at least one genome. Typically, these correspond to regions containing MGEs or unaligned repetitive regions.

Identification of TFBSs and TSSs. Identification of TFBSs was performed by retrieving *C. difficile* 630 regulatory sites in FASTA format from the RegPrecise database⁹¹ (accessed July 2017). These were converted to position-specific scoring matrices (PSSMs) using in-house-developed scripts. This led to a total of 21 PSSMs pertaining to 14 distinct TF families (Supplementary Table 7b). Matches between these matrices and *C. difficile* genomes were performed using MAST⁹² (default settings). MAST output was filtered on the basis of P value. Hits with $P < 10^{-9}$ were considered to be positive, whereas hits of $P > 10^{-5}$ were considered to be negative. Hits with intermediate P values were only considered to be positive if the P value of the hit divided by the P value of the worst positive hit was lower than 100. For the CcpA, LexA, NrdR and CodY TFs (which have shorter binding sites), hits with $P < 10^{-8}$ were considered to be positive. TSSs were predicted with Parseq⁹³ using the 'fast' speed option from multiple RNA-seq datasets (see below). Transcription and breakpoint probabilities were computed using a background expression level threshold of 0.1 and a score penalty of 0.05. We retained only high-confidence 5' breakpoint hits, located at a maximum distance of 200 bp from the nearest start codon. A ± 5 bp window around the TSS was considered if only one single predicted value was obtained; otherwise, we considered an interval that was delimited by the minimum and maximum values predicted by Parseq.

RNA processing. For analyses of sporulating cell transcriptomes, RNA was extracted from three biological replicates of WT and Δ camA growing on 70:30 sporulation media after 9 h and 10.5 h of growth using the FastRNA Pro Blue Kit (MP Biomedical) and the FastPrep-24 automated homogenizer (MP Biomedical), similar to previous research³². For analyses of the mid-log and early stationary phase cultures, overnight cultures of WT and Δ camA in BHIS were back-diluted 1:50 into three biological replicates of 30 ml of BHIS in 125 ml Erlenmeyer flasks. The cultures were grown until mid-log phase ($OD_{600} = 0.5-0.6$) and early stationary phase ($OD_{600} = 1.3-1.4$). RNA was collected from 15 ml and 10 ml of the same cultures for the mid-log and early stationary-phase cultures, respectively. Contaminating genomic DNA was depleted using three successive DNase treatments, with the final treatment being on-column using the Qiagen RNeasy kit. The samples were tested for genomic DNA contamination using qPCR for 16S rRNA and the *sleC* gene. DNase-treated RNA (15 μ g) was enriched for mRNA using the Ribo-Zero Magnetic Kit (Epicentre) for the broth-grown cultures. Ribosomal RNA was depleted from RNA that was collected from sporulating cultures using the Ambion MICROExpress Bacterial mRNA Enrichment Kit (Thermo Fisher) because Ribo-Zero kits were temporarily discontinued. The quality of total RNA was validated using an Agilent 2100 Bioanalyzer. Samples for RT-qPCR analyses were collected in triplicate from a separate set of three biological replicates that was grown identically to the cultures used for RNA-seq analyses. The RNA was processed similarly except that mRNA enrichment was performed using a MICROExpress Bacterial mRNA Enrichment Kit, and the DNase-treated RNA samples for RT-qPCR analyses were tested for genomic DNA contamination using qPCR for *rpoB*.

RNA-seq, read alignment and differential-expression analysis. Purified RNA was extracted from three biological replicates of sporulating (9 h and 10.5 h) and exponential and stationary growth cultures of *C. difficile* 630 Δ erm and *C. difficile* 630 Δ erm Δ camA, DNase-treated, ribosomal RNA-depleted and converted to cDNA as previously described³². RNA-seq was performed using a HiSeq 2500, yielding 29.4 ± 4.5 million (mean \pm s.d.) 100 bp single-end reads per sample (exponential and stationary growth time points) and 26.9 ± 4.3 million (mean \pm s.d.) 150 bp paired-end reads per sample (sporulation time points). Read quality was checked using FastQC v.0.11.5 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>). We used Trimmomatic⁹⁴ v.0.39 to remove adapters and low-quality reads (parameters: PE, -phred33, ILLUMINACLIP:<adapters.fa>:2:30:10:8:True, SLIDINGWINDOW:4:15, LEADING:20 TRAILING:20, MINLEN:50). Subsequently, rRNA sequences were filtered from the dataset using SortMeRNA⁹⁵ v.2.1 on the basis of the SILVA 16S and 23S rRNA databases⁹⁶ and the Rfam 5S rRNA database⁹⁷. The resulting non-rRNA reads were mapped to the *C. difficile* 630 reference genome using BWA-MEM v.0.17-r1198 (ref.⁹⁸). The resulting BAM files were sorted and indexed using SAMTOOLS, and read assignment was performed using featureCounts⁹⁹ v.1.6.4 (excluding multi-mapping and multi-overlapping reads). A gene was included for differential-expression analysis if it had more than one count in all of the samples. Normalization and differential-expression testing were performed using the Bioconductor package DESeq2 v.1.18.1 (ref.¹⁰⁰). DE genes were defined as genes with an FDR-corrected $P < 0.05$ and $|\log_2[FC]| > 1$. Functional classification of genes was performed using the DAVID online database (<https://david.ncicrf.gov/>)⁹⁸. GO annotation terms with a gene count of at least 5 and $P < 0.05$ (one-tailed Fisher's exact test, FDR corrected) were considered to be significant. The reproducibility of DAVID's functional classification was tested with Blast2GO¹⁰¹ v.5.2 and Panther¹⁰² v.14. In brief, for Blast2GO, we ran BLASTX searches of the *C. difficile* 630 genome against the entire GenBank bacterial protein database (as of September 2018). The output, in XML format, was loaded into Blast2GO, and mapping, annotation and enrichment analysis was performed as indicated (<http://docs.blast2go.com/user-manual/quick-start/>). For Panther, we downloaded the most recent HMM library (ftp.pantherdb.org/hmm_scoring/13.1/PANTHER13.1_hmmscoring.tgz) and annotated our *C. difficile* 630 protein set with pantherScore2.1.pl. Both input and background gene lists were formatted to the Panther Generic Mapping File type as described at <http://www.pantherdb.org>. To assess the significance of the intersection between multiple datasets of DE genes (typically observed during *C. difficile* colonization and infection), we collected gene-expression data from in vivo and in vitro studies⁴⁰⁻⁴², in which key factors for gut colonization (such as time after infection, antibiotic exposure and spatial structure (planktonic and biofilm growth)) were tested. DE genes were called using the same conditions as described above. Statistical analyses and graphical representation of multiset intersections were performed using the R package SuperExactTest¹⁰³.

RT-qPCR. Transcript levels were determined from cDNA templates that were prepared from the three biological replicates described above. Gene-specific primer pairs are provided in Supplementary Table 5b. RT-qPCR was performed as described previously³³, except that we used iTaq Universal SYBR Green supermix (BioRad), 50 nM of gene specific primers and a Mx3005P qPCR system (Stratagene) in a total volume of 25 μ l. The following cycling conditions were used: 95 °C for 2 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 min. Transcript levels were normalized to the housekeeping gene *rpoB* using the standard curve method.

Western blots. Sporulation protein analyses. Sporulation was induced as indicated, and samples were collected and processed for immunoblotting as described previously⁸⁶. Total protein in each sample was quantified using the Pierce 660 nm protein assay with ionic detergent compatibility reagent (Thermo Fisher), and 5 μ g of protein was loaded for each sample. σ^F , σ^E and Spo0A were resolved using 15% SDS-polyacrylamide gel electrophoresis (SDS-PAGE) gels, whereas SpoIIQ and SpoIVA were resolved using 12% SDS-PAGE gels. Proteins were transferred to polyvinylidene difluoride membranes, which were subsequently probed with rabbit (anti- σ^F , anti- σ^E and anti-SpoIIQ) and mouse (anti-Spo0A and anti-SpoIVA) polyclonal primary antibodies, and anti-rabbit IR800 and anti-mouse IR680 secondary antibodies (LI-COR). Blots were imaged using an LiCor Odyssey CLx imaging system. The results shown are representative of analyses of two biological replicates.

Toxin analyses. Overnight cultures of *C. difficile* were diluted 1:50 in TY medium and incubated at 37 °C for 24 h. Cells were collected using centrifugation, suspended in SDS-PAGE buffer and boiled for 10 min. The samples were then run on 4-20% Mini-PROTEAN TGX Precast Protein Gels (Bio Rad) and transferred to a nitrocellulose membrane. TcdA was detected as described previously using mouse anti-TcdA primary antibodies (Novus Biologicals) and goat anti-mouse IgG conjugated with IR800 (Thermo Fisher)¹⁰⁴.

Animal infection studies. All of the animal experiments was performed under the guidance of veterinarians and trained animal technicians within the University of North Carolina Division of Comparative Medicine. Animal experiments were

performed with prior approval from the UNC Institutional Animal Care and Use Committee. Animals considered to be moribund as defined in the protocols were euthanized by CO₂ asphyxiation followed by a secondary physical method in accordance with the Panel on Euthanasia of the American Veterinary Medical Association. The University complies with state and federal Animal Welfare Acts, the standards and policies of the Public Health Service.

Mouse model. The parental *C. difficile* strain 630Δ*erm*, the MTase mutant 630Δ*erm*Δ*camA* and the MTase complemented strain were evaluated in an antibiotic-treated mouse model as previously described^{105,106}. Groups of 8-to-10-week old female and male C57BL/6 mice (*Mus musculus*; Charles River Laboratories) were administered a cocktail of antibiotics (kanamycin (400 μg ml⁻¹), gentamicin (35 μg ml⁻¹), colistin (850 U ml⁻¹), vancomycin (45 μg ml⁻¹) and metronidazole (215 μg ml⁻¹) in their water ad libitum 7 d before inoculation for 3 d, followed by a single intra-peritoneal dose of clindamycin (10 mg kg⁻¹ body weight) 2 d before inoculation. The mice were randomly assigned into groups, with 2 female mice assigned to the mock condition and 6 mice (3 male and 3 female) to each infection condition. The experiment was independently repeated to assess the consistency of the data. The data from the experiments were combined for analysis for a total of 12 mice (6 male and 6 female) in each infection condition. Mice were inoculated with 10⁵ spores by oral gavage. Mock-inoculated mice were included as controls. Cage changes were performed every 48 h after inoculation. Faecal samples were collected every 24 h for 7 d after inoculation. Dilutions were plated on BHIS agar containing 0.1% of the germinant TA to enumerate spores as CFU g⁻¹ of faeces.

Hamster model. The above strains were tested in Syrian golden hamster strain LVG (*Mesocricetus auratus*; Charles River Laboratories) as described previously¹⁰⁷. Hamsters were randomly assigned into groups, with 2 assigned to the mock condition and 6 (3 male and 3 female) to each infection condition. Hamsters were administered a single dose of clindamycin (30 mg kg⁻¹ body weight) by oral gavage, then inoculated with approximately 5,000 spores of the above strains 5 d later. Hamsters were monitored for weight loss and diarrhoeal symptoms and were considered to be moribund after 15–20% weight loss from their maximum body weight, with or without concurrent diarrhoea.

Biofilm assays. Biofilm assays were performed as previously described¹⁰⁸. In brief, overnight cultures of *C. difficile* were diluted 1:100 in BHIS supplemented with 1% glucose and 50 mM sodium phosphate buffer (pH 7.5) in 24-well polystyrene plates. After 24 h of growth at 37 °C, supernatants were removed, the biofilms were washed once with PBS and then stained for 30 min with 0.1% (w/v) crystal violet. After 30 min, the biofilms were washed again with PBS, and the crystal violet was solubilized with ethanol. Absorbance was read at 570 nm. Three independent experiments were performed, and each strain was assayed in quadruplicate in each experiment.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Genome assemblies and methylation data are available from NCBI under BioProject ID PRJNA448390. RNA-seq data are available under project ID PRJNA445308. Additional data are available from the corresponding authors on reasonable request.

Code availability

Scripts and a tutorial supporting all of the key analyses of this research are publicly available as a package named Bacterial Epigenome Analysis SuiTe (BEAST) at <http://github.com/fanglab/>.

Received: 14 August 2018; Accepted: 18 October 2019;

Published online: 25 November 2019

References

- Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H. & Kuijper, E. J. *Clostridium difficile* infection. *Nat. Rev. Dis. Primers* **2**, 16020 (2016).
- Sebahia, M. et al. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **38**, 779–786 (2006).
- He, M. et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* **45**, 109–113 (2013).
- Herbert, M., O’Keefe, T. A., Purdy, D., Elmore, M. & Minton, N. P. Gene transfer into *Clostridium difficile* CD630 and characterisation of its methylase genes. *FEMS Microbiol. Lett.* **229**, 103–110 (2003).
- van Eijk, E. et al. Complete genome sequence of the *Clostridium difficile* laboratory strain 630Δ*erm* reveals differences from strain 630, including translocation of the mobile element CTn5. *BMC Genom.* **16**, 31 (2015).
- Hargreaves, K. R., Thanki, A. M., Jose, B. R., Oggioni, M. R. & Clokie, M. R. Use of single molecule sequencing for comparative genomics of an environmental and a clinical isolate of *Clostridium difficile* ribotype 078. *BMC Genom.* **17**, 1020 (2016).
- Casadesus, J. & Low, D. Epigenetic gene regulation in the bacterial world. *Microbiol. Mol. Biol. Rev.* **70**, 830–856 (2006).
- Low, D. A., Weyand, N. J. & Mahan, M. J. Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect. Immun.* **69**, 7197–7204 (2001).
- Cohen, N. R. et al. A role for the bacterial GATC methylome in antibiotic stress survival. *Nat. Genet.* **48**, 581–586 (2016).
- Manso, A. S. et al. A random six-phase switch regulates pneumococcal virulence via global epigenetic changes. *Nat. Commun.* **5**, 5055 (2014).
- Atack, J. M. et al. A biphasic epigenetic switch controls immunoevasion, virulence and niche adaptation in non-typeable *Haemophilus influenzae*. *Nat. Commun.* **6**, 7828 (2015).
- Wion, D. & Casadesus, J. N⁶-methyl-adenine: an epigenetic signal for DNA-protein interactions. *Nat. Rev. Microbiol.* **4**, 183–192 (2006).
- Oliveira, P. H., Touchon, M. & Rocha, E. P. Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl Acad. Sci. USA* **113**, 5658–5663 (2016).
- Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
- Beaulaurier, J., Schadt, E. E. & Fang, G. Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* **20**, 157–172 (2019).
- Fang, G. et al. Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* **30**, 1232–1239 (2012).
- Murray, I. A. et al. The methylomes of six bacteria. *Nucleic Acids Res.* **40**, 11450–11462 (2012).
- Davis, B. M., Chao, M. C. & Waldor, M. K. Entering the era of bacterial epigenomics with single molecule real time DNA sequencing. *Curr. Opin. Microbiol.* **16**, 192–198 (2013).
- Smits, W. K. Hype or hypervirulence: a reflection on problematic *C. difficile* strains. *Virulence* **4**, 592–596 (2013).
- Roberts, R. J. et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* **31**, 1805–1812 (2003).
- Wust, J., Sullivan, N. M., Hardegger, U. & Wilkins, T. D. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. *J. Clin. Microbiol.* **16**, 1096–1101 (1982).
- Barra-Carrasco, J. & Paredes-Sabja, D. *Clostridium difficile* spores: a major threat to the hospital environment. *Future Microbiol.* **9**, 475–486 (2014).
- Dembek, M. et al. High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. *mBio* **6**, e02383 (2015).
- Donnelly, M. L., Fimlaid, K. A. & Shen, A. Characterization of *Clostridium difficile* spores lacking either SpoVAC or dipicolinic acid synthetase. *J. Bacteriol.* **198**, 1694–1707 (2016).
- Shen, A., Fimlaid, K. A. & Pishdadian, K. Inducing and quantifying *Clostridium difficile* spore formation. *Methods Mol. Biol.* **1476**, 129–142 (2016).
- Schbath, S. & Hoebeke, M. in *Advances in Genomic Sequence Analysis and Pattern Discovery* Vol. 7 (eds Elmlits, L. et al.) 25–64 (World Scientific, 2011).
- Knijnenburg, T. A. et al. Multiscale representation of genomic signals. *Nat. Methods* **11**, 689–694 (2014).
- Huang, D. W. et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.* **35**, W169–W175 (2007).
- Lim, H. N. & van Oudenaarden, A. A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nat. Genet.* **39**, 269–275 (2007).
- Ardissone, S. et al. Cell cycle constraints and environmental control of local DNA hypomethylation in α-Proteobacteria. *PLoS Genet.* **12**, e1006499 (2016).
- Cota, I. et al. OxyR-dependent formation of DNA methylation patterns in OpvABOFF and OpvABON cell lineages of *Salmonella enterica*. *Nucleic Acids Res.* **44**, 3595–3609 (2016).
- Fimlaid, K. A. et al. Global analysis of the sporulation pathway of *Clostridium difficile*. *PLoS Genet.* **9**, e1003660 (2013).
- Pishdadian, K., Fimlaid, K. A. & Shen, A. SpoIIID-mediated regulation of σK function during *Clostridium difficile* sporulation. *Mol. Microbiol.* **95**, 189–208 (2015).
- Fimlaid, K. A. & Shen, A. Diverse mechanisms regulate sporulation sigma factor activity in the Firmicutes. *Curr. Opin. Microbiol.* **24**, 88–95 (2015).
- Saujet, L., Pereira, F. C., Henriques, A. O. & Martin-Verstraete, I. The regulatory network controlling spore formation in *Clostridium difficile*. *FEMS Microbiol. Lett.* **358**, 1–10 (2014).

36. Saujet, L. et al. Genome-wide analysis of cell type-specific gene transcription during spore formation in *Clostridium difficile*. *PLoS Genet.* **9**, e1003756 (2013).
37. Rosenbusch, K. E., Bakker, D., Kuijper, E. J. & Smits, W. K. C. *difficile* 630Δerm Spo0A regulates sporulation, but does not contribute to toxin production, by direct high-affinity binding to target DNA. *PLoS ONE* **7**, e48608 (2012).
38. Fimlaid, K. A., Jensen, O., Donnelly, M. L., Siegrist, M. S. & Shen, A. Regulation of *Clostridium difficile* spore formation by the SpoIIQ and SpoIIIA proteins. *PLoS Genet.* **11**, e1005562 (2015).
39. Ribis, J. W., Fimlaid, K. A. & Shen, A. Differential requirements for conserved peptidoglycan remodeling enzymes during *Clostridioides difficile* spore formation. *Mol. Microbiol.* **110**, 370–389 (2018).
40. Maldarelli, G. A. et al. Type IV pili promote early biofilm formation by *Clostridium difficile*. *Pathog. Dis.* **74**, ftw061 (2016).
41. Jenior, M. L., Leslie, J. L., Young, V. B. & Schloss, P. D. *Clostridium difficile* colonizes alternative nutrient niches during infection across distinct murine gut microbiomes. *mSystems* **2**, e00063-17 (2017).
42. Fletcher, J. R., Erwin, S., Lanzas, C. & Theriot, C. M. Shifts in the gut metabolome and *Clostridium difficile* transcriptome throughout colonization and infection in a mouse model. *mSphere* **3**, e00089-18 (2018).
43. Lessa, F. C. et al. Burden of *Clostridium difficile* infection in the United States. *N. Engl. J. Med.* **372**, 825–834 (2015).
44. Deakin, L. J. et al. The *Clostridium difficile* spo0A gene is a persistence and transmission factor. *Infect. Immun.* **80**, 2704–2711 (2012).
45. Lewis, B. B. & Pamer, E. G. Microbiota-based therapies for *Clostridium difficile* and antibiotic-resistant enteric infections. *Annu. Rev. Microbiol.* **71**, 157–178 (2017).
46. Abt, M. C., McKenney, P. T. & Pamer, E. G. *Clostridium difficile* colitis: pathogenesis and host defence. *Nat. Rev. Microbiol.* **14**, 609–620 (2016).
47. Sanchez-Romero, M. A., Cota, I. & Casadesus, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).
48. Griffiths, D. et al. Multilocus sequence typing of *Clostridium difficile*. *J. Clin. Microbiol.* **48**, 770–778 (2010).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
52. Oliveira, P. H., Touchon, M. & Rocha, E. P. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* **42**, 10618–10631 (2014).
53. Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* **43**, D298–D299 (2015).
54. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
55. Katoh, K. & Standley, D. M. MAFFT: iterative refinement and additional methods. *Methods Mol. Biol.* **1079**, 131–146 (2014).
56. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* **27**, 221–224 (2010).
57. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
58. Bland, C. et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **8**, 209 (2007).
59. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
60. Goldfarb, T. et al. BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).
61. Ofir, G. et al. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90–98 (2018).
62. Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol. Direct* **4**, 29 (2009).
63. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
64. Xie, Y. et al. TADB 2.0: an updated database of bacterial type II toxin-antitoxin loci. *Nucleic Acids Res.* **46**, D749–D753 (2018).
65. Fouts, D. E. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res.* **34**, 5839–5851 (2006).
66. Arndt, D. et al. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res.* **44**, W16–W21 (2016).
67. Cury, J., Jove, T., Touchon, M., Neron, B. & Rocha, E. P. Identification and analysis of integrons and cassette arrays in bacterial genomes. *Nucleic Acids Res.* **44**, 4539–4550 (2016).
68. Cury, J., Touchon, M. & Rocha, E. P. C. Integrative and conjugative elements and their hosts: composition, distribution and organization. *Nucleic Acids Res.* **45**, 8943–8956 (2017).
69. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinform.* **5**, 113 (2004).
70. Crisculo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
71. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
72. Touchon, M. et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).
73. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinform.* **12**, 116 (2011).
74. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).
75. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
76. Sawyer, S. Statistical tests for detecting gene conversion. *Mol. Biol. Evol.* **6**, 526–538 (1989).
77. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
78. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
79. Csuros, M. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* **26**, 1910–1912 (2010).
80. Ng, Y. K. et al. Expanding the repertoire of gene tools for precise manipulation of the *Clostridium difficile* genome: allelic exchange using *pyrE* alleles. *PLoS ONE* **8**, e56051 (2013).
81. Sorg, J. A. & Dineen, S. S. Laboratory maintenance of *Clostridium difficile*. *Curr. Protoc. Microbiol.* **12**, 9A.1.1–9A.1.10 (2009).
82. Cartman, S. T. & Minton, N. P. A mariner-based transposon system for in vivo random mutagenesis of *Clostridium difficile*. *Appl. Environ. Microbiol.* **76**, 1103–1109 (2010).
83. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
84. Donnelly, M. L. et al. A *Clostridium difficile*-specific, gel-forming protein required for optimal spore germination. *mBio* **8**, e02085-16 (2017).
85. Ducret, A., Quardokus, E. M. & Brun, Y. V. MicroBeJ, a tool for high throughput bacterial cell detection and quantitative analysis. *Nat. Microbiol.* **1**, 16077 (2016).
86. Ribis, J. W., Ravichandran, P., Putnam, E. E., Pishdadian, K. & Shen, A. The conserved spore coat protein SpoVM is largely dispensable in *Clostridium difficile* spore formation. *mSphere* **2**, e00315-17 (2017).
87. Edwards, A. N. et al. Chemical and stress resistances of *Clostridium difficile* spores and vegetative cells. *Front. Microbiol.* **7**, 1698 (2016).
88. Darling, A. E., Mau, B. & Perna, N. T. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
89. Rissman, A. I. et al. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics* **25**, 2071–2073 (2009).
90. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
91. Novichkov, P. S. et al. RegPrecise 3.0—a resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genom.* **14**, 745 (2013).
92. Bailey, T. L. & Gribskov, M. Combining evidence using *p*-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
93. Mirault, B., Nicolas, P. & Richard, H. Parseq: reconstruction of microbial transcription landscape from RNA-seq read counts using state-space models. *Bioinformatics* **30**, 1409–1416 (2014).
94. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
95. Kopylova, E., Noe, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211–3217 (2012).
96. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
97. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. & Eddy, S. R. Rfam: an RNA family database. *Nucleic Acids Res.* **31**, 439–441 (2003).
98. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

99. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
100. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
101. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
102. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2019).
103. Wang, M., Zhao, Y. & Zhang, B. Efficient test and visualization of multi-set intersections. *Sci. Rep.* **5**, 16923 (2015).
104. Anjuwon-Foster, B. R., Maldonado-Vazquez, N. & Tamayo, R. Characterization of flagellum and toxin phase variation in *Clostridioides difficile* ribotype 012 isolates. *J. Bacteriol.* **200**, e00056-18 (2018).
105. Chen, X. et al. A mouse model of *Clostridium difficile*-associated disease. *Gastroenterology* **135**, 1984–1992 (2008).
106. McKee, R. W., Aleksanyan, N., Garrett, E. M. & Tamayo, R. Type IV pili promote *Clostridium difficile* adherence and persistence in a mouse model of infection. *Infect. Immun.* **86**, e00943-17 (2018).
107. Woods, E. C., Edwards, A. N., Childress, K. O., Jones, J. B. & McBride, S. M. The *C. difficile* *chnRAB* operon initiates adaptations to the host environment in response to LL-37. *PLoS Pathog.* **14**, e1007153 (2018).
108. Purcell, E. B. et al. A nutrient-regulated cyclic diguanylate phosphodiesterase controls *Clostridium difficile* biofilm and toxin production during stationary phase. *Infect. Immun.* **85**, e00347-17 (2017).
109. Pereira, F. C. et al. The spore differentiation pathway in the enteric pathogen *Clostridium difficile*. *PLoS Genet.* **9**, e1003782 (2013).
110. Serrano, M. et al. A recombination directionality factor controls the cell type-specific activation of σ^k and the fidelity of spore development in *Clostridium difficile*. *PLoS Genet.* **12**, e1006312 (2016).
111. Theriot, C. M. et al. Cefoperazone-treated mice as an experimental platform to assess differential virulence of *Clostridium difficile* strains. *Gut Microbes* **2**, 326–334 (2011).

Acknowledgements

We thank R. J. Roberts (New England Biolabs) for his help with the prediction of R–M systems and orphan MTases in *C. difficile* genomes using REBASE Tools and for providing comments. He was originally an author of this manuscript; however, as a staunch supporter of the open access movement, he will not author a paper that is not open access. We also thank E. P. C. Rocha (Institut Pasteur, Paris, France) for reading the manuscript and for providing comments. The research was primarily funded by R01 GM114472 (to G.F.) from the National Institutes of Health and Icahn Institute for Genomics and Multiscale Biology. The research was also funded by NIH grants R01 AI119145 (to H.v.B and A.B.), R01 AI22232 (to A.S.), R01 AI107029 (to R.T.) and R35

GM131780 (to A.K.A.), a Hirschl Research Scholar award from the Irma T. Hirschl/Monique Weill-Caulier Trust (to G.F.), a Pew Scholar in the Biomedical Sciences grant from the Pew Charitable (to A.S.). G.F. is a Nash Family Research Scholar. A.S. holds an Investigators in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome Fund. J.W.R. was supported by an NIH training grant 5T32GM007310-42. The participation of R. J. Roberts in this project was funded by New England Biolabs. This research was also supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Author contributions

G.F. conceived the hypothesis. A.S. and G.F. supervised the project. P.H.O. and G.F. designed the computational methods. P.H.O., R.T., A.S. and G.F. designed the experiments. P.H.O. performed most of the computational analyses and developed most of the scripts that support the analyses. J.W.R. performed the growth curves, microscopy analyses (fluorescence and phase contrast), analyses of cell length and sporulation stage, isolation of some of the RNA and processed it for RT–qPCR studies, and RT–qPCR analyses of sporulation genes. A.S. constructed the deletion and catalytic *AcamA* mutants, performed complementation, isolated and processed the RNA for several of the RNA analyses, and performed many of the sporulation phenotypic assays. E.M.G. and D.T. performed the animal infection experiment and analysed the data under the supervision of R.T. A.Kim and G.F. performed methylation motif discovery and refinement. O.S. and E.A.M. performed RT–qPCR controls for RNA-seq analyses. O.S., E.A.M., G.D., M.L.-S., C.B., N.E.Z., D.R.A., I.O., G.P., F.W., C.H., S.H., R.S., H.v.B. and A.S. contributed to the other experiments. G.D., I.O. and R.S. designed and conducted SMRT-seq. P.H.O., J.W.R., E.M.G., D.T., A.Kim, O.S., T.P., S.Z., E.A.M., M.T., C.B., S.B., A.K.A., A.B., R.T., E.E.S., R.S., H.v.B., A.Kasarskis, R.T., A.S. and G.F. analysed the data. P.H.O., R.T., A.S. and G.F. wrote the manuscript with additional information inputs from other co-authors.

Competing interests

A.S. has a consultant role for BioVector, a diagnostic start-up. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41564-019-0613-4>.

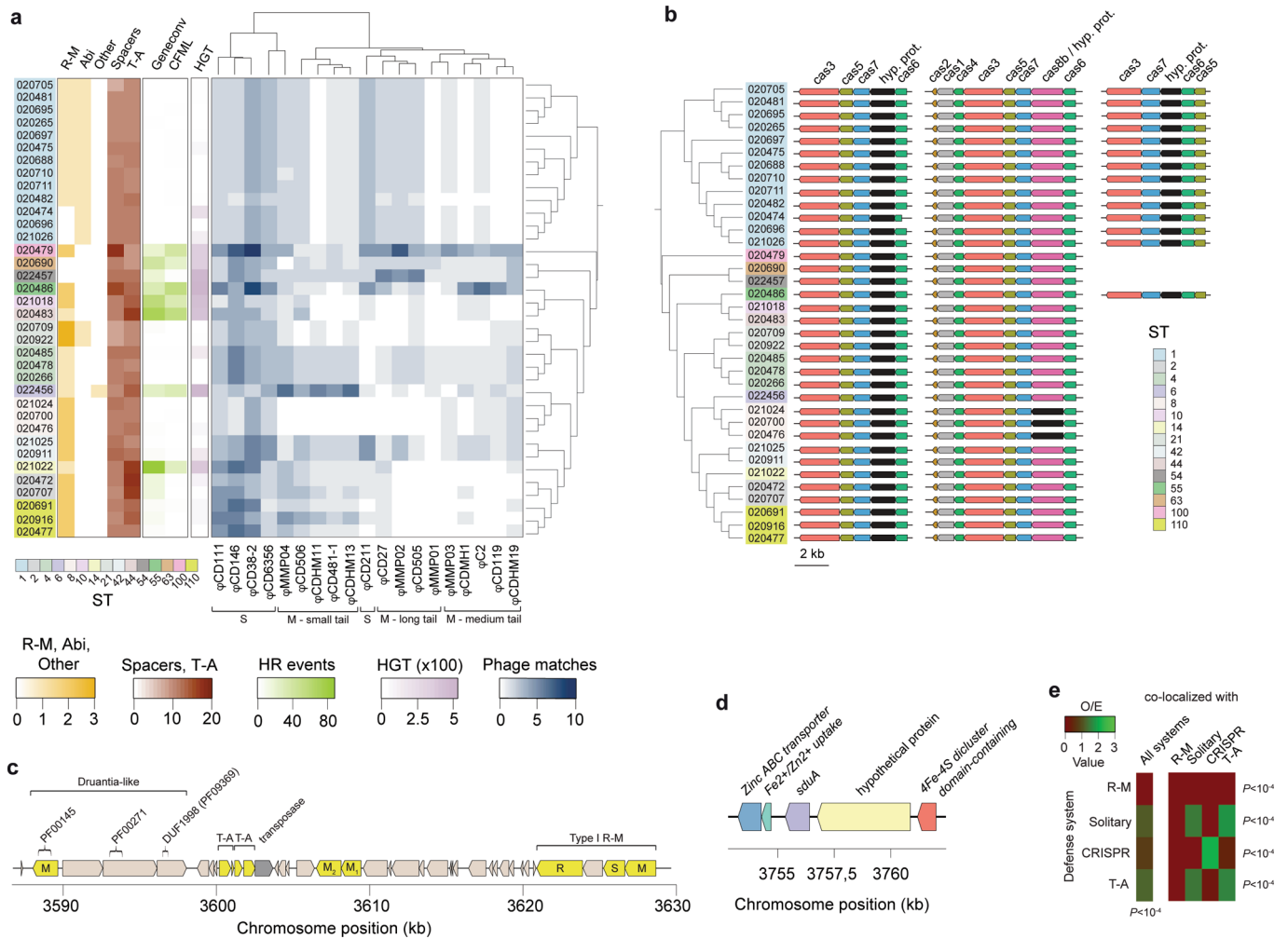
Supplementary information is available for this paper at <https://doi.org/10.1038/s41564-019-0613-4>.

Correspondence and requests for materials should be addressed to A.S. or G.F.

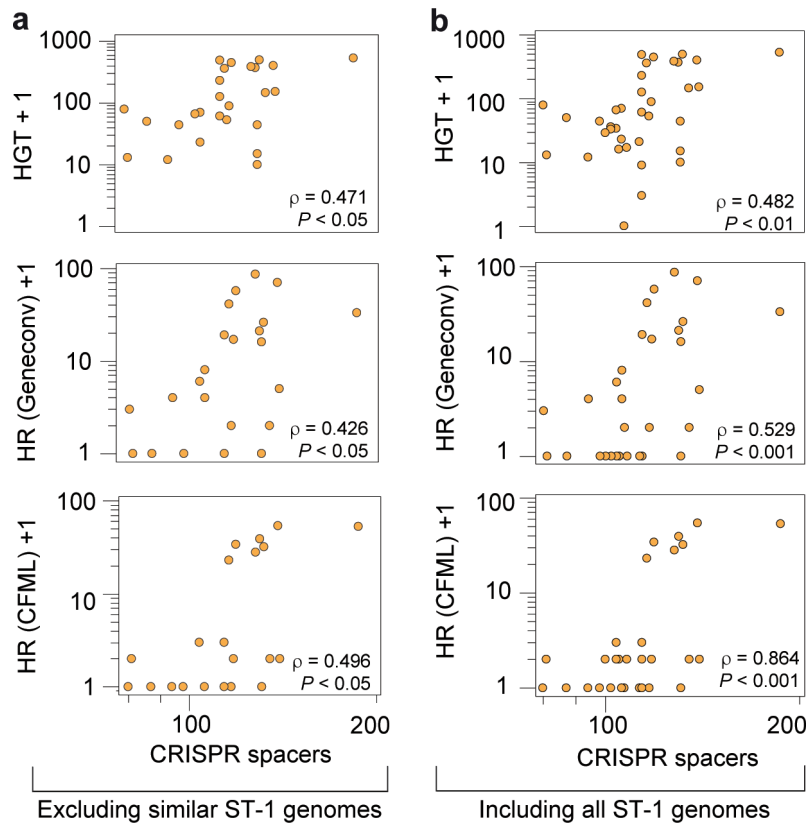
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

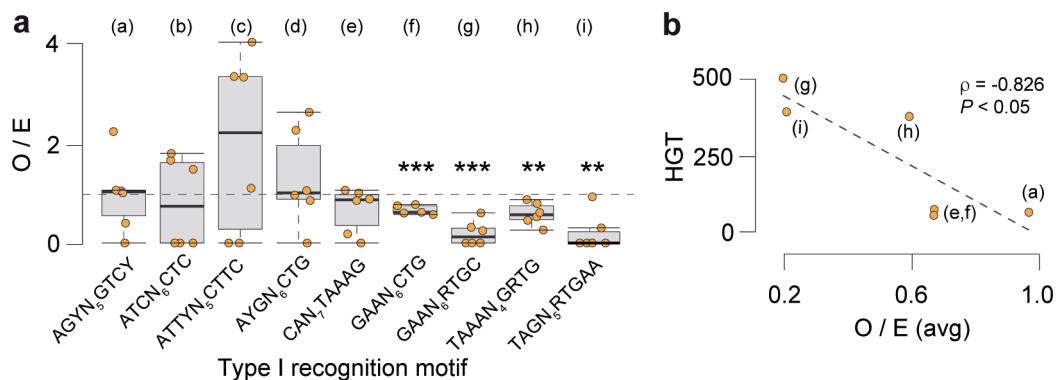
© The Author(s), under exclusive licence to Springer Nature Limited 2019



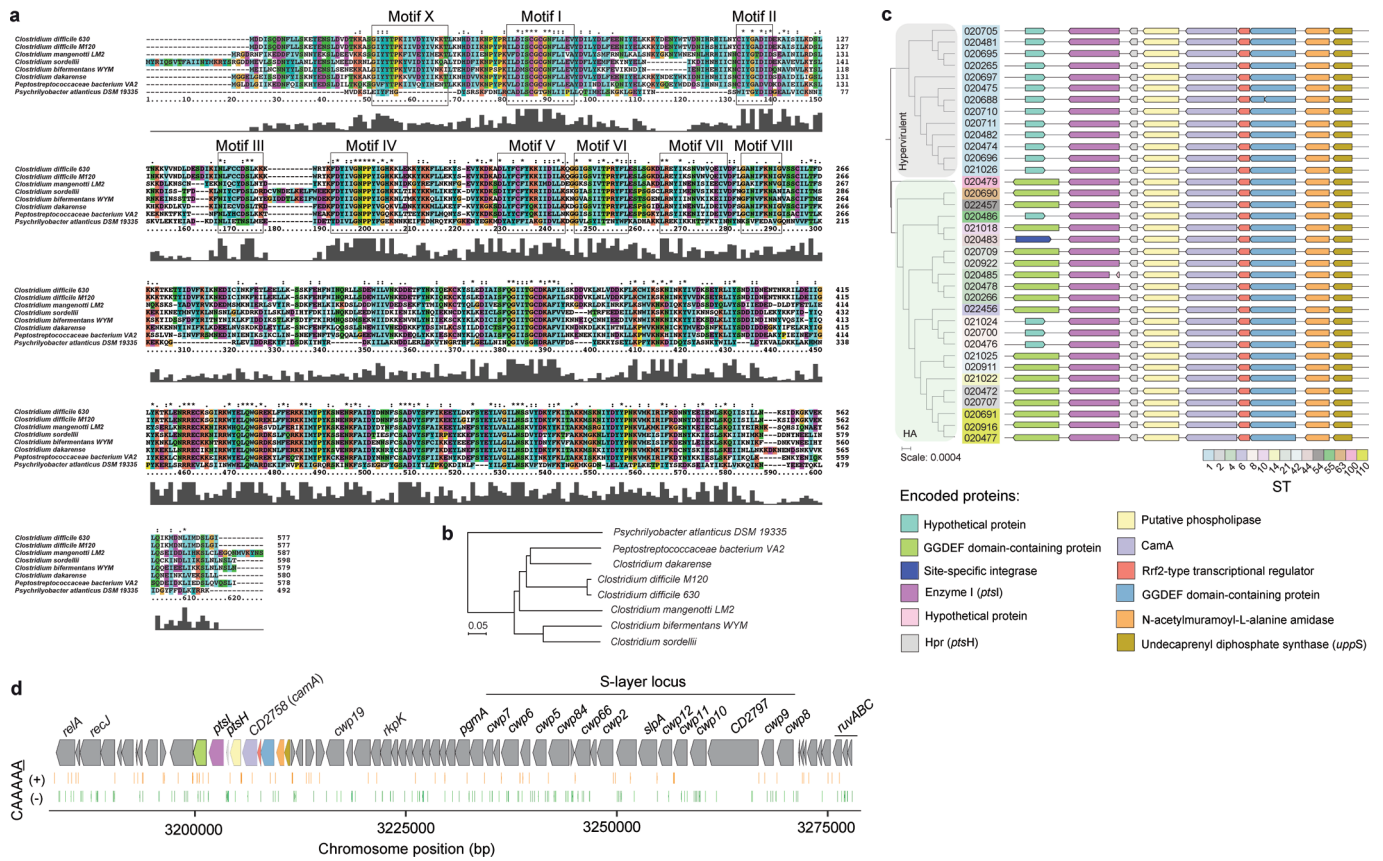
Extended Data Fig. 1 | Multiple defense systems and gene flux control in *C. difficile*. (a) Heatmap aggregate depicts: abundance of defense systems (R-M, abortive infection (Abi), average number of spacers per CRISPR, toxin-antitoxin (T-A), and Shedu systems (other)), homologous recombination (HR) events (given by Geneconv and ClonalFrameML (CFML)), horizontal gene transfer (HGT, given by Wagner parsimony), and number of phage-targeting CRISPR spacers (Supplementary Notes). Phages were clustered according to their family (*Siphoviridae* (S), *Myoviridae* (M)), and tail type. (b) Cas genes detected in *C. difficile*. Apart from the complete Type-IB gene cluster (*cas1-cas8*), we also observed two truncated gene clusters lacking *cas1*, *cas2*, and *cas4*. One of the truncated operons was present across all genomes, while the second was restricted to ST-1 and ST-55. (c) Example of a putative 'defense island' detected in CD_020472 harboring: a Druantia-like system, two T-A systems, two solitary MTases, and one Type I R-M system. The Druantia-like system is similar to the previously reported Type II Druantia systems⁶³ in the sense that a PF00271 helicase conserved C-terminal domain and DUF1998 (PF09369) are associated with a nearby cytosine methylase. However, it lacks a PF00270 DEAx box helicase. (d) Genomic context of the *sduA* gene in CD_22456 pertaining to the newly identified Shedu defense system. The gene is located in an integrative conjugative element (ICE) (Supplementary Table 2d). (e) Observed/expected (O/E) ratios for co-localized defense systems (maximum of 10 genes apart). Only the most abundant systems were included in the analysis. Expected values were obtained by multiplying the total number of defense systems by the fraction of co-localized defense systems. *P* values correspond to the Chi-square test.



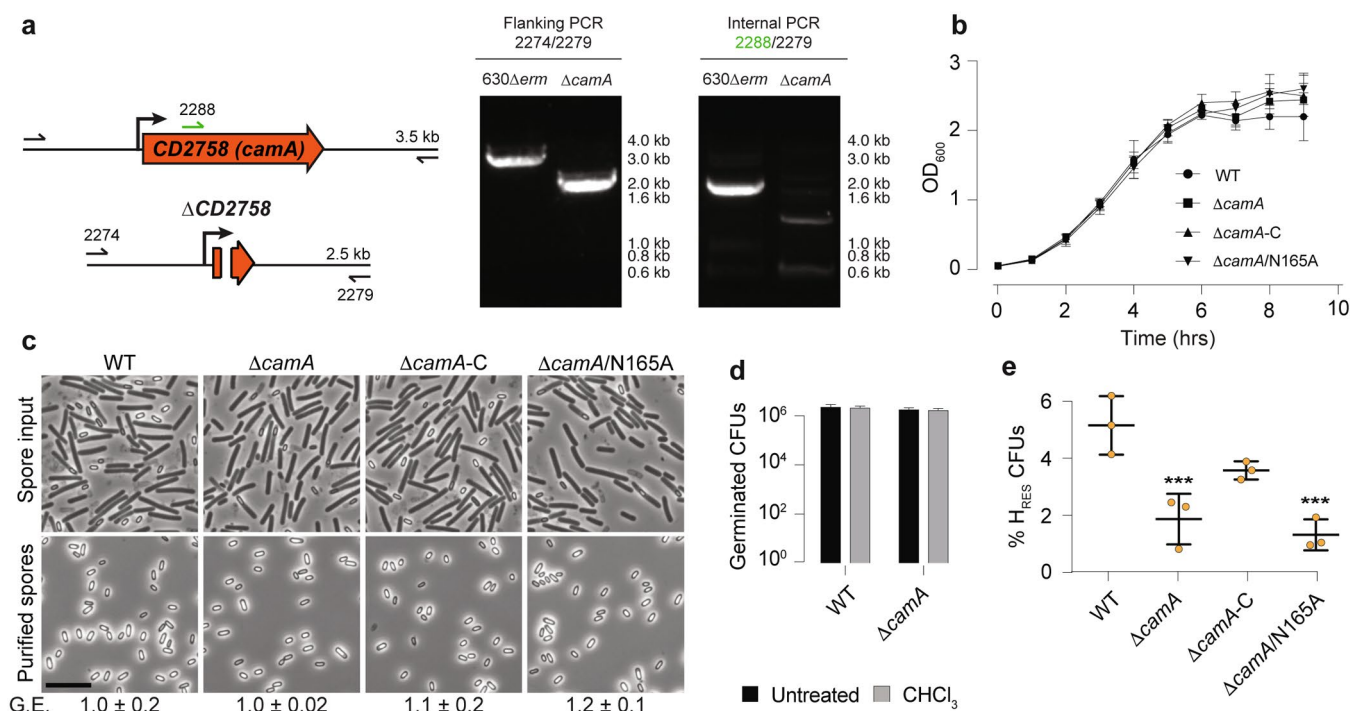
Extended Data Fig. 2 | Relation between gene flux and CRISPR spacer content. (a) Association between genetic flux (horizontal gene transfer (HGT) and homologous recombination (HR, computed using both ClonalFrameML (CFML) and Geneconv)) and number of CRISPR spacers. The latter were used as proxy of their activity. Data was plotted after excluding very similar ST-1 genomes. The criteria to remove these genomes were based on similarities in R-M content, and gene flux, that is, all ST-1 genomes but CD_020475, CD_020474, CD_021026 were removed ($n=26$). (b) Same as (a) but considering the complete genome dataset ($n=36$). Spearman's rank correlation coefficients (ρ) and associated P values (two-sided) are shown in each graph.



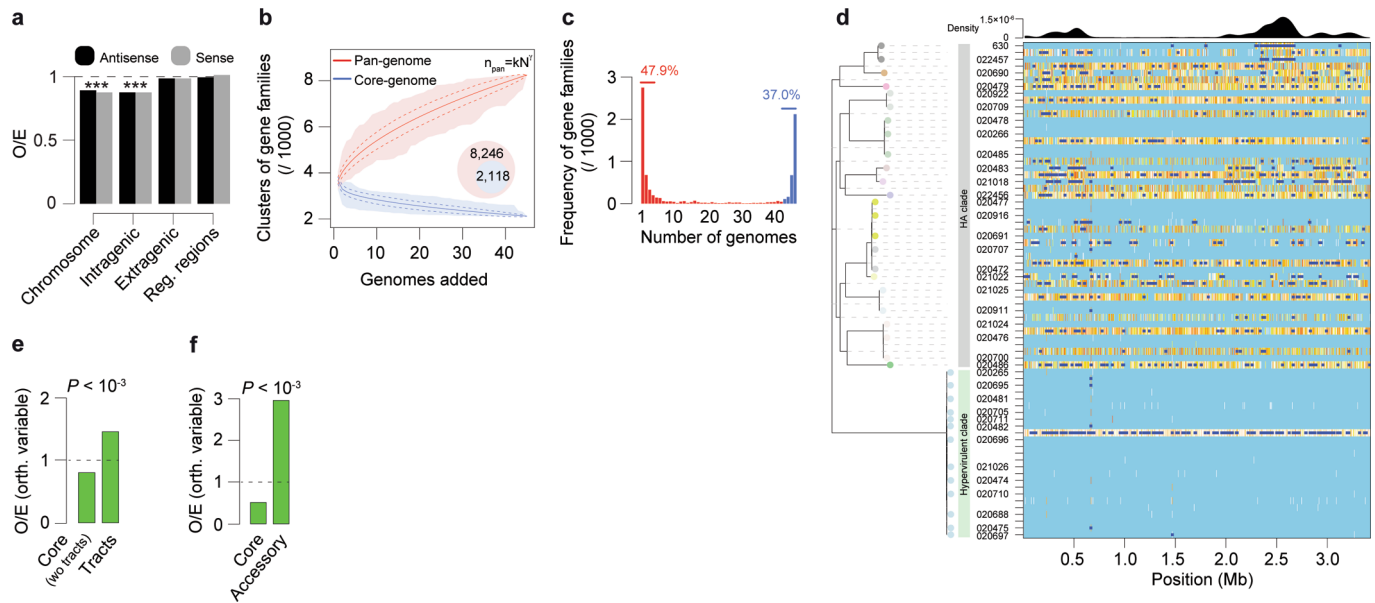
Extended Data Fig. 3 | Interplay between Type I R-M systems and gene flux in *C. difficile*. (a) Observed/expected (O/E) ratios for Type I target recognition motifs in *Clostridioides* phage genomes. 6 phage genomes representative of *Siphoviridae* and *Myoviridae* families and tail types were analyzed (ϕ CD111, ϕ CDHM11, ϕ MMP01, ϕ MMP04, ϕ C2, ϕ CD38). O/E values were obtained with R'MES using Markov chain models that take into consideration oligonucleotide composition. For each motif, we tested if the median value of the O/E ratio in phage genomes was significantly different from 1. In box plots, the middle line indicates the median value, boxes are 25th and 75th quartiles, and whiskers indicate 1.5 times the interquartile range. *** $P < 10^{-3}$; ** $P < 10^{-2}$ (one-sided one-sample t-test). (b) Relation between HGT and O/E ratio for Type I target recognition motifs. For those *C. difficile* genomes harboring a single Type I R-M system (that is, without the confounding effect of multiple systems), we computed the average values of HGT, and plotted these values against the average O/E ratio for the corresponding target recognition motif in phage genomes. This was only possible for the $n = 6$ motifs indicated in brackets. The spearman's rank correlation coefficient (ρ) and associated P value (two-sided) is shown.



Extended Data Fig. 4 | Genomic context and conservation of *camA*. (a) *CamA* protein alignment among *Clostridiales* (*C. mangentii* LM2 (587 aa, 56% identity), *C. sordellii* (598 aa, 53% identity), *C. bifementans* WYM (579 aa, 53% identity), *C. dakarensis* sp. nov (580 aa, 63% identity), *Peptostreptococcaceae bacterium* VA2) and *Fusobacteriales* (*Psychrobacter atlanticus* DSM 19335) using ClustalX. The nine conserved motifs (I–VIII and X) typically found in MTases are highlighted. (b) Phylogenetic tree obtained from the MTase alignment. (c) Phylogenetic tree of the 36 *C. difficile* strains colored by clade (hypervirulent, human/animal (HA) associated) and MLST sequence type (ST). Shown is the genomic context of *camA* across the entire dataset. (d) Expanded view of the region shown in Fig. 1f. The example shown (including coordinates) refers to the reference genome of *C. difficile* 630. + and – signs correspond to the sense and antisense strands respectively. Vertical bars correspond to the distribution of the CAAAAA motif.

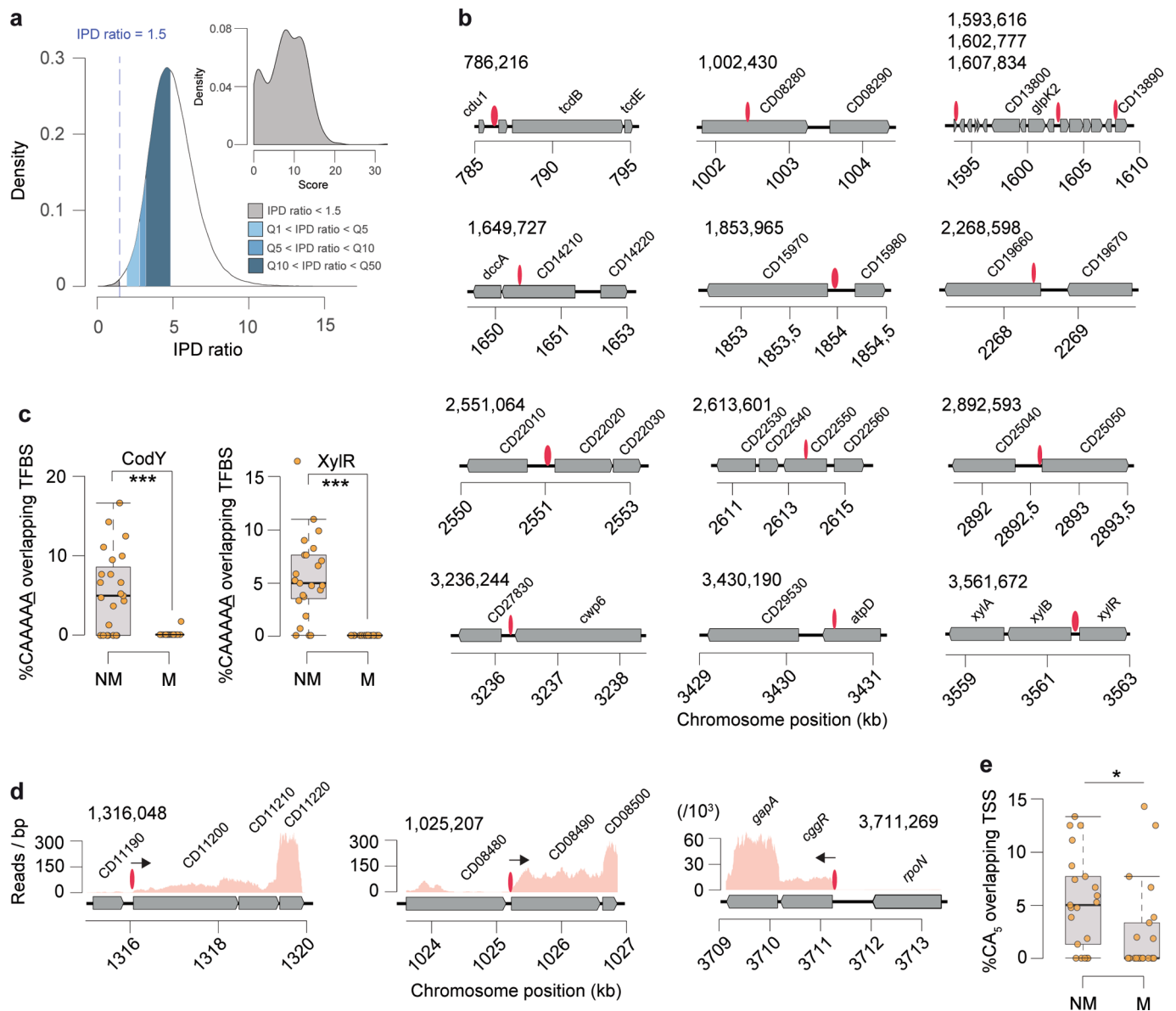


Extended Data Fig. 5 | $\Delta camA$ construction, purified spore analyses, broth culture growth, and sporulation kinetics. (a) PCR to distinguish between wild-type *camA* and $\Delta camA$ using flanking primers and primers internal to the deletion. PCRs were performed twice independently. (b) Growth curves comparing wild-type *camA*, $\Delta camA$, $\Delta camA$ -C, and *camA*/N165A cultures grown in BHIS liquid media. Early stationary-phase cultures were diluted to a starting O.D. of 0.05 in BHIS media and growth was measured over 9 h. Each pair genotype / timepoint correspond to mean of $n=3$ independent biological replicates. Error bars correspond to standard deviation. (c) Phase-contrast microscopy analyses of sporulating culture samples prior to and after spore purification on a density gradient. No gross differences in spore morphology were observed between wild type and the MTase mutant. The germination efficiency (G.E.) of purified spores from the indicated strains is shown below. Scale bar represents 5 μ m. Microscopy analyses were performed on three independent spore preparations. (d) Chloroform resistance of purified $\Delta camA$ spores relative to wild type. Spores were treated with 10 % chloroform for 15 min after which spore viability was measured by plating untreated and chloroform-treated spores on media containing germinant and measuring colony forming units. No significant differences in germination efficiency or chloroform resistance were observed. Data are presented as mean \pm standard deviation of four independent biological replicates. (e) Heat-resistance (H_{RES}) efficiencies of sporulating cultures 22 h after sporulation induction were determined relative to wild-type. Data are presented as mean \pm standard deviation. Three independent biological replicates per group were used. *** $P < 10^{-3}$, one-way ANOVA with Tukey's test.

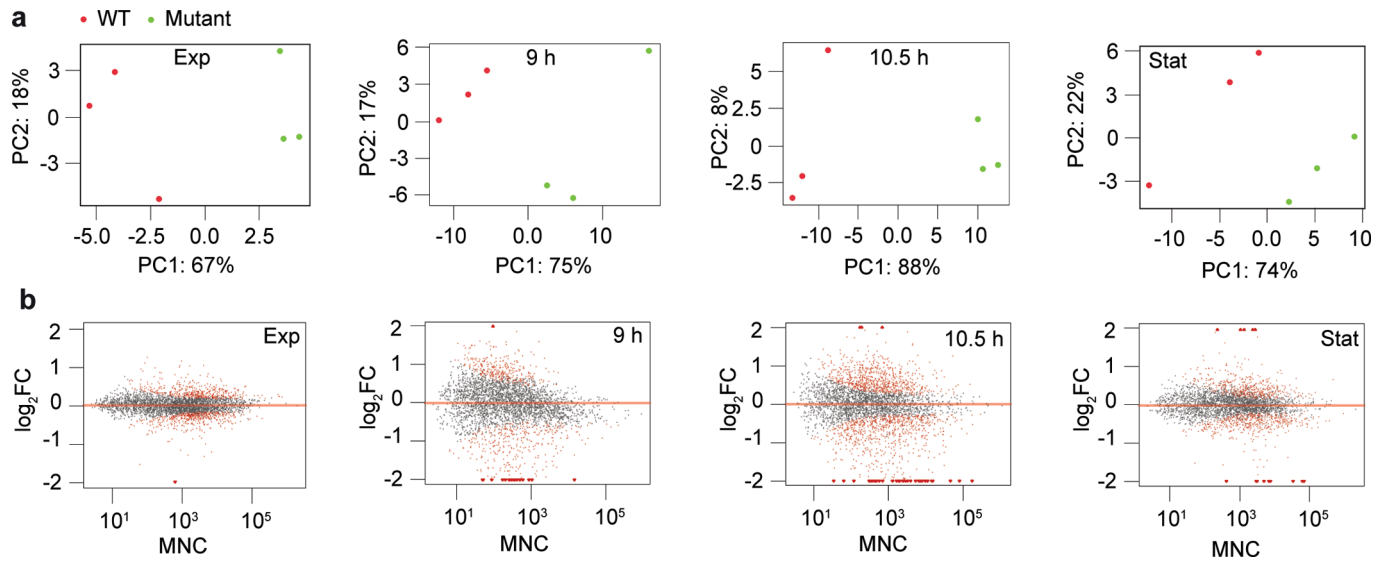


Extended Data Fig. 6 | CAAAAA exceptionality, core- / pan-genome analyses of *C. difficile*, and homologous recombination (HR) landscape. (a)

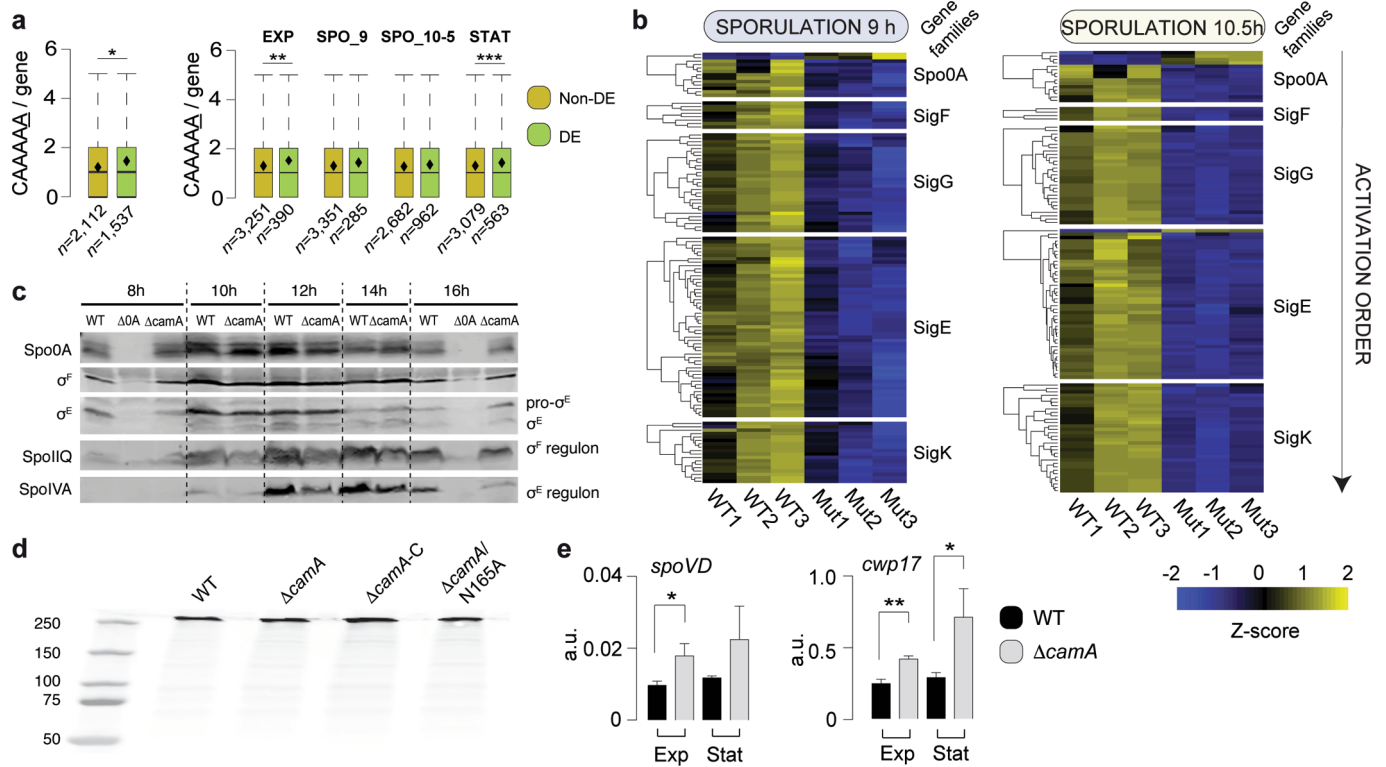
Observed (O) numbers of CAAAAA motifs in the *C. difficile* chromosome ($n = 7,824$), intragenic ($n = 6,131$), extragenic ($n = 1,693$), and regulatory regions ($n = 794$, defined as the windows spanning 100 bp upstream the start codon to 50 bp downstream) were compared with expected (E) values computed in random sequences showing the same oligonucleotide composition. The significance of the difference between O/E was evaluated by computing a P value based on a Gaussian approximation of motif counts under a Markov model of order 4 ($*** P < 10^{-3}$). (b) Core- and pan-genome sizes of *C. difficile*. The pan- and core-genomes were used to perform gene accumulation curves. These curves describe the number of new genes (pan-genome) and genes in common (core-genome) obtained by adding a new genome to a previous set. The procedure was repeated 1,000 times by randomly modifying the order of integration of the $n = 45$ genomes in the analysis. Solid lines correspond to the average number of gene families obtained across all permutations, dashed lines indicate standard deviation of the mean, and shaded regions indicate range. The values for the specific constants obtained after Heap's law fitting are 2,887 and 0.271, respectively for the k and γ , thus implying an open pan-genome. (c) Spectrum of frequencies for *C. difficile* gene repertoires. It represents the number of genomes where the families of the pan-genome can be found, from 1 for strain-specific genes to 45 for core-genes. Red indicates accessory genes and blue the genes that are highly persistent in *C. difficile*. (d) Graphical representation of the recombinational events in the core genome of *C. difficile* (inferred by ClonalFrameML). The HA and hypervirulent branches of the tree are depicted in colors. Substitutions are represented by vertical lines and recombination events by dark blue horizontal bars. Light blue vertical lines represent the absence of substitutions, and white lines refer to non-homoplasic substitutions. All other colors represent homoplasic substitutions, with increases in homoplasy associated with increases in the degree of redness (from white to red). (e) O/E ratios of orthologous variable CAAAAA motifs (compared to orthologous conserved) in the core-genome (excluding recombination tracts) ($n = 770$) and recombination tracts ($n = 325$), or (f) core ($n = 1,095$) and accessory genome ($n = 1,415$). P values correspond to the Chi-square test.



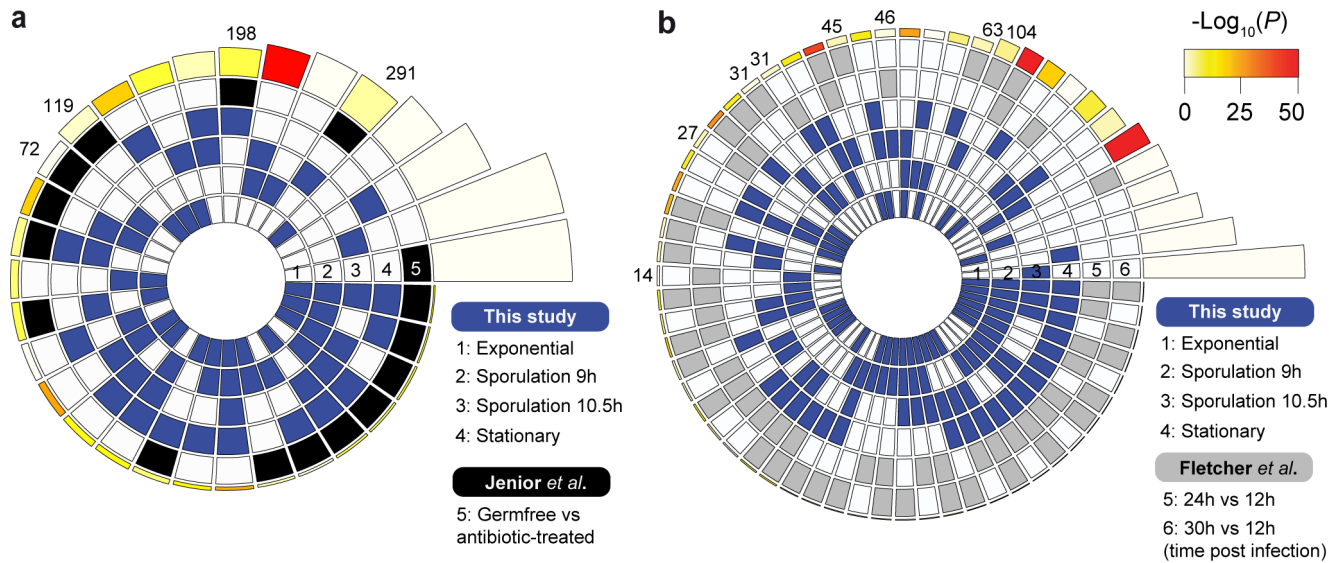
Extended Data Fig. 7 | Non-methylated CAAAAA motif sites overlapping TFBSs and TSSs. (a) Interpulse duration ratio (ipdR) density distribution of the terminal adenine of CAAAAA. Motifs were considered as non-methylated if the terminal adenine had IPD ratios < 1.5 (stippled line), coverage > 20 \times , and methylation scores < 20 (gray distribution). Also shown for comparison are the sections delimited by quantiles (Q) 1, 5, 10, and 50. (b) Additional examples of highly conserved non-methylated CAAAAA motif sites (red ovals) and corresponding genetic context. Positions indicated above the graph correspond to the non-methylated base. (c) %CAAAA motif sites (non-methylated (NM) and methylated (M)) overlapping CodY and XylR TFBS for each *C. difficile* isolate excluding ST1 genomes ($n = 23$). (d) Additional examples of chromosomal regions for which non-methylated CAAAAA motif sites overlap TSSs (shown as arrows). (e) %CAAAA motif sites (non-methylated and methylated) overlapping TSSs for each *C. difficile* isolate excluding ST1 genomes ($n = 23$). For box plots the middle line indicates the median value, boxes are 25th and 75th quartiles, and whiskers indicate 1.5 times the interquartile range. * $P < 0.05$, *** $P < 10^{-3}$ (one-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction).



Extended Data Fig. 8 | Principal Component Analysis (PCA) and MA-plots for RNA-seq data. (a) PCA performed using DESeq2 rlog-normalized RNA-seq data ($n=3$ biological replicates for each genotype). (b) MA-plots showing the variation of fold change with mean normalized counts (MNC). Number of genes represented: 3,532 (Exp), 3,426 (9 h), 3,523 (10.5), and 3,510 (Stat). Red-colored points have P values < 0.1 (Wald test, Benjamini-Hochberg adjusted). Points that fall out of the window are plotted as open triangles pointing either up or down.



Extended Data Fig. 9 | DE, gene, and protein expression analyses. (a) Enrichment of the CAAAAA motif in DE genes compared to non-DE ones either globally (left, $n = 3,649$ genes) or at each time point studied (right, $n_{\text{EXP}} = 3,641$, $n_{\text{SPO}_9} = 3,636$, $n_{\text{SPO}_{10.5}} = 3,644$, $n_{\text{STAT}} = 3,642$). For box plots, the middle line indicates the median value, boxes are 25th and 75th quartiles, and whiskers indicate 1.5 times the interquartile range. * $P < 0.05$, ** $P < 10^{-2}$, *** $P < 10^{-3}$ (one-sided Mann-Whitney-Wilcoxon rank sum test with continuity correction). (b) Time-course change in the expression of genes under the control of the specific sigma factors (σ^f , σ^E , σ^G , and σ^K) and master transcriptional activator Spo0A at both 9 and 10.5 h after sporulation induction (respectively $n = 121$ and $n = 124$ genes). (c) Representative immunoblot time-course (from $n = 2$ independent biological replicates with similar results) comparing the levels of the early sporulation proteins σ^f , SpoIIQ, σ^E , and SpoIVA in WT and Δ camA at 8, 10, 12, 14, and 16 h following induction of sporulation. (d) Western blot for TcdA for each *C. difficile* genotype. (e) RT-qPCR of *spoVD* and *cwp17* genes ($n = 3$ independent biological replicates) of exponential and stationary phase liquid broth cultures. Data is presented as mean \pm standard deviation. * $P < 0.05$, ** $P < 10^{-2}$, two-tailed unpaired Student's t-test.



Extended Data Fig. 10 | Overlap between multiple datasets of differentially expressed (DE) genes. Comparisons were performed between DE genes called in this study for each time point (blue-shaded, $n=1,537$) and those obtained from (a) Jenior *et al.* (black-shaded, $n=971$) and (b) Fletcher *et al.* (gray-shaded, 299). Color intensities of the outermost layer represent the P value significance of the intersections (3,896 genes were used as background). The height of the corresponding bars is proportional to the number of common genes in the intersection (indicated at the top of the bars for pairwise comparisons between the different studies). Significant overlaps were found between our DE dataset and either (a) genes DE during infection in different mice gut microbiome compositions ($P < 10^{-6}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted), or (b) DE genes obtained from mice gut isolates at increasing time points after infection ($P < 10^{-4}$, one-tailed hypergeometric test implemented in *SuperExactTest*, Bonferroni adjusted).

Opinion

Conserved DNA Methyltransferases: A Window into Fundamental Mechanisms of Epigenetic Regulation in Bacteria

Pedro H. Oliveira ^{1,*} and Gang Fang^{1,*}

An increasing number of studies have reported that bacterial DNA methylation has important functions beyond the roles in restriction-modification systems, including the ability of affecting clinically relevant phenotypes such as virulence, host colonization, sporulation, biofilm formation, among others. Although insightful, such studies have a largely *ad hoc* nature and would benefit from a systematic strategy enabling a joint functional characterization of bacterial methylomes by the microbiology community. In this opinion article, we propose that highly conserved DNA methyltransferases (MTases) represent a unique opportunity for bacterial epigenomic studies. These MTases are rather common in bacteria, span various taxonomic scales, and are present in multiple human pathogens. Apart from well-characterized core DNA MTases, like those from *Vibrio cholerae*, *Salmonella enterica*, *Clostridioides difficile*, or *Streptococcus pyogenes*, multiple highly conserved DNA MTases are also found in numerous human pathogens, including those belonging to the genera *Burkholderia* and *Acinetobacter*. We discuss why and how these MTases can be prioritized to enable a community-wide, integrative approach for functional epigenomic studies. Ultimately, we discuss how some highly conserved DNA MTases may emerge as promising targets for the development of novel epigenetic inhibitors for biomedical applications.

Introduction

The information content of DNA is not limited to that contained within the primary nucleotide sequence. Instead, significant meaning can also be conveyed through the epigenetic states of DNA (e.g., by chemical modification such as DNA methylation). In bacteria, DNA methylation is typically associated with **restriction-modification (R-M) systems** (see [Glossary](#)), which operate as key moderators of the flow of genetic information between cells by horizontal gene transfer (HGT) [1,2]. R-M systems typically encode a **DNA methyltransferase** (MTase) that modifies particular DNA sequences in function of the presence of target recognition sites and a restriction endonuclease (REase) that cleaves them when they are unmethylated [3] ([Box 1](#)). Some DNA MTases, known as solitary or orphan, were also identified as apparently lacking a cognate REase [4]. DNA methylation performed either by R-M or orphan MTases were properly discussed in a few seminal works [5–7].

The bacterial genome has three major forms of DNA methylation: *N*6-methyladenine (6mA), *N*4-methylcytosine (4mC), and 5-methylcytosine (5mC), with 6mA being the most prevalent form. While 5mC may be detected with bisulfite sequencing, 6mA and 4mC events have been challenging to map at the genome-wide scale [8], limiting the comprehensive study of bacterial **epigenomes**. The study of bacterial **methylomes** entered a new era in 2012 when a new

Highlights

DNA methylation is the epigenetic mark most commonly found throughout the living world. In bacteria, it is responsible for a variety of functional roles, including defense against foreign DNA, regulation of chromosome replication and segregation, mismatch repair, and control of virulence gene expression, among others.

DNA methyltransferases (MTases) are responsible for transferring a methyl group from an *S*-adenosyl-*L*-methionine (AdoMet) donor to DNA. Dam, Dcm, and CcrM are examples of bacterial DNA MTases that have been comprehensively characterized for their roles in gene regulation.

Here, we summarized the landscape of DNA MTase conservation in bacteria and observed that MTase conservation is more common than previously portrayed, spanning several phylogenetic levels, and being present in multiple human and animal pathogens. Information on the functional relevance of these MTases is virtually inexistent, but they are expected to play key functional roles.

We also discuss why and how these MTases can be prioritized to enable a community-wide, integrative approach for functional epigenomic studies. Ultimately, we discuss how some highly

¹Department of Genetics and Genomic Sciences, Institute for Genomics and Multiscale Biology, Mount Sinai School of Medicine, New York, NY, USA

*Correspondence: pcphco@gmail.com (P.H. Oliveira) and fanggang@gmail.com (G. Fang).



Box 1. Restriction-Modification Types and Key Definitions on Gene Persistence and Essentiality**Restriction-Modification (R-M) Types**

The three classical types of R-M systems differ in their molecular structure, sequence recognition, cleavage position, and cofactor requirements [6,101–103]. Type I systems are complex hetero-oligomers either comprising one DNA sequence specificity (S), two REase and two MTase subunits with restriction and modification activities, or two MTase and one S subunits with modification activity only. Type II systems encoded on separate genes are composed of one homodimeric or homotetrameric REase and one monomeric MTase and in most cases are able to operate separately and independently from each other at least *in vitro*. Some type II systems, particularly types IIB, IIG, IIL, and some IIH (collectively termed IIC) encode both restriction and modification domains within the same protein. Type III systems are heterotrimers or heterotetramers of products of two genes, *res* and *mod*, involved in restriction and modification, respectively. Both subunits are required for restriction, whereas Mod is sufficient to produce a modification. Finally, type IV 'restriction systems', as opposed to R-M systems, are composed of one or two REases that cleave modified recognition sites.

Core Genes

Genes common to all genomes in a phylogenetically coherent group. They should contain the essential genes particular to that group as well as some nonessential ones.

Essential Genes

Typically involved in basic cellular processes such as translation, transcription, and replication. The concept of essentiality is not an intrinsic property of a gene, but instead a function of genetic and environmental factors. Essential genes can be essential in one species but not another, or under a defined growth condition but not in others.

Persistent Genes

Conserved above a predefined cutoff threshold of bacterial genomes. Although somewhat arbitrary, such threshold should take into consideration certain criteria, such as phylogenetic relatedness between organisms and gene organization within genomes. By definition, persistent genes include core genes.

conserved DNA MTases may emerge as promising targets for the development of novel epigenetic inhibitors for biomedical applications.

technology called **single molecule real-time (SMRT) sequencing** [9] enabled the detection of all three major forms of bacterial DNA methylation. Since then, >2470 (as of 05/2020) bacterial and archaeal methylomes [10,11] have been determined at a quasi-exponential pace. Propelled by SMRT sequencing, an increasing number of studies documented the involvement of DNA methylation in often critical aspects of cell biology. Some examples include gene expression changes affecting cell motility [11], sporulation [12], virulence [13,14], and in providing structural support for bacterial survival during antibiotic stress [15].

Previous bacterial epigenome studies have a largely *ad hoc* nature, in that most have performed methylome mapping in one or few strains of the same species and, less frequently, across multiple species. A systematic examination of MTases across a large number of strains in a single species was only determined in few occasions and in an even lower number of studies were MTase mutants constructed for phenotypic and molecular characterization [11,16–19]. Such studies are insightful as they provide a comprehensive snapshot of MTase diversity and some have been indeed capable of linking individual MTases to specific functions in the cell. But they face major challenges. For example, it is usually difficult for one single study to obtain sufficiently deep mechanistic insight, or comprehensively uncover phenotypes impacted by the loss of an MTase. It is also conceptually challenging to integrate epigenomic information stemming from different studies dealing with MTases present in few strains. More importantly, there have been limited attempts to identify specific methylation sites and mechanisms, underlying the epigenetic regulation of genes linked to defined phenotypes. Due to these limitations, some fundamental questions still remain unanswered: What phenotypes (in a particular species) are impacted by DNA MTases? Which specific methylation sites play important regulatory roles? What are

the underlying epigenetic mechanisms regulating cellular phenotypes by specific methylation events?

Among all the diversity of DNA MTases in bacteria [10,20], some are highly conserved at the species level or at higher taxonomic ranks. Examples of well characterized ones include the *Escherichia coli* Dam enzyme (methylating at 5'-GATC-3') and the *Caulobacter crescentus* CcrM enzyme (methylating at 5'-GAN_TC-3'). Dam and CcrM homologs are widespread in γ - and α -Proteobacteria, respectively [21]. Both are encoded by core genes [22,23] (Box 1) and recognized as conditionally essential for the viability of several species [19,24–26], typically via mutation or overexpression approaches coupled to gene expression profile analyses. We recently witnessed a surge of studies focusing on less known conserved MTases belonging to different R-M types and operating the three major forms of DNA methylation [12,14,27–29]. Despite multiple evidence suggesting that R-M genes are frequently exchanged between species [30,31,45], and evolve very quickly [32,33], the above-mentioned examples illustrate how certain MTases may endure strong selective pressure for retention in genomes. Several possibilities may account for such retention, including the involvement in epigenetic regulation of functionally relevant genes [12,27], the ability of certain selfish R-M systems to induce postsegregational killing [7], or in shaping gene flux and host genome composition [34]. Whether Dam, CcrM, and the few other recent examples are merely outliers, or actually representatives of a broader set of conserved MTases (and eventually full R-Ms), is currently not clear.

In this opinion article, we summarize the landscape of DNA MTase conservation in the bacterial kingdom. We observed that MTase conservation is more common than previously portrayed, spanning multiple phylogenetic levels, and being present in multiple human pathogens. We then propose that prioritizing conserved MTases can facilitate community-wide efforts for integrating experimental and multiple omics data (e.g., genomic, transcriptomic, epigenomic) to more effectively address the fundamental questions laid earlier. Ultimately, we discuss how some of these targets may emerge as promising targets for the development of novel epigenetic inhibitors.

Conserved DNA MTases Are Abundant in Bacteria

A total of 26 582 MTases are found in 5568 complete bacterial genomes available in GenBank (considering only species with at least ten complete genomes) (Figure 1A, Tables S1–S3 in the supplemental information online). Type II MTases are present at the highest densities, in what is likely a consequence of type II R-M systems' ability to induce genetic addiction. Conversely, types IIC and III are the least abundant. A total of 52% of the species harbor persistent MTases (here defined as those conserved in at least 80% of each species' genomes) (Box 1, Figure 1B, Table S3 in the supplemental information online). The frequency of persistent MTases varies widely among large bacterial phyla and is unrelated to the density of total MTases (Figure 1A). For example, α -Proteobacteria harbor multiple persistent MTases, but show an overall low density of total MTases. However, phyla such as Fusobacteria and Chloroflexi are devoid of persistent MTases, but are rich in other MTases (Figure 1A). In 27% of the species, more than one persistent MTase is present (either belonging to the same or different types) (Tables S3 and S4 in the supplemental information online). A total of 36% of the species harbor MTases that are consistently present across all genomes (core), the majority being of type II (Figure 1C). These core MTases represent 8.5% from the total MTase dataset. The human obligate pathogen *Neisseria gonorrhoeae* stands out as the species harboring the most profuse arsenal of persistent/core MTases ($n = 10$) spanning types I and II/IIC. Since we have only included bacterial species with at least ten complete genomes available at GenBank, we expect our estimate on the number and diversity of core/persistent MTases to increase in the future as more genomes get

Glossary

DNA methyltransferase: family of enzymes that catalyze the transfer of a methyl group from an S-adenosyl-L-methionine (AdoMet) donor to DNA.

Epigenome: complete record of all chemical modifications to DNA.

Together with the epitranscriptome (chemical modifications of RNA) and epiproteome (chemical modifications of proteins), makes up the epi-ome.

Methylome: complete record of all methyl modifications to either DNA, RNA, or proteins in a particular cell or organism.

Restriction-modification (R-M) systems: almost ubiquitous in prokaryotes, these systems consist of a DNA methyltransferase that methylates a specific target sequence in the host genome and a cognate restriction endonuclease that cleaves unmethylated or inappropriately methylated targets from exogenous DNA. They are thus typically regarded as innate defense systems and, depending on type, as molecular parasites.

Single molecule real-time (SMRT) sequencing: third generation long-read sequencing-by-synthesis technology, based on the real-time imaging of fluorescently tagged nucleotides as they are synthesized along individual DNA template molecules. The duration between consecutive pulses of light directly reflects the DNA polymerase kinetics, including the impact caused by DNA modification events.

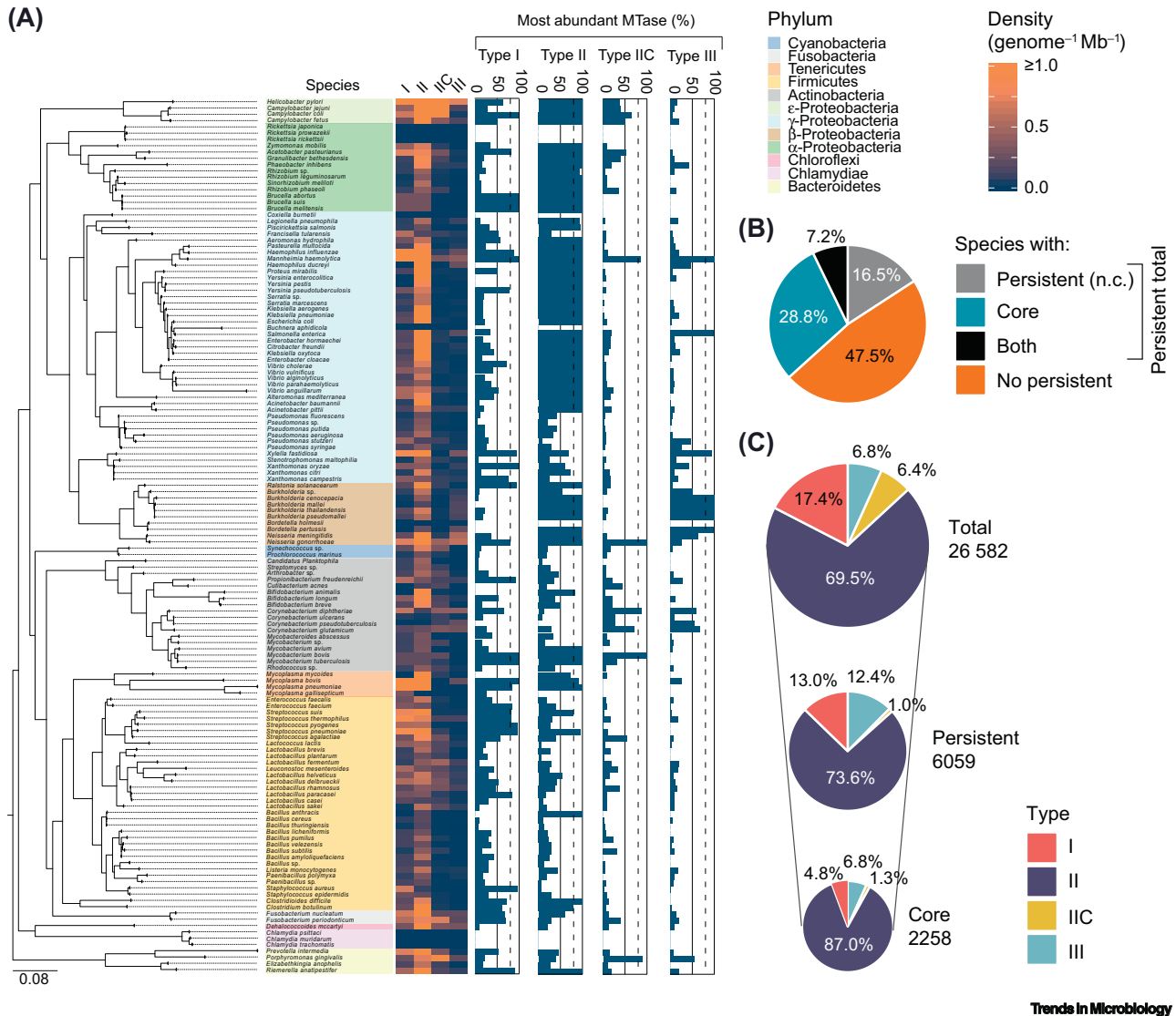


Figure 1. Summary of Methyltransferase (MTase) Conservation in Bacterial Genomes from GenBank. (A) Phylogenetic tree of the 139 bacterial species (colored by phylum), for which at least ten complete genomes were available at GenBank (corresponding to a total of 5568 genomes). Heatmap corresponds to the density (per genome per Mb) of types I, II, III MTases and type IIC restriction-modification (R-M) systems for each species. Bar plots indicate the percentage of the most abundant MTase(s) found in each species, assuming as inclusion criteria a minimum of 80% similarity in amino acid sequence and less than 20% difference in protein length. Stippled lines indicate a threshold of 80%, above which an MTase can be considered persistent. A core gene is denoted by 100%. (B) Pie-chart summarizing the percentages of species analyzed containing either persistent non-core (n.c.) MTases, core MTases, both, or none. (C) Pie-charts showing the breakdown of total, persistent, and core MTases per type.

sequenced and novel MTases are found. On top of this, there is the possibility that small non-canonical MTases may have gone unnoticed, as recently pinpointed in a large-scale analysis in human microbiomes [35]. Certain core/persistent MTase genes may also undergo structural variations (e.g., at the level of the target recognition domain) capable of changing their recognition motif or rendering their products inactive in some genomes, while still being subtle enough to be classified in the same gene family. This is the case of, for example, the persistent type II MTase from *Mycobacterium tuberculosis* recognizing CTGGAG. Hence, core and persistent MTases are abundant in bacteria.

Core and Persistent DNA MTases Differ Substantially in Their Organization and Sequence Recognition

We next summarize the diversity of core and persistent MTases in terms of their organization (orphan versus part of an R-M system) and target sequence recognition (see the supplemental information online). Across strains of the same species, MTases are found predominantly organized as part of complete R-M systems or as orphans, but less frequently as both (Figure 2). This suggests that for orphan MTases, loss of the cognate REase likely occurred early in the evolutionary history of these species. Alternatively, orphan MTases may have been acquired as such by HGT

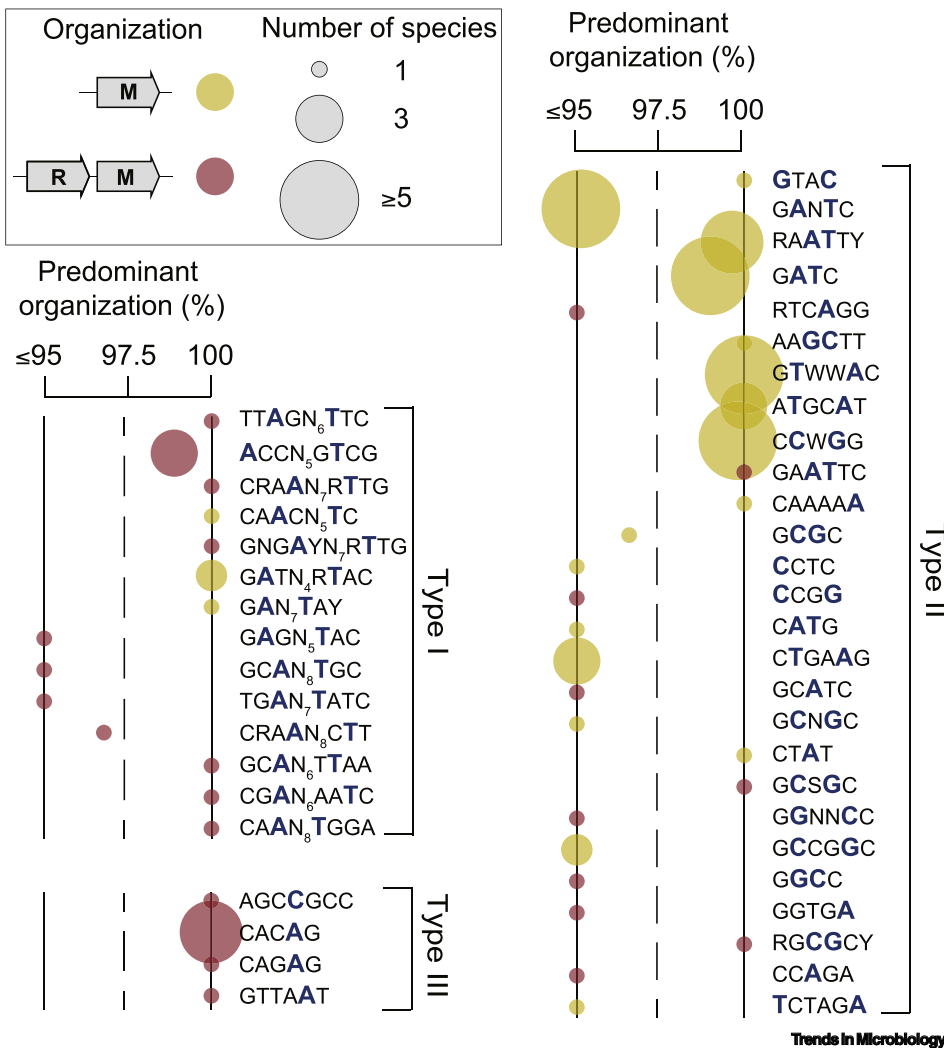


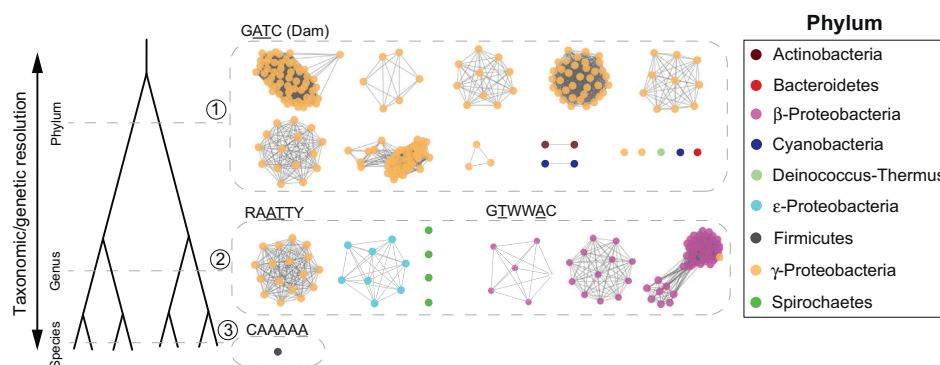
Figure 2. Summary of the Organization and Target Recognition Motifs of Persistent Methyltransferases (MTases) Based on the REBASE Database. Yellow circles represent solitary MTases, whereas red ones represent complete systems. Predominant organization of 100% means that the MTase is always found either as solitary (without a cognate endonuclease) or as part of a complete restriction-modification (R-M) system. Values below 100% indicate that both organizations are present, with the most predominant one highlighted. For example, type II GTWWAC-recognizing MTases are exclusively solitary, whereas type III CACAG-recognizing MTases are exclusively found within complete R-M systems. GATC-recognizing MTases are found as solitary in 98.9% of the species analyzed and the remaining 1.1% in complete systems. Target recognition motifs shown are based on the REBASE database. Circle radius is proportional to the number of species in which the MTase is present.

and further kept under strong selective pressure [20]. The existence of multiple core and persistent complete R-M systems suggests alternative roles such as gene expression regulation or subversion of host genome integrity during infection. For example, although it is not entirely clear if all R-M systems are active in *N. gonorrhoeae*, at least some of its type II REases are known to be released in an active way during infection of host cells and to enter the nucleus through nuclear pores, inducing double strand breaks in DNA during mitosis [36].

Persistent MTases are also very diverse in terms of sequence recognition (Figure 2, Table S4 in the supplemental information online). We observed a total of 48 different methylation motifs belonging to the three major R-M types, among which 73% methylate at 6mA. These observations are expected to be conservative as they correspond solely to MTases whose recognition sequence has been confirmed by SMRT sequencing [37]. As expected, the MTases for which more functional studies have been published [38] (namely Dam, Dcm, CcrM) also correspond to those most widespread across a higher number of species (Figure 2). Hence, core and persistent MTases are diverse in terms of organization and target recognition sequence.

Core and Persistent DNA MTases Are Found at Multiple Taxonomic Scales

Genes can differ significantly in their taxonomic distributions, with more broadly conserved genes having 'housekeeping' functions and less conserved genes being responsible for the phenotypic differences observed between organisms. In this regard, persistent genes can be restricted to any taxonomic level (e.g., domain-, family-, genus-, species-, or strain-specific). Once persistent genes have been defined to identify a related group of organisms, the biological roles performed by these genes' products can provide insights into functions and phenotypes that may be characteristic (and even critical) to those groups. One example, is that of Dam MTase, conserved in a large subset of γ -Proteobacteria (Figure 3), including the clinically relevant genera *Escherichia*, *Salmonella*, *Vibrio*, and *Yersinia*. Its acquisition might have been the key evolutionary moment that created a new mechanism capable of DNA strand discrimination based on the hemimethylated state of newly replicated DNA [39]. Such mechanism is critical for the regulation of multiple cellular processes. For example, during DNA mismatch repair in *E. coli*, the MutH protein recognizes hemi-methylated DNA and cuts the nonmethylated daughter strand, ensuring that the methylated parental strand will be used as template for repair-associated DNA synthesis [40]. In



Trends in Microbiology

Figure 3. Illustrative Examples of Sequence Similarity Networks of Persistent Methyltransferases (MTases) Conserved at Different Taxonomic Resolutions. See Table S5 in the supplemental information online. Each node represents one protein. To avoid redundancy and improve visualization, only one genome per species is shown (typically the reference/representative genome). Edges correspond to pairwise protein sequence identity >60%. Node colors correspond to different phyla.

addition, hemi-methylated GATC sites can activate gene expression upon passage of the replication fork [41,42] and coordinate the initiation of replication within cell cycle in *E. coli* [43].

Closer to the genus level conservation, we can highlight, as illustrative examples, those of RAATTY- and GTWWAC-recognizing MTases. The former are pervasive in the genera *Acinetobacter* and *Campylobacter*, while the latter are often found in *Burkholderia*. Information on the functional relevance of these MTases is virtually inexistent, but they are expected to play specific roles that help maintain the identity of these genera. In line with this hypothesis is the recent observation that RAATTY methylation is required for efficient transformation in *Campylobacter jejuni* [44]. Genes mainly preserved at the species and strain level are also of interest, as they may be involved in exclusive ecological adaptations to particular niches. One example is that of CamA, a 6mA persistent MTase recognizing CAAAAA involved in the sporulation and biofilm formation in *Clostridioides difficile* [12] (Figure 3) and also the most species-specific persistent MTase currently known.

The acquisition of a new functional R-M system by a bacterial clone may significantly reduce its ability of engaging in genetic exchanges with conspecific bacteria [45]. This may help carving preferential routes of DNA exchange between its offspring (which inherited this R-M system), favor the maintenance of cohesive population structures, and eventually give rise to a new lineage in the population [46]. Specific lineages of important pathogens that have recently changed their R-M repertoires and show higher sexual isolation include *Burkholderia pseudomallei*, *E. coli*, *Neisseria meningitidis*, *Staphylococcus aureus*, and *Streptococcus pneumoniae* [47–50]. A Type I R-M system for example, decreased transfer to and from a major methicillin-resistant *S. aureus* lineage [51]. Hence, core and persistent MTases are found at multiple taxonomic scales, where they are expected to play roles that help shape phylogenetic structure.

Core and Persistent MTases as an Opportunity for Integrative Studies of Bacterial Epigenomes

Persistent genes, as orthologs shared by all (or almost all) members of an evolutionarily coherent group, likely reflect the important functions positively selected over time [52,53]. They are also more likely to facilitate standardization and extrapolation from well-studied bacterial strains to newly sequenced ones using systems-level approaches, rendering possible direct comparisons of findings from different laboratories (Figure 4). In this regard, core and persistent MTases appear as particularly attractive targets to be prioritized in bacterial epigenomic studies, as they allow the integration and analysis of multidimensional omics data to retrieve meaningful information from bacterial epigenomes and to ultimately address the questions laid down in the introductory section.

How Can Phenotypes That Are Impacted by DNA MTases Be Identified?

Our understanding of the genetic mechanisms that underlie biological processes has relied extensively on loss-of-function approaches that reduce or ablate gene function. Through the analysis of the phenotypes caused by such perturbations, one can elucidate the wild type function of a given gene. For example, nontargeted DNA mutagenesis approaches, such as large scale random transposon insertion mutagenesis coupled with deep sequencing (TIS), have become powerful tools to simultaneously assess the essentiality of genes under defined experimental conditions and to rapidly connect genotype to phenotype in a wide range of bacteria [54]. Several variants of TIS have been independently developed [55–58] and applied to a variety of bacteria, allowing assessment of the role of certain DNA MTases as controllers of critical cellular processes [59] and/or as conditionally essential genes [60–62]. Relevant functional information has also been obtained by targeted mutagenesis or overexpression of DNA MTases [11,16–19]. A

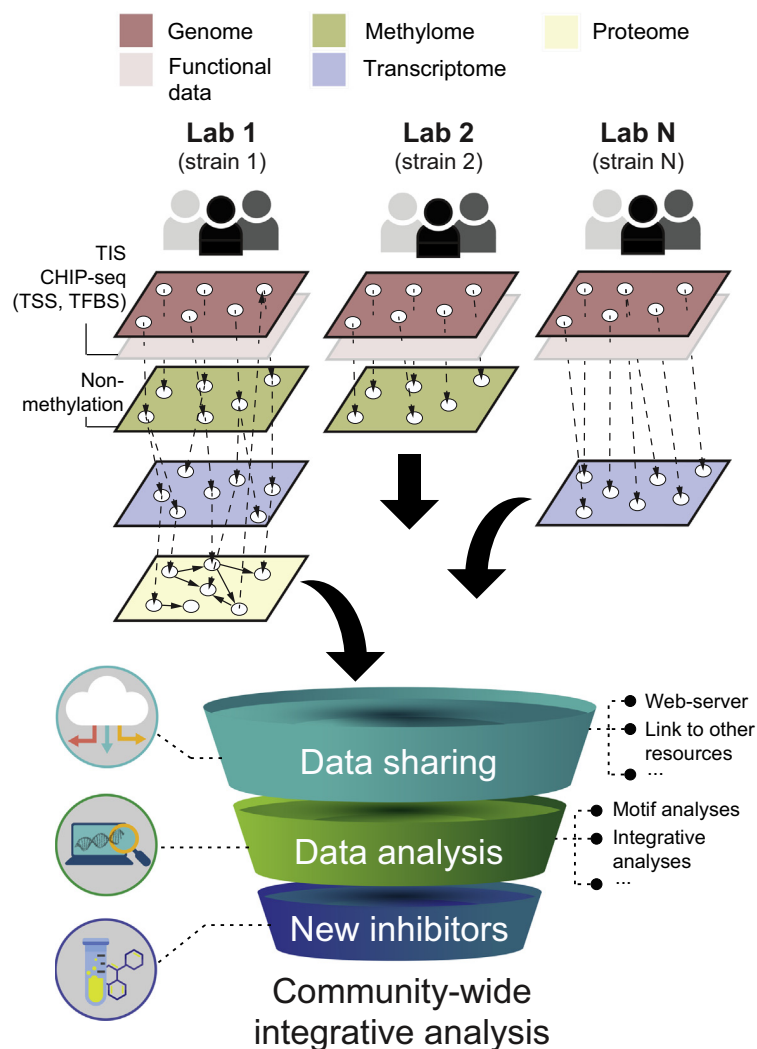


Figure 4. Overview of a Large-Scale Community-Wide Integrative Approach for Bacterial Methylome Analyses. Core and persistent methyltransferases (MTases) can be prioritized to build a trans-omic network across multiple laboratories merging multiple functional data gathered at different experimental conditions. The latter may build upon MTase mutants generated by transposon insertion mutagenesis coupled with deep sequencing (TIS), site-directed mutagenesis of methylation sites, genome-wide profiling of DNA binding proteins [chromatin immunoprecipitation (ChIP)-seq], transcription start site (TSS) mapping, and identification of methylation-sensitive transcription factors. Multiple layers of omics data may ultimately be commonly shared, linked to other resources (e.g., REBASE), allow for an in-depth analysis of, for example, methylation motif conservation, phase-variable DNA methyltransferases, and accelerate the research of novel epigenomic inhibitors. Abbreviations: TFBS, Transcription factor binding site.

comprehensive global transcriptome and functional profiling by RNA-seq offers the opportunity to further dissect the range of differentially expressed genes in a methylation-free strain. Integrative analyses that incorporate RNA-seq data and other omics experiments are also becoming prevalent. For example, pairwise integration of RNA-seq and DNA methylation is typically performed by the analysis of correlation between differentially expressed genes and methylation patterns (e.g., using linear models, logistic regression, or empirical Bayes models), or alternatively, through the identification of sets of genes that have coordinated differential expression and methylation [63]. For an in-depth understanding of the complex relationships between multiple omics sets,

tools such as MultiDataSet [64], CNAMet [65], and SuperExactTest [67] can be used. The latter, for example, has been recently used to aid in the identification of novel functional roles of a bacterial MTase [12].

How Can Specific Methylation Sites Playing Important Regulatory Roles Be Identified?

Another outstanding question concerns the different regulatory roles played by distinctive subsets of methylation sites in a genome. Two approaches, based on intra- and intergenome analyses, can be considered. The former stems from the observation of a positive correlation between the number of methylation sites in a gene and the fold change of expression between wild type and MTase mutants [11,59], suggesting that epigenetic regulation of expression may be driven by multiple methylation sites, particularly in promoter regions. In this case, genomic regions with significant high density of methylation sites should be targeted for site-directed mutagenesis or genetic editing in order to gauge the impact of each methylation site mutation [27,67–69]. A second orthogonal approach, inspired by phylogenetic footprinting, deduces functional relevance based on the degree of conservation of orthologous methylation motifs across multiple genomes. By comparing multiple methylomes associated with a persistent MTase, one can distinguish between strictly conserved orthologous target methylation sites and variable ones (e.g., harboring single nucleotide polymorphisms or indels). While the former are likely to preferentially play housekeeping roles, at least some of the latter are expected to serve as ON/OFF regulators through phase variation. An additional benefit of an orthogonal approach conducted across a substantial number of same-species genomes, is to gain sufficient statistical power to perform a systematic interrogation of nonmethylated motifs sites. Such an approach has recently allowed for a more systematic detection and analysis of both highly conserved and nonmethylated sites in methylomes associated with persistent MTases [12,70].

How Can Epigenetic Mechanisms Regulating Cellular Phenotypes by Methylation Be Identified?

Finally, we are left with the question of the mechanisms of epigenetic regulation. Here, more comprehensive studies will be necessary to fully characterize the precise mechanisms by which DNA methylation modulates gene expression and alters bacterial phenotypes. Such studies would benefit from the integration of methylome information with other assays, such as high-confidence genome-wide transcriptional landscape inference and transcription start site calling [71,72], or mapping of transcription factor binding sites (TFBSs). Our understanding of the latter, for example, has been mainly achieved by means of chromatin immunoprecipitation (ChIP) [73] assays eventually coupled to next-generation sequencing (ChIP-seq) [74]. The growing number of available bacterial epigenomes has not only spurred a surge in comparative epigenomic studies, but also calls for additional integration with fine-resolution TFBS maps, which in bacteria is still limited to a few species, namely *E. coli* [75,76], *Bacillus subtilis* [77], and *M. tuberculosis* [78]. While an alternative strategy would be to use comparative genomics across a large genomic dataset to identify putative TFBSs [12], the generation of additional CHIP-seq data would provide valuable insight and stimulate sharing across laboratories.

Overlaying comprehensive TFBS and methylation maps becomes critical for elucidating complex transcriptional networks and, in few cases, has allowed characterizing multiple ON/OFF methylation-dependent phase variation systems [79–81]. The variable expression of MTases via, for example, slipped-strand mispairing of simple sequence repeats (SSRs), may lead to genome-wide methylation changes and to altered expression of multiple genes (commonly termed phasevarions) [82]. In an appraisal of the potential for phase-variation in bacterial methyltransferases, two recent studies revealed the presence of SSRs in as much as 2% and 17.4% of type I *hsdM* and type III *mod* genes, respectively [83,84]. Such type of systematic studies, coupled with information provided by long-read sequencing technologies, will likely set the

stage for further large-scale analyses of whole bacterial phasomes and development of controllable toggle switches.

Another interesting point would be to test the hypothesis that the thermodynamic effect of DNA methylation induces conformational changes to a bacterial chromosome, increasing gene accessibility to the transcriptional machinery [85,86]. Generation of methylation-induced non-B topologies [85,87] is likely to take place at higher methylation densities [88] and should provide key insight on how structural changes can alter the repertoire of genes exposed to the cellular transcriptional machinery. Techniques such as circular dichroism and chromatin conformation capture (e.g., Hi-C) can be used to elucidate the effects of bacterial DNA methylation on DNA conformation and, consequently, on gene expression [89,90]. Additionally, it would be worth testing the extent to which non-canonical (non-B) DNA conformations contribute to the occurrence of nonmethylated sites, particularly for those cases that cannot be explained by protein competitive binding. Hence, all three above-mentioned questions would strongly benefit from a community-wide analysis of core and persistent MTases.

Concluding Remarks and Future Perspectives

In this opinion article we propose that core and persistent DNA MTases should be prioritized in community-wide integrative studies to better understand bacterial epigenomes as well as the drivers behind MTase conservation. To illustrate this, we provided a comprehensive summary of the MTase conservation landscape in bacteria and highlight a catalog of 145 core and persistent MTases across 139 unique species, as well as a framework to guide future methylome analyses. These core and persistent MTases include not only well-characterized ones, but also multiple previously unknown ones in human and animal pathogens. These observations open a new window to more effectively study the basic science and translational aspects of epigenetic regulation in bacteria and call for a community-wide integrative effort using a data and knowledge sharing strategy such as the one we have outlined (see Outstanding Questions).

Due to their indispensability in bacteria, essential MTases (which are often core) are potential targets for the development of epigenetic inhibitors capable of, for example, enhancing the therapeutic activity of antimicrobials. For instance, Dam inhibition reportedly weakens bacterial pathogenicity *in vivo*, as GATC methylation controls virulence gene expression in various organisms [91–95]. GATC methylation was also found to play a role in drug potentiation, by curbing the therapeutic activity of the β -lactam and quinolone classes of antibiotics [15]. Indeed, Dam represents an attractive target for epigenetic inhibition of the multiple biological processes it regulates (e.g., virulence), as it lacks mammalian homologs while being conserved in several enteric pathogens [96–98]. Unlike Dam, CcrM has been found to be essential for viability in multiple bacteria [25,99,100], thus raising the possibility that inhibitors of methylation may be bactericidal in some cases. Although very promising, Dam, CcrM, and other similar MTases are prevalent across multiple bacterial species. From the point of view of the development of more targeted epigenetic inhibitors, other core/persistent MTases specific to only one of few species may hold greater interest. One example is that of the CAAAAA MTase of *C. difficile*, involved in the sporulation and biofilm formation in *C. difficile* [12].

We should emphasize that by proposing the prioritization of core and persistent MTases in methylome studies, we are by no means devaluing research focusing on *ad hoc* MTases. Such studies should be encouraged as they provide important contributes towards the understanding of MTase diversity and their specific roles in, for example, the emergence/maintenance of genetic cohesion of particularly virulent lineages [47–50] and genetic regulation in their natural (i.e., nonexperimentally perturbed) environment. Under such circumstances, recently acquired

Outstanding Questions

To what extent do additional DNA chemical modifications (beyond 6mA, 5mC, and 4mC) play regulatory roles in bacteria?

How can the functions of persistent MTases be identified?

What targets of these MTases have a regulatory nature?

How can the difference between epigenetic pathways be disentangled?

MTases may also represent good candidates for inhibitor development.

We anticipate that in the next few years, advances in existing and forthcoming long-read sequencing technologies, concurrently with additional progress in the understanding of multiple functional roles of core/persistent MTases, will offer unprecedented opportunities for achieving a more complete snapshot of bacterial methylomes, especially in human pathogens. These current and future advances make the present times an exciting period for studying and harnessing bacteria epigenomics for medical and clinical impact.

Author Contributions

P.H.O. and G.F. designed the study. P.H.O. performed all computational analyses. P.H.O. and G.F. wrote the manuscript.

Disclaimer Statement

The authors declare no competing financial interests.

Acknowledgments

We acknowledge Eduardo P.C. Rocha (Institut Pasteur, Paris, France) and Mi Ni, Yangmei Li from Fang lab for critical reading and for providing helpful comments/suggestions. We would also like to acknowledge the anonymous reviewers for their constructive comments. The work was funded by R01 GM114472 (G.F.) and R01 GM128955 (G.F.) from the National Institutes of Health. G.F. is an Irma T. Hirsch/Monique Weill-Caulier Trust Research Scholar. This work was also supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

Supplemental Information

Supplemental information associated with this article can be found online at <https://doi.org/10.1016/j.tim.2020.04.007>.

References

- Thomas, C.M. and Nielsen, K.M. (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3, 711–721
- Labrie, S.J. *et al.* (2010) Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* 8, 317–327
- Mruk, I. and Kobayashi, I. (2014) To be or not to be: regulation of restriction-modification systems and other toxin-antitoxin systems. *Nucleic Acids Res.* 42, 70–86
- Murphy, J. *et al.* (2013) Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence. *Appl. Environ. Microbiol.* 79, 7547–7555
- Rocha, E.P. *et al.* (2001) Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.* 11, 946–958
- Pingoud, A. *et al.* (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.* 62, 685–707
- Kobayashi, I. (2001) Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* 29, 3742–3756
- Beaulaurier, J. *et al.* (2019) Deciphering bacterial epigenomes using modern sequencing technologies. *Nat. Rev. Genet.* 20, 157–172
- Flusberg, B.A. *et al.* (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* 7, 461–465
- Blow, M.J. *et al.* (2016) The epigenomic landscape of prokaryotes. *PLoS Genet.* 12, e1005854
- Fang, G. *et al.* (2012) Genome-wide mapping of methylated adenine residues in pathogenic *Escherichia coli* using single-molecule real-time sequencing. *Nat. Biotechnol.* 30, 1232–1239
- Oliveira, P.H. *et al.* (2020) Epigenomic characterization of *Clostridioides difficile* finds a conserved DNA methyltransferase that mediates sporulation and pathogenesis. *Nat. Microbiol.* 5, 166–180
- Kumar, S. *et al.* (2018) N4-cytosine DNA methylation regulates transcription and pathogenesis in *Helicobacter pylori*. *Nucleic Acids Res.* 46, 3429–3445
- Nye, T.M. *et al.* (2019) DNA methylation from a type I restriction modification system influences gene expression and virulence in *Streptococcus pyogenes*. *PLoS Pathog.* 15, e1007841
- Cohen, N.R. *et al.* (2016) A role for the bacterial GATC methylome in antibiotic stress survival. *Nat. Genet.* 48, 581–586
- Kahramanoglou, C. *et al.* (2012) Genomics of DNA cytosine methylation in *Escherichia coli* reveals its role in stationary phase transcription. *Nat. Commun.* 3, 886
- Kwiatek, A. *et al.* (2015) Type III methyltransferase M.NgoAX from *Neisseria gonorrhoeae* FA1090 regulates biofilm formation and interactions with human cells. *Front. Microbiol.* 6, 1426
- Stephenson, S.A. and Brown, P.D. (2016) Epigenetic influence of Dam methylation on gene expression and attachment in uropathogenic *Escherichia coli*. *Front. Public Health* 4, 131
- Payelleville, A. *et al.* (2017) DNA adenine methyltransferase (Dam) overexpression impairs *Photobacterium luminescens* motility and virulence. *Front. Microbiol.* 8, 1671
- Oliveira, P.H. *et al.* (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.* 42, 10618–10631
- Mouammine, A. and Collier, J. (2018) The impact of DNA methylation in alphaproteobacteria. *Mol. Microbiol.* 110, 1–10
- Payelleville, A. *et al.* (2018) The complete methylome of an entomopathogenic bacterium reveals the existence of loci with unmethylated adenines. *Sci. Rep.* 8, 12091

23. Christen, B. *et al.* (2011) The essential genome of a bacterium. *Mol. Syst. Biol.* 7, 528
24. Julio, S.M. *et al.* (2001) DNA adenine methylase is essential for viability and plays a role in the pathogenesis of *Yersinia pseudotuberculosis* and *Vibrio cholerae*. *Infect. Immun.* 69, 7610–7615
25. Stephens, C. *et al.* (1996) A cell cycle-regulated bacterial DNA methyltransferase is essential for viability. *Proc. Natl. Acad. Sci. U. S. A.* 93, 1210–1214
26. Gonzalez, D. *et al.* (2014) The functions of DNA methylation by CcrM in *Caulobacter crescentus*: a global approach. *Nucleic Acids Res.* 42, 3720–3735
27. Estibariz, I. *et al.* (2019) The core genome m5C methyltransferase JHP1050 (M.Hpy99III) plays an important role in orchestrating gene expression in *Helicobacter pylori*. *Nucleic Acids Res.* 47, 2336–2348
28. Gartner, K. *et al.* (2019) Cytosine N4-methylation via M. Ssp6803II is involved in the regulation of transcription, fine-tuning of DNA replication and DNA repair in the cyanobacterium *Synechocystis* sp. PCC 6803. *Front. Microbiol.* 10, 1233
29. Hagemann, M. *et al.* (2018) Identification of the DNA methyltransferases establishing the methylome of the cyanobacterium *Synechocystis* sp. PCC 6803. *DNA Res.* 25, 343–352
30. Kita, K. *et al.* (1999) Evidence of horizontal transfer of the EcoO109I restriction-modification gene to *Escherichia coli* chromosomal DNA. *J. Bacteriol.* 181, 6822–6827
31. Kobayashi, I. *et al.* (1999) Shaping the genome–restriction-modification systems as mobile genetic elements. *Curr. Opin. Genet. Dev.* 9, 649–656
32. Lauster, R. (1989) Evolution of type II DNA methyltransferases. A gene duplication model. *J. Mol. Biol.* 206, 313–321
33. Jeltsch, A. and Pingoud, A. (1996) Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J. Mol. Evol.* 42, 91–96
34. Seshasayee, A.S. *et al.* (2012) Context-dependent conservation of DNA methyltransferases in bacteria. *Nucleic Acids Res.* 40, 7066–7073
35. Sberro, H. *et al.* (2019) Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* 178, 1245–1259
36. Weyler, L. *et al.* (2014) Restriction endonucleases from invasive *Neisseria gonorrhoeae* cause double-strand breaks and distort mitosis in epithelial cells during infection. *PLoS One* 9, e114208
37. Roberts, R.J. *et al.* (2015) REBASE – a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.* 43, D298–D299
38. Sanchez-Romero, M.A. and Casadesus, J. (2020) The bacterial epigenome. *Nat. Rev. Microbiol.* 18, 7–20
39. Putnam, C.D. (2016) Evolution of the methyl directed mismatch repair system in *Escherichia coli*. *DNA Repair* 38, 32–41
40. Kunkel, T.A. and Erie, D.A. (2005) DNA mismatch repair. *Annu. Rev. Biochem.* 74, 681–710
41. Roberts, D. *et al.* (1985) IS10 transposition is regulated by DNA adenine methylation. *Cell* 43, 117–130
42. Camacho, E.M. and Casadesus, J. (2005) Regulation of *traJ* transcription in the *Salmonella* virulence plasmid by strand-specific DNA adenine hemimethylation. *Mol. Microbiol.* 57, 1700–1718
43. Slater, S. *et al.* (1995) *E. coli* SeqA protein binds oriC in two different methyl-modulated reactions appropriate to its roles in DNA replication initiation and origin sequestration. *Cell* 82, 927–936
44. Beauchamp, J.M. *et al.* (2017) Methylation-dependent DNA discrimination in natural transformation of *Campylobacter jejuni*. *Proc. Natl. Acad. Sci. U. S. A.* 114, E8053–E8061
45. Jeltsch, A. (2003) Maintenance of species identity and controlling speciation of bacteria: a new function for restriction/modification systems? *Gene* 317, 13–16
46. Oliveira, P.H. *et al.* (2016) Regulation of genetic flux between bacteria by restriction-modification systems. *Proc. Natl. Acad. Sci. U. S. A.* 113, 5658–5663
47. Budroni, S. *et al.* (2011) *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4494–4499
48. Croucher, N.J. *et al.* (2014) Diversification of bacterial genome content through distinct mechanisms over different timescales. *Nat. Commun.* 5, 5471
49. Nandi, T. *et al.* (2015) *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res.* 25, 129–141
50. Forde, B.M. *et al.* (2015) Lineage-specific methyltransferases define the methylome of the globally disseminated *Escherichia coli* ST131 clone. *MBio* 6, e01602-15
51. Roberts, G.A. *et al.* (2013) Impact of target site distribution for Type I restriction enzymes on the evolution of methicillin-resistant *Staphylococcus aureus* (MRSA) populations. *Nucleic Acids Res.* 41, 7472–7484
52. Maddamssetti, R. *et al.* (2017) Core genes evolve rapidly in the long-term evolution experiment with *Escherichia coli*. *Genome Biol. Evol.* 9, 1072–1083
53. Oliveira, P.H. *et al.* (2017) The chromosomal organization of horizontal gene transfer in bacteria. *Nat. Commun.* 8, 841
54. Chao, M.C. *et al.* (2016) The design and analysis of transposon insertion sequencing experiments. *Nat. Rev. Microbiol.* 14, 119–128
55. van Opijnen, T. *et al.* (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. *Nat. Methods* 6, 767–772
56. Goodman, A.L. *et al.* (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe* 6, 279–289
57. Gawronski, J.D. *et al.* (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for *Haemophilus* genes required in the lung. *Proc. Natl. Acad. Sci. U. S. A.* 106, 16422–16427
58. Langridge, G.C. *et al.* (2009) Simultaneous assay of every *Salmonella typhi* gene using one million transposon mutants. *Genome Res.* 19, 2308–2316
59. Chao, M.C. *et al.* (2015) A cytosine methyltransferase modulates the cell envelope stress response in the *Cholera* pathogen. *PLoS Genet.* 11, e1005739
60. Stemon, J.F. *et al.* (2018) Transposon sequencing of *Brucella abortus* uncovers essential genes for growth *in vitro* and inside macrophages. *Infect. Immun.* 86, e00312-18
61. Phan, M.D. *et al.* (2013) The serum resistome of a globally disseminated multidrug resistant uropathogenic *Escherichia coli* clone. *PLoS Genet.* 9, e1003834
62. Dembek, M. *et al.* (2015) High-throughput analysis of gene essentiality and sporulation in *Clostridium difficile*. *MBio* 6, e02383
63. Jiao, Y. *et al.* (2014) A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics* 30, 2360–2366
64. Hernandez-Ferrer, C. *et al.* (2017) MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. *BMC Bioinformatics* 18, 36
65. Louhimo, R. and Hautaniemi, S. (2011) CNAmets: an R package for integrating copy number, methylation and expression data. *Bioinformatics* 27, 887–888
66. Wang, M. *et al.* (2015) Efficient test and visualization of multi-set intersections. *Sci. Rep.* 5, 16923
67. van der Woude, M. *et al.* (1996) Epigenetic phase variation of the pap operon in *Escherichia coli*. *Trends Microbiol.* 4, 5–9
68. Wallecha, A. *et al.* (2002) Dam- and OxyR-dependent phase variation of *agn43*: essential elements and evidence for a new role of DNA methylation. *J. Bacteriol.* 184, 3338–3347
69. Lim, H.N. and van Oudenaarden, A. (2007) A multistep epigenetic switch enables the stable inheritance of DNA methylation states. *Nat. Genet.* 39, 269–275
70. Erill, I. *et al.* (2017) Comparative analysis of *Ralstonia solanacearum* methylomes. *Front. Plant Sci.* 8, 504
71. Mirauta, B. *et al.* (2014) Parseq: reconstruction of microbial transcription landscape from RNA-seq read counts using state-space models. *Bioinformatics* 30, 1409–1416
72. Jorjani, H. and Zavolan, M. (2014) TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics* 30, 971–974

73. Solomon, M.J. *et al.* (1988) Mapping protein-DNA interactions *in vivo* with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. *Cell* 53, 937–947
74. Mikkelsen, T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* 448, 553–560
75. Gama-Castro, S. *et al.* (2016) RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res.* 44, D133–D143
76. Ishihama, A. *et al.* (2016) Transcription profile of *Escherichia coli*: genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Res.* 44, 2058–2074
77. Siervo, N. *et al.* (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res.* 36, D93–D96
78. Minch, K.J. *et al.* (2015) The DNA-binding network of *Mycobacterium tuberculosis*. *Nat. Commun.* 6, 5829
79. Casadesu, J. and Low, D.A. (2013) Programmed heterogeneity: epigenetic mechanisms in bacteria. *J. Biol. Chem.* 288, 13929–13935
80. De Ste Croix, M. *et al.* (2017) Phase-variable methylation and epigenetic regulation by type I restriction-modification systems. *FEMS Microbiol. Rev.* 41, S3–S15
81. Atack, J.M. *et al.* (2018) Phasevarions of bacterial pathogens: methylomics sheds new light on old enemies. *Trends Microbiol.* 26, 715–726
82. Phillips, Z.N. *et al.* (2019) Phasevarions of bacterial pathogens - phase-variable epigenetic regulators evolving from restriction-modification systems. *Microbiology* 165, 917–928
83. Atack, J.M. *et al.* (2018) A survey of type III restriction-modification systems reveals numerous, novel epigenetic regulators controlling phase-variable regulons; phasevarions. *Nucleic Acids Res.* 46, 3532–3542
84. Atack, J.M. *et al.* (2020) DNA sequence repeats identify numerous type I restriction-modification systems that are potential epigenetic regulators controlling phase-variable regulons; phasevarions. *FASEB J.* 34, 1038–1051
85. Polaczek, P. *et al.* (1998) GATC motifs may alter the conformation of DNA depending on sequence context and N6-adenine methylation status: possible implications for DNA-protein recognition. *Mol. Gen. Genet.* 258, 488–493
86. Ngo, T.T. *et al.* (2016) Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* 7, 10813
87. Diekmann, S. (1987) DNA methylation can enhance or induce DNA curvature. *EMBO J.* 6, 4213–4217
88. Moller, A. *et al.* (1981) 7-Methylguanine in poly(dG-dC).poly(dG-dC) facilitates z-DNA formation. *Proc. Natl. Acad. Sci. U. S. A.* 78, 4777–4781
89. Hognon, C. *et al.* (2019) Cooperative effects of cytosine methylation on DNA structure and dynamics. *J. Phys. Chem. B* 123, 7365–7371
90. Lee, D.S. *et al.* (2019) Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nat. Methods* 16, 999–1006
91. Garcia-Del Portillo, F. *et al.* (1999) DNA adenine methylase mutants of *Salmonella typhimurium* show defects in protein secretion, cell invasion, and M cell cytotoxicity. *Proc. Natl. Acad. Sci. U. S. A.* 96, 11578–11583
92. Pucciarelli, M.G. *et al.* (2002) Envelope instability in DNA adenine methylase mutants of *Salmonella enterica*. *Microbiology* 148, 1171–1182
93. Heithoff, D.M. *et al.* (1999) An essential role for DNA adenine methylation in bacterial virulence. *Science* 284, 967–970
94. Watson Jr., M.E. *et al.* (2004) Inactivation of deoxyadenosine methyltransferase (dam) attenuates *Haemophilus influenzae* virulence. *Mol. Microbiol.* 53, 651–664
95. Robinson, V.L. *et al.* (2005) A dam mutant of *Yersinia pestis* is attenuated and induces protection against plague. *FEMS Microbiol. Lett.* 252, 251–256
96. Luo, N. *et al.* (2018) DNA methyltransferase inhibition upregulates MHC-I to potentiate cytotoxic T lymphocyte responses in breast cancer. *Nat. Commun.* 9, 248
97. Stressemann, C. *et al.* (2006) Functional diversity of DNA methyltransferase inhibitors in human cancer cell lines. *Cancer Res.* 66, 2794–2800
98. Brueckner, B. and Lyko, F. (2004) DNA methyltransferase inhibitors: old and new drugs for an epigenetic cancer therapy. *Trends Pharmacol. Sci.* 25, 551–554
99. Kahng, L.S. and Shapiro, L. (2001) The CcrM DNA methyltransferase of *Agrobacterium tumefaciens* is essential, and its activity is cell cycle regulated. *J. Bacteriol.* 183, 3065–3075
100. Robertson, G.T. *et al.* (2000) The *Brucella abortus* CcrM DNA methyltransferase is essential for viability, and its overexpression attenuates intracellular replication in murine macrophages. *J. Bacteriol.* 182, 3482–3489
101. Roberts, R.J. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* 31, 1805–1812
102. Rao, D.N. *et al.* (2014) Type III restriction-modification enzymes: a historical perspective. *Nucleic Acids Res.* 42, 45–55
103. Murray, N.E. (2000) Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.* 64, 412–434

