Evolution and Impact of Transposable Elements and Viruses

Florian Maumus

HABILITATION À DIRIGER DES RECHERCHES
de l'Université Paris-Sud

Spécialité :
Biologie Végétale, Ecologie, Evolution

Présentée par

Florian MAUMUS
Unité de recherche en Génomique-Info (UR1164) - Institut National de la Recherche Agronomique

Evolution and Impact of Transposable Elements and Viruses

Soutenance prévue le 16 novembre 2018 à l'INRA de Versailles, devant le jury d'examen :

Mr. Yves Bigot, Directeur de recherche - Centre INRA Val de Loire - rapporteur

Mr. Richard Cordaux, Directeur de recherche - Université de Poitiers - examinateur

Mme Aurélie Hua-Van, Maître de conférence – Université Paris Sud – examinateur

Mr. Carlos Llorens, CEO/CSO at Biotechvana – examinateur

Mr. Pierre Pontarotti, Directeur de recherche – Université d'Aix-Marseille – rapporteur

Pr. Jean-Nicolas Volff, Professeur – ENS Lyon – rapporteur

TABLE DES MATIERES

CURICULUM VITAE

Florian Maumus
Researcher at INRA
Born in Paris, France on 1980 July 19th
Father of two

URGI-INRA - Route de St Cyr, 78026 Versailles Cedex, France
Florian.maumus@inra.fr
+33 1 30 83 31 74

CURRENT POSITION

Since July 2014 **Researcher** at INRA Versailles-Grignon, Versailles, France
Co-evolution of selfish genetic elements and their host

DEGREES

2009	PhD in Biology, Paris Sud Orsay University (Paris XI), France
2005	Master 2 in Plant Science, Jussieu University (Paris VI), France
2004	Master 1 in Cellular Biology, Jussieu University (Paris VI), France
2003	Licence in Biology and Biochemistry, Descartes University (Paris V)

PREVIOUS POSITIONS AND EDUCATION

2011-2014	Postdoctoral researcher supervised by Dr. Hadi Quesneville at INRA Versailles-Grignon, Versailles, France # Evolution of plant genomes and epigenomes, Paleovirology
2009-2010	Postdoctoral researcher supervised by Dr. Hervé Vaucheret at INRA Versailles-Grignon, Versailles, France # Epigenetic regulation in plants (AGO1 homeostasis in Arabidopsis thaliana)
2005-2009	PhD supervised by Dr. Chris Bowler at École Normale Supérieure, Paris, France and Stazione Zoologica Anton Dohrn, Naples, Italy # Transposable elements and epigenetic regulation in marine algae
2000-2005	Master degree in plant science at Pierre & Marie Curie University, Paris, France. Internship (2005) supervised by Chris Bowler at École Normale Supérieure, Paris, France and Angela Falciatore at Stazione Zoologica Anton Dohrn, Naples, Italy # Photobiology in marine diatoms (Cryptochromes in Phaedodactylum tricornutum)

SUPERVISION OF STUDENTS

Master 2 students:

Ophélie Jouffroy (2015) – Impact of transposable elements on tomato ripening
Seydine Diop (2017) – Macroevolution of pararetroviruses

PhD students:

Ophélie Jouffroy (until mid-2018) – Footprints of purifying selection over epigenome and sequences of transposable elements

TEACHING ACTIVITIES

Since 2014: Invited lectures for the European Master of Genetics at Paris-Diderot University on the bioinformatics analysis of the impact of transposable elements in genomes. Three hours per year.

OTHER RESPONSABILITIES

From 2018 Organizer of the Journal Club at URGI

From 2018 Principal investigator on TE annotation for the Open Green Genomes (OGG) project funded by the Joint Genome Institute (JGI, Walnut Creek, CA, USA) and coordinated by Jim Leebens-Mack (University of Georgia, USA)

From 2018 Editorial board member for Scientific Reports

2016 Member of the jury of the thesis of Sébastien Guizard, supervised by Yves Bigot (INRA)

From 2016 PhD advisor of Roland Akapo supervised by Dr. Karine Alix (INRA-ArgoParisTech) & Dr Xavier Vigouroux (IRD)

From 2016 Principal investigator on TE annotation for the Brassicales Map Alignment Program (BMAP) funded by the Joint Genome Institute (JGI, Walnut Creek, CA, USA) and coordinated by Stephen Wright (University of Toronto, Canada)

From 2015 PhD advisor of Victoire Baillet supervised by Dr. Vincent Colot (École Normale Supérieure)

2011-2014 Board member of the French Society of Genetics

GRANTS

2018-2020 Work package leader for ANR project EVENTS (18 months postdoc supervision)
Endogenous viral elements: role in virus evolution and functions in plants

2018-2019 Work package leader for BAP INRA project methylTOM
Methylation across the genome of tomato ddm1 mutants

2017-2018 Work package leader for BAP INRA project LTR-HYBRID
Retrotransposon-associated changes during allopolyploidy

AWARDS

2017 Promising researcher award (INRA Laurier) attributed each year by an international jury to one top early career INRA researcher

INDEX

Profile on Google Scholar:

40 scientific publications

Over 3,000 citations

Hirsch index = 22

PUBLICATIONS

† 1st or co-1st author

‡ Corresponding or co-corresponding author

Articles

1- A. Rastogi, U. Maheswari, R.G. Dorrell, F.R.J. Vieira, **F. Maumus**, A. Kustka, J. McCarthy, A.E. Allen, P. Kersey, C. Bowler, L. Tirichine, Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms, Sci Rep, 8 (2018) 4834.

2- D. Lang, K.K. Ullrich, F. Murat, J. Fuchs, J. Jenkins, F.B. Haas, M. Piednoel, H. Gundlach, M. Van Bel, R. Meyberg, C. Vives, J. Morata, A. Symeonidi, M. Hiss, W. Muchero, Y. Kamisugi, O. Saleh, G. Blanc, E.L. Decker, N. van Gessel, J. Grimwood, R.D. Hayes, S.W. Graham, L.E. Gunter, S.F. McDaniel, S.N.W. Hoernstein, A. Larsson, F.W. Li, P.F. Perroud, J. Phillips, P. Ranjan, D.S. Rokshar, C.J. Rothfels, L. Schneider, S. Shu, D.W. Stevenson, F. Thummler, M. Tillich, J.C. Villarreal Aguilar, T. Widiez, G.K.

Wong, A. Wymore, Y. Zhang, A.D. Zimmer, R.S. Quatrano, K.F.X. Mayer, D. Goodstein, J.M. Casacuberta, K. Vandepoele, R. Reski, A.C. Cuming, G.A. Tuskan, **F. Maumus**, J. Salse, J. Schmutz, S.A. Rensing, The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution, Plant J, 93 (2018) 515-533.

3- S.I. Diop, A.D.W. Geering, F. Alfama-Depauw, M. Loaec, P.Y. Teycheney, **F. Maumus**[‡], Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae, Sci Rep, 8 (2018) 572.

4- M.H. Schmidt, A. Vogel, A.K. Denton, B. Istace, A. Wormit, H. van de Geest, M.E. Bolger, S. Alseekh, J. Mass, C. Pfaff, U. Schurr, R. Chetelat, **F. Maumus**, J.M. Aury, S. Koren, A.R. Fernie, D. Zamir, A.M. Bolger, B. Usadel, De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing, Plant Cell, 29 (2017) 2336-2348.

5- T. Mock, R.P. Otillar, J. Strauss, M. McMullan, P. Paajanen, J. Schmutz, A. Salamov, R. Sanges, A. Toseland, B.J. Ward, A.E. Allen, C.L. Dupont, S. Frickenhaus, **F. Maumus**, A. Veluchamy, T. Wu, K.W. Barry, A. Falciatore, M.I. Ferrante, A.E. Fortunato, G. Glockner, A. Gruber, R. Hipkin, M.G. Janech, P.G. Kroth, F. Leese, E.A. Lindquist, B.R. Lyon, J. Martin, C. Mayer, M. Parker, H. Quesneville, J.A. Raymond, C. Uhlig, R.E. Valas, K.U. Valentin, A.Z. Worden, E.V. Armbrust, M.D. Clark, C. Bowler, B.R. Green, V. Moulton, C. van Oosterhout, I.V. Grigoriev, Evolutionary genomics of the cold-adapted diatom Fragilariopsis cylindrus, Nature, 541 (2017) 536-540.

6- T.C. Mathers, Y. Chen, G. Kaithakottil, F. Legeai, S.T. Mugford, P. Baa-Puyoulet, A. Bretaudeau, B. Clavijo, S. Colella, O. Collin, T. Dalmay, T. Derrien, H. Feng, T. Gabaldon, A. Jordan, I. Julca, G.J. Kettles, K. Kowitwanich, D. Lavenier, P. Lenzi, S. Lopez-Gomollon, D. Loska, D. Mapleson, **F. Maumus**, S. Moxon, D.R. Price, A. Sugio, M. van Munster, M. Uzest, D. Waite, G. Jander, D. Tagu, A.C. Wilson, C. van Oosterhout, D. Swarbreck, S.A. Hogenhout, Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species, Genome Biol, 18 (2017) 27.

7- A. Gouin, A. Bretaudeau, K. Nam, S. Gimenez, J.M. Aury, B. Duvic, F. Hilliou, N. Durand, N. Montagne, I. Darboux, S. Kuwar, T. Chertemps, D. Siaussat, A. Bretschneider, Y. Mone, S.J. Ahn, S. Hanniger, A.G. Grenet, D. Neunemann, **F. Maumus**, I. Luyten, K. Labadie, W. Xu, F. Koutroumpa, J.M. Escoubas, A. Llopis, M. Maibeche-Coisne, F. Salasc, A. Tomar, A.R. Anderson, S.A. Khan, P. Dumas, M. Orsucci, J. Guy, C. Belser, A. Alberti, B. Noel, A. Couloux, J. Mercier, S. Nidelet, E. Dubois, N.Y. Liu, I. Boulogne, O. Mirabeau, G. Le Goff, K. Gordon, J. Oakeshott, F.L. Consoli, A.N. Volkoff, H.W. Fescemyer, J.H. Marden, D.S. Luthe, S. Herrero, D.G. Heckel, P. Wincker, G.J. Kergoat, J. Amselem, H. Quesneville, A.T. Groot, E. Jacquin-Joly, N. Negre, C. Lemaitre, F. Legeai, E. d'Alencon, P. Fournier, Two genomes of highly polyphagous lepidopteran pests (Spodoptera frugiperda, Noctuidae) with different host-plant ranges, Sci Rep, 7 (2017) 11816.

8- S. Basu, S. Patil, D. Mapleson, M.T. Russo, L. Vitale, C. Fevola, **F. Maumus**, R. Casotti, T. Mock, M. Caccamo, M. Montresor, R. Sanges, M.I. Ferrante, Finding a partner in the ocean: molecular and evolutionary bases of the response to sexual cues in a planktonic diatom, New Phytol, 215 (2017) 140-156.

9- J.L. Olsen, P. Rouze, B. Verhelst, Y.C. Lin, T. Bayer, J. Collen, E. Dattolo, E. De Paoli, S. Dittami, **F. Maumus**, G. Michel, A. Kersting, C. Lauritano, R. Lohaus, M. Topel, T. Tonon, K. Vanneste, M.

Amirebrahimi, J. Brakel, C. Bostrom, M. Chovatia, J. Grimwood, J.W. Jenkins, A. Jueterbock, A. Mraz, W.T. Stam, H. Tice, E. Bornberg-Bauer, P.J. Green, G.A. Pearson, G. Procaccini, C.M. Duarte, J. Schmutz, T.B. Reusch, Y. Van de Peer, The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea, Nature, 530 (2016) 331-335.

10- **F. Maumus**, G. Blanc, Study of Gene Trafficking between Acanthamoeba and Giant Viruses Suggests an Undiscovered Family of Amoeba-Infecting Viruses, Genome Biol Evol, 8 (2016) 3351-3363.

11- O. Jouffroy, S. Saha, L. Mueller, H. Quesneville, **F. Maumus**[‡], Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening, BMC Genomics, 17 (2016) 624.

12- E.M. Willing, V. Rawat, T. Mandakova, **F. Maumus**, G.V. James, K.J. Nordstrom, C. Becker, N. Warthmann, C. Chica, B. Szarzynska, M. Zytnicki, M.C. Albani, C. Kiefer, S. Bergonzi, L. Castaings, J.L. Mateos, M.C. Berns, N. Bujdoso, T. Piofczyk, L. de Lorenzo, C. Barrero-Sicilia, I. Mateos, M. Piednoel, J. Hagmann, R. Chen-Min-Tao, R. Iglesias-Fernandez, S.C. Schuster, C. Alonso-Blanco, F. Roudier, P. Carbonero, J. Paz-Ares, S.J. Davis, A. Pecinka, H. Quesneville, V. Colot, M.A. Lysak, D. Weigel, G. Coupland, K. Schneeberger, Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation, Nat Plants, 1 (2015) 14023.

13- F. Murat[†], A. Louis[†], **F. Maumus**[†], A. Armero, R. Cooke, H. Quesneville, H. Roest Crollius, J. Salse, Understanding Brassicaceae evolution through ancestral genome reconstruction, Genome Biol, 16 (2015) 262.

14- D.R. Hoen, G. Hickey, G. Bourque, J. Casacuberta, R. Cordaux, C. Feschotte, A.S. Fiston-Lavier, A. Hua-Van, R. Hubley, A. Kapusta, E. Lerat, **F. Maumus**, D.D. Pollock, H. Quesneville, A. Smit, T.J. Wheeler, T.E. Bureau, M. Blanchette, A call for benchmarking transposable element annotation methods, Mob DNA, 6 (2015) 13.

15- M. El Baidouri, K.D. Kim, B. Abernathy, S. Arikit, **F. Maumus**, O. Panaud, B.C. Meyers, S.A. Jackson, A new approach for annotation of transposable elements using small RNA mapping, Nucleic Acids Res, 43 (2015) e84.

16- G. Blanc[†‡], L. Gallot-Lavallee, **F. Maumus**[†‡], Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses, Proc Natl Acad Sci U S A, 112 (2015) E5318-5326.

17- **F. Maumus**[‡], H. Quesneville, Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter, PLoS One, 9 (2014) e94101.

18- **F. Maumus**[‡], H. Quesneville, Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana, Nat Commun, 5 (2014) 4104.

19- **F. Maumus**[†‡], A. Epert, F. Nogue, G. Blanc[†‡], Plant genomes enclose footprints of past infections by giant virus relatives, Nat Commun, 5 (2014) 4268.

20- S. Lopez-Gomollon, M. Beckers, T. Rathjen, S. Moxon, **F. Maumus**, I. Mohorianu, V. Moulton, T. Dalmay, T. Mock, Global discovery and characterization of small non-coding RNAs in marine microalgae, BMC Genomics, 15 (2014) 697.

21- A.D. Geering†, **F. Maumus**†, D. Copetti, N. Choisne, D.J. Zwickl, M. Zytnicki, A.R. McTaggart, S. Scalabrin, S. Vezzulli, R.A. Wing, H. Quesneville, P.Y. Teycheney, Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution, Nat Commun, 5 (2014) 5269.

22- A. Bolger, F. Scossa, M.E. Bolger, C. Lanz, **F. Maumus**, T. Tohge, H. Quesneville, S. Alseekh, I. Sorensen, G. Lichtenstein, E.A. Fich, M. Conte, H. Keller, K. Schneeberger, R. Schwacke, I. Ofner, J. Vrebalov, Y. Xu, S. Osorio, S.A. Aflitos, E. Schijlen, J.M. Jimenez-Gomez, M. Ryngajllo, S. Kimura, R. Kumar, D. Koenig, L.R. Headland, J.N. Maloof, N. Sinha, R.C. van Ham, R.K. Lankhorst, L. Mao, A. Vogel, B. Arsova, R. Panstruga, Z. Fei, J.K. Rose, D. Zamir, F. Carrari, J.J. Giovannoni, D. Weigel, B. Usadel, A.R. Fernie, The genome of the stress-tolerant wild tomato species Solanum pennellii, Nat Genet, 46 (2014) 1034-1038.

23- A. Veluchamy, X. Lin, **F. Maumus**, M. Rivarola, J. Bhavsar, T. Creasy, K. O'Brien, N.A. Sengamalay, L.J. Tallon, A.D. Smith, E. Rayko, I. Ahmed, S. Le Crom, G.K. Farrant, J.Y. Sgro, S.A. Olson, S.S. Bondurant, A.E. Allen, P.D. Rabinowicz, M.R. Sussman, C. Bowler, L. Tirichine, Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum, Nat Commun, 4 (2013) 2091.

24- T. Slotte, K.M. Hazzouri, J.A. Agren, D. Koenig, **F. Maumus**, Y.L. Guo, K. Steige, A.E. Platts, J.S. Escobar, L.K. Newman, W. Wang, T. Mandakova, E. Vello, L.M. Smith, S.R. Henz, J. Steffen, S. Takuno, Y. Brandvain, G. Coop, P. Andolfatto, T.T. Hu, M. Blanchette, R.M. Clark, H. Quesneville, M. Nordborg, B.S. Gaut, M.A. Lysak, J. Jenkins, J. Grimwood, J. Chapman, S. Prochnik, S. Shu, D. Rokhsar, J. Schmutz, D. Weigel, S.I. Wright, The Capsella rubella genome and the genomic consequences of rapid mating system evolution, Nat Genet, 45 (2013) 831-835.

25- B.A. Read, J. Kegel, M.J. Klute, A. Kuo, S.C. Lefebvre, **F. Maumus**, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J.M. Claverie, S. Frickenhaus, K. Gonzalez, E.K. Herman, Y.C. Lin, J. Napier, H. Ogata, A.F. Sarno, J. Shmutz, D. Schroeder, C. de Vargas, F. Verret, P. von Dassow, K. Valentin, Y. Van de Peer, G. Wheeler, C. Emiliania huxleyi Annotation, J.B. Dacks, C.F. Delwiche, S.T. Dyhrman, G. Glockner, U. John, T. Richards, A.Z. Worden, X. Zhang, I.V. Grigoriev, Pan genome of the phytoplankton Emiliania underpins its global distribution, Nature, 499 (2013) 209-213.

26- A. Zambounis, M. Elias, L. Sterck, **F. Maumus**, C.M. Gachon, Highly dynamic exon shuffling in candidate pathogen receptors ... what if brown algae were capable of adaptive immunity?, Mol Biol Evol, 29 (2012) 1263-1276.

27- C. Llorens, R. Futami, L. Covelli, L. Dominguez-Escriba, J.M. Viu, D. Tamarit, J. Aguilar-Rodriguez, M. Vicente-Ripolles, G. Fuster, G.P. Bernet, **F. Maumus**, A. Munoz-Pomer, J.M. Sempere, A. Latorre, A. Moya, The Gypsy Database (GyDB) of mobile genetic elements: release 2.0, Nucleic Acids Res, 39 (2011) D70-74.

28- E. Rayko, **F. Maumus**, U. Maheswari, K. Jabbari, C. Bowler, Transcription factor families inferred from genome sequences of photosynthetic stramenopiles, New Phytol, 188 (2010) 52-66.

29- J.M. Cock, L. Sterck, P. Rouze, D. Scornet, A.E. Allen, G. Amoutzias, V. Anthouard, F. Artiguenave, J.M. Aury, J.H. Badger, B. Beszteri, K. Billiau, E. Bonnet, J.H. Bothwell, C. Bowler, C. Boyen, C. Brownlee, C.J. Carrano, B. Charrier, G.Y. Cho, S.M. Coelho, J. Collen, E. Corre, C. Da Silva, L. Delage, N. Delaroque, S.M. Dittami, S. Doulbeau, M. Elias, G. Farnham, C.M. Gachon, B. Gschloessl, S. Heesch, K.

Jabbari, C. Jubin, H. Kawai, K. Kimura, B. Kloareg, F.C. Kupper, D. Lang, A. Le Bail, C. Leblanc, P. Lerouge, M. Lohr, P.J. Lopez, C. Martens, **F. Maumus**, G. Michel, D. Miranda-Saavedra, J. Morales, H. Moreau, T. Motomura, C. Nagasato, C.A. Napoli, D.R. Nelson, P. Nyvall-Collen, A.F. Peters, C. Pommier, P. Potin, J. Poulain, H. Quesneville, B. Read, S.A. Rensing, A. Ritter, S. Rousvoal, M. Samanta, G. Samson, D.C. Schroeder, B. Segurens, M. Strittmatter, T. Tonon, J.W. Tregear, K. Valentin, P. von Dassow, T. Yamagishi, Y. Van de Peer, P. Wincker, The Ectocarpus genome and the independent evolution of multicellularity in brown algae, Nature, 465 (2010) 617-621.

30- **F. Maumus**, A.E. Allen, C. Mhiri, H. Hu, K. Jabbari, A. Vardi, M.A. Grandbastien, C. Bowler, Potential impact of stress activated retrotransposons on genome evolution in a marine diatom, BMC Genomics, 10 (2009) 624.

31- V. De Riso, R. Raniello, **F. Maumus**, A. Rogato, C. Bowler, A. Falciatore, Gene silencing in the marine diatom Phaeodactylum tricornutum, Nucleic Acids Res, 37 (2009) e96.

32- C. Bowler, A.E. Allen, J.H. Badger, J. Grimwood, K. Jabbari, A. Kuo, U. Maheswari, C. Martens, **F. Maumus**, R.P. Otillar, E. Rayko, A. Salamov, K. Vandepoele, B. Beszteri, A. Gruber, M. Heijde, M. Katinka, T. Mock, K. Valentin, F. Verret, J.A. Berges, C. Brownlee, J.P. Cadoret, A. Chiovitti, C.J. Choi, S. Coesel, A. De Martino, J.C. Detter, C. Durkin, A. Falciatore, J. Fournet, M. Haruta, M.J. Huysman, B.D. Jenkins, K. Jiroutova, R.E. Jorgensen, Y. Joubert, A. Kaplan, N. Kroger, P.G. Kroth, J. La Roche, E. Lindquist, M. Lommer, V. Martin-Jezequel, P.J. Lopez, S. Lucas, M. Mangogna, K. McGinnis, L.K. Medlin, A. Montsant, M.P. Oudot-Le Secq, C. Napoli, M. Obornik, M.S. Parker, J.L. Petit, B.M. Porcel, N. Poulsen, M. Robison, L. Rychlewski, T.A. Rynearson, J. Schmutz, H. Shapiro, M. Siaut, M. Stanley, M.R. Sussman, A.R. Taylor, A. Vardi, P. von Dassow, W. Vyverman, A. Willis, L.S. Wyrwicz, D.S. Rokhsar, J. Weissenbach, E.V. Armbrust, B.R. Green, Y. Van de Peer, I.V. Grigoriev, The Phaeodactylum genome reveals the evolutionary history of diatom genomes, Nature, 456 (2008) 239-244.

Reviews

1- A. Villain, L. Gallot-Lavallee, G. Blanc, **F. Maumus**, Giant viruses at the core of microscopic wars with global impacts, Curr Opin Virol, 17 (2016) 130-137.

2- **F. Maumus**, H. Quesneville, Impact and insights from ancient repetitive elements in plant genomes, Curr Opin Plant Biol, 30 (2016) 41-46.

3- **F. Maumus**, A.-S. Fiston-Lavier, H. Quesneville, Impact of transposable elements on insect genomes and biology, Current Opinion in Insect Science, 7 (2015) 30-36.

4- **F. Maumus**, P. Rabinowicz, C. Bowler, M. Rivarola, Stemming epigenetics in marine stramenopiles, Curr Genomics, 12 (2011) 357-370.

RESEARCH

SYNOPSIS

Since my early career, I have been fascinated by the evolution of Selfish Genetic Elements (SGEs - including viruses and transposable elements) and their impact on the evolution of the genomes and epigenomes of their host organisms. By integrating host genomes repeatedly in a stochastic manner, SGEs can profoundly influence the biology of their hosts. They represent a predominant part of most eukaryotic genomes and constitute a major source of genetic and epigenetic changes. Friends or foes? The integration of SGEs in genomes causes deleterious mutations most of the time, but it occasionally mediates key evolutionary adaptations. My goal in research is to help elucidating the modes and consequences of these evolutionary tradeoffs, especially in plants and algae.

EARLY CAREER

I was beginning my PhD when I first heard about transposable elements (TEs). A postdoc from the lab told me “look, this thing is transcriptionally activated in response to nitrogen starvation!!”. That was Andrew E. Allen, now group leader at J. Craig Venter Institute (San Diego, USA). We were studying diatoms, more specifically the genome and transcriptomes of the model species *Phaeodactylum tricornutum* (Pt, Figure 1)). “But it does not seem to be just a gene, it has best hit against hopscotch, which is a transposable element”, he added. I had studied plant science at the University and it was the first time I heard about a transposable element; it sounded so exciting! I took this over as a starting point in my PhD project. Several libraries of expressed sequence tags (ESTs) had been generated from Pt under a variety of growth conditions. I looked for similar sequences across the EST libraries and found that a second transposable element was transcribed in response to an aldehyde reactive diatoms use to fight grazers. We first confirmed the transcriptional activation of both elements in response to respective conditions using quantitative RT-PCR. Because these elements were genetic parasites of a marine diatom, we called them *Blackbeard* and *Surcouf*, respectively, as referring to the infamous pirate and corsair (Figure 1).

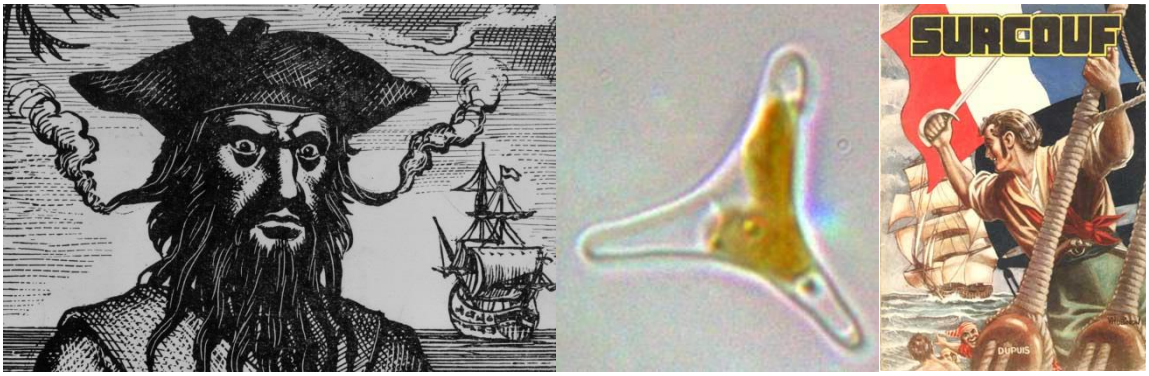

Figure 1: Illustration of legendary pirate Blackbeard (left) and French corsair Surcouf (right) and picture of the diatom *Phaeodactylum tricornutum* with triradiate morphology (middle).

It was not really new to find a TE in a eukaryotic genome, but they had never been thoroughly characterized in diatoms. A drudgery of manual sequence annotation and curation enabled to identify all the (almost) intact TEs in the genome. By chance, it was only 27 megabases of DNA! It was

followed by the phylogenetic analysis of these elements and I discovered that *Blackbeard*, *Surcouf*, and all Ty1/Copia-type Long Terminal Repeats (LTR) retrotransposons were distantly related to known lineages and they were classified in six different clades named CoDis (Copia from Diatoms) (Maumus et al., 2009) (Llorens et al., 2010).

It was known that DNA methylation plays a major role in TE silencing in plants, animals and fungi. Nothing was known about DNA methylation in diatoms except that it does exist as measured by flow cytometry decades ago (Jarvis et al., 1992). I found out that most TEs, and especially LTR retrotransposons, were heavily methylated in the Pt genome. I next addressed whether any changes in DNA methylation would occur at *Blackbeard* locus in response to nitrate starvation. The signal was very strong: transcriptional activation was accompanied by substantial hypomethylation of this locus. It constituted the first proof that environmental changes can lead to epigenetic modifications in diatoms.

And more, TE transcriptional activation could lead to new insertions that could impact the diatom's biology, the big dogma. We therefore performed what has been one of the most exciting series of experiments during my thesis to address whether the *Blackbeard* element would be mobile in response to nitrate starvation. During one of these, we isolated one hundred single Pt cells under the microscope and put them to grow into clonal colonies. We had no idea what the insertion frequency could be but one hundred were worth the try. After isolating DNA from each colony, we performed transposon display experiments in collaboration with the laboratory of Marie-Angèle Grandbastien (INRA, France). In the end, we could not clearly identify new *Blackbeard* insertions in any of the Pt colonies. One of the explanations could be that *Blackbeard* transposes at low frequency and/or that the Pt genome is so compact that most insertions are deleterious and rapidly eliminated from the population. Although we could not detect new TE insertions over such short evolutionary times, using transposon display, we observed high polymorphism among *Blackbeard* loci across several Pt accessions isolated from different parts of the world, thereby demonstrating that *Blackbeard* is an active element (Maumus et al., 2009).

Combining these results, I wrote my first article as first author. After submitting to several prestigious journals, it was eventually published in BMC Genomics (Maumus et al., 2009). This paper is about to reach 70 citations, underlying its significance in the field.

While locus-specific measurements indicated that TEs were commonly methylated in Pt, we had no idea how DNA methylation was distributed genome-wide. It was for instance unknown whether some gene bodies would be methylated as well. I adapted a restriction-based protocol (Lippman et al., 2005) to produce a whole genome map of DNA methylation for the Pt genome. In collaboration with NimbleGen Company, we generated a high density tiling array to which uncut DNA was hybridized. And now we had obtained the raw results from peak calling, we found out it would be impossible to have it analyzed by whom we had thought of. The project was then stalled for years, leading to loss of novelty and conflicts regarding authorships. Fortunately, the project followed up and it was eventually published in 2013 (Veluchamy et al., 2013).

After my PhD, I have joined the team of Hervé Vaucheret (IJPB, INRA) for a first postdoc (18 months). The team has extensive expertise and is among world leaders in plant epigenetics, more specifically small RNA pathways in *Arabidopsis thaliana*. That was a great opportunity to develop my knowledge about epigenetic pathways and to discover the power of using genetics and mutant lines.

Working with such a model system seemed much different compared to the drudgery of working with emerging models that are diatoms for countless aspects. It has been a chance to unfold a variety of experiments in molecular biology, transgenesis and so on within short time. During my project, I have addressed some facets of the homeostasis of Argonaute 1, which is a key factor that regulates the production of most microRNAs in plants (Mallory and Vaucheret, 2009). Argonaute 1 transcripts are themselves the target of two miRNA isoforms: miR168a and miR168b. Intriguingly, each of the two isoforms produce different ratio of 21nt and 22nt small RNA that are respectively responsible of simple transcript cleavage and transcript cleavage priming amplification loop. I have investigated the determinism of the ratio of 21nt versus 22nt and found that this was depending on both the hairpin structure and the bases found at the extremity of the mature miRNA. As latest news, this work is still ongoing and has expanded into a vast story in Vaucheret's lab.

As a side project (or homework), I continued working on the annotation of repetitive elements in algal genomes during this first postdoc. Confronted to escalating amount of data, it became evident to me that I would benefit from switching to bioinformatics analyses to lever my capability to study evolution, transposable elements, epigenetics and much more.

SWITCHING TO BIOINFORMATICS

I switched to bioinformatics during my second postdoctoral contract (from 2011), in the INRA Unité de Recherche en Génomique-Info (URGI) which has developed a robust package for TE detection and annotation called REPET. It was a great opportunity to get my hands on bioinformatics while feeding my curiosity for SGEs.

EVOLUTION AND IMPACT OF THE PLANT REPEATOME

GENOME ANNOTATION

Owing to my expertise on transposable elements and genome evolution, I have been involved in several international genome projects, leading to publications in high impact journals including Cell (1), Nature (5), Nature genetics (2) and Nature plants (1). I have annotated several genomes from plants, insects, and various algae: small ones, bigger ones, A+T rich or G+C-rich, TE-rich or TE-poor. Every genome came with its specificities and represented a new challenge as well as an opportunity to learn and improve my methodology.

As recent examples (published in 2018), I was leading the repeat analysis of two basal plant genomes. The genome of the moss *Physcomitrella patens* was first published in 2008 as a draft composed of over 2,000 scaffolds (Rensing et al., 2008). While providing most of the gene set to the community, such a fragmented assembly was limiting the possibilities to address genome structure and evolution. In the last years, Sefan Rensing (University of Marburg, Germany) coordinated an effort to generate a chromosome-scale assembly for *P. patens* to offer access to such studies. Surprisingly, we first observed that, in contrast to plant genomes of similar size (~500 Mb), the distribution of repetitive elements is homogenous along the chromosome: no density peak corresponding to potential centromeres could be detected and no density gradient reminiscent of

pericentromeres could be observed. Nevertheless, FISH experiments targeting centromeric histone variants clearly established the *P. patens* chromosomes to be monocentric. Interestingly, detailed analysis of TE distribution revealed that a specific family of Copia-type TEs, comprising both autonomous and non-autonomous elements, presents a single peak in most chromosomes, suggesting that they might correspond to or be tightly associated with centromeric functions. We have recently published this new assembly in The Plant Journal (Lang et al., 2018) and I am signing as a senior author.

As another recent example, I was in charge of repeat analysis in the genome of the charophyte algae *Chara braunii*. Charophytes represent the lineage of green algae that is closest to that from which land plants emerged so that its genome is expected to reveal early plant traits and main terrestrialization events. I like basal stuff in general ;) so I boarded in, again with Stefan Rensing as coordinator. The *C. braunii* is about 2 Gb, of which 1.4 Gb were assembled into contigs with 75% corresponding to repetitive elements. However this estimate is probably low, given that highly similar repeats are challenging to assemble. Intriguingly in this species we found no Copia-type LTR RTs unlike in most plants and green algae, while Gypsy-type LTR RTs are predominant. In addition, we could detect repetitive elements with putative GIY-YIG homing endonuclease and reverse transcriptase domains, which are hallmarks of Penelope retrotransposons and group II introns, both being hitherto highly uncommon in plant genomes. Among a range of genomic features, comparative analysis with other plant genomes revealed that introns are remarkably long in *C. braunii* and we found that Penelope-like elements are way more frequent in introns than expected by chance, suggesting insertion bias or function in the turnover of introns. The *C. braunii* genome paper comes with a lot more exciting findings! It has just been accepted in Cell and I am signing as a senior author. More profound investigation of the SGE content in such basal plant genomes should be implemented in future.

With every genome analyzed, it is a challenge annotating right but also interpreting, mining the data towards pulling some biology, some hypothesis out of such analyses. There is always some novelty to catch, sometimes independent of the specific genome that is being analyzed.

Over the years, I became an internationally recognized expert in the annotation of repeated DNA in eukaryotic genomes. In 2014, I was among the few experts attending an international meeting in the field of transposable element annotation (TEAM, coordinated by Mathieu Blanchette, McGill University, Canada) to discuss the need for new strategies towards standardized characterization and quality improvement of TE annotation. We discussed questions such as how to limit false positives in repeat annotations and how to reach a maximum of sensitivity. These are central questions in the field, though largely ignored by those who use these annotations in their analyses. Nevertheless, high quality annotation of repetitive elements is crucial to support a range of epigenetics, genetics and transcriptomic analyses. A white paper reported the consortium's main discussions and proposed work ahead to the community such as the establishment of benchmark annotations (Hoen et al., 2015).

Conceptually, there are different ways to detect repetitive elements in a genome that can fold into two main categories: either based on sequence alignments or based on k-mer counting. Further, some other tools such as P-clouds build clouds of highly similar k-mers while RepeatScout uses k-mers to identify seeds from which alignments are extended. Programs also exist to build

consensus of repetitive elements by assembling highly frequent k-mers found from DNA sequencing reads. Of yet another kind, the TASR pipeline proposes to identify repeated elements by searching clusters of 24-nt small RNAs, which are commonly associated with silencing of repetitive elements by targeting DNA methylation. After testing a variety of tools, my experience is that alignment-based annotations produce higher quality and sensitivity as compare to common k-mer based approaches. The construction of a library of consensus sequences also allows their classification. By contrast, the k-mer strategies are typically faster and offer high specificity. Both types of approaches are complementary and should be applied to answer different questions.

PUSHING THE LIMITS OF DETECTION

Non-annotated sequences in genomes are known as genomic dark matter. Because of their abundance in eukaryotic genomes the decay of repetitive elements is thought to be a major source of genomic dark matter. Indeed, eukaryotic genomes contain a diversity of repetitive elements of different age, and the vast majority of this genetic material is not functional and thus accumulates mutations and deletions over time. Following such a sequence decay model, TE copies are expected to accumulate mutations and loose repetitiveness until genetic information enabling their characterization completely melts in the form of random DNA sequence (Figure 2). Because TEs and other repeats are abundant in most eukaryotic genomes, it is hypothesized that a large fraction of un-annotated segments in genomes, a.k.a. the genomic dark matter, find their origin in the long drift of repetitive elements (Maumus and Quesneville, 2016). TE sequences can occasionally become domesticated into new cellular functions. In such cases, very ancient repetitive elements may be unrecognizable as such but annotated for various functions such as promoter or intron. The comprehensive detection of repeated and repeat-derived sequences hence represents a difficult task that has challenged scientists since the beginning of the genomic era.

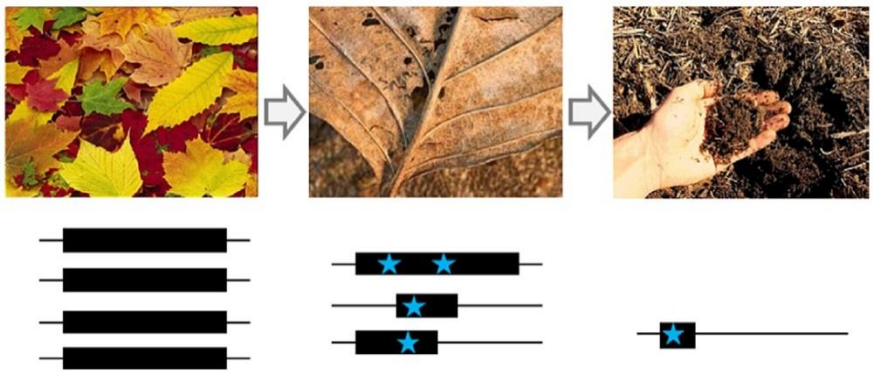

Figure 2: Repeat decay. Metaphorically, repeats are composted over time towards becoming genomic humus. Image taken from publication (Maumus and Quesneville, 2016).

Owing to its small genome and limited recent TE activity, *A. thaliana* provides an excellent model to study the long-term evolution of repetitive elements. Consensus sequences are derived from the alignment of several copies of a given TE family. Mutations being independent in each copy, the consensus sequences approximate the ancestral sequences. As a result, the identity between a copy and the cognate consensus sequence is negatively correlated with the time of integration. Thus high copy vs consensus identities indicate recent copy insertions, while low identities indicate more ancient events. Measuring the identity between repeat copies and consensus from the annotation of the *A. thaliana* genome with REPET, I could establish that most repeats are ancient and that recent integrations are relatively poorly frequent, consistent with its downsizing trend (Oyama et al., 2008).

By contrast, the distribution of copy vs consensus identity values in *A. lyrata* shows a significant peak of recent integrations (Figure 3), as documented (Hu et al., 2011).

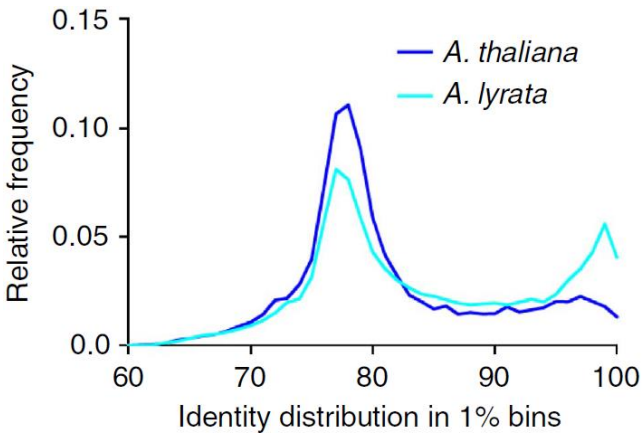

Figure 3: Distribution in 1% bins of the identity values between genomic copies and consensus sequences in *A. thaliana* and *A. lyrata*. Image taken from publication (Maumus and Quesneville, 2016).

Considering that most detected repeats in *A. thaliana* are ancient, it is likely that a significant fraction remains undetected using standard parameters, i.e. default REPET parameters. Comparing REPET to other pipelines for de novo repeat annotation, I demonstrated the greater sensitivity, specificity, and annotation quality offered by REPET. To be noted, however, that the best results were obtained when combining consensus generated by different programs (Maumus and Quesneville, 2014b).

I next used REPET to explore the nature of the dark matter of the *A. thaliana* genome (i.e. the fraction that is pristine of annotation). Using a series of creative strategies, I was able to demonstrate that a significant amount of the *A. thaliana* dark matter is of repetitive origin. Altogether, I was able to assign a repetitive origin to about 33% of the genome, as compared to 24% obtained with standard approaches. To achieve this, I have tested four different annotation strategies based on scientific ground, which all turned out to provide relevant results. Of the most efficient approaches, the reiterative annotation proved to produce annotation of high sensitivity and very good specificity. It is based on the evidence that repeat copies as a whole hold more information than consensus sequences. Consensus sequences and genome annotation are established with default settings. The sequences from all the repeat copies of sufficient quality are then extracted to build a new repeat library that is used to run a new genome annotation. The process can be repeated a number of times. At the first iteration, genome coverage was increased by 20%, and tended to reach a plateau at the fourth iteration.

For another strategy, I reasoned that, owing to the reducing trend of the *A. thaliana* genome, several sequences of repetitive origin could remain undetected because they are now low- or single-copy elements. However, each repeat family from the genome of a common ancestral host has evolved independently up to the modern species. Hence, the genomes of closely related species such as *Arabidopsis lyrata* could contain repeats with significant similarity to low copy repeat-derived elements in *A. thaliana* and thereby would allow their detection in the latter. I therefore constructed a library comprising consensus sequences representative of repeated elements from eight *Brassicaceae* as well as from four *A. thaliana* ecotypes and used this library to annotate the *A. thaliana* genome. Importantly, REPET applies a “best score wins” post-process, implying that if a locus is hit by several consensus, only the annotation giving highest score is kept. Plotting the distribution of the consensus-vs-copy identity values obtained per consensus derived from each

source *Brassicaceae* genome reveals the power of such an approach. Indeed, while all the high (>90%) identity hits are attributed to consensus derived from *A. thaliana*, the vast majority of the low identity hits are attributed to consensus inferred from other species (Figure 4), suggesting that genetic information of ancient repeat families was better conserved in the genomes of different *Brassicaceae* species than it was in *A. thaliana*. This analysis therefore provides independent evidence that most *A. thaliana* repeats are ancient.

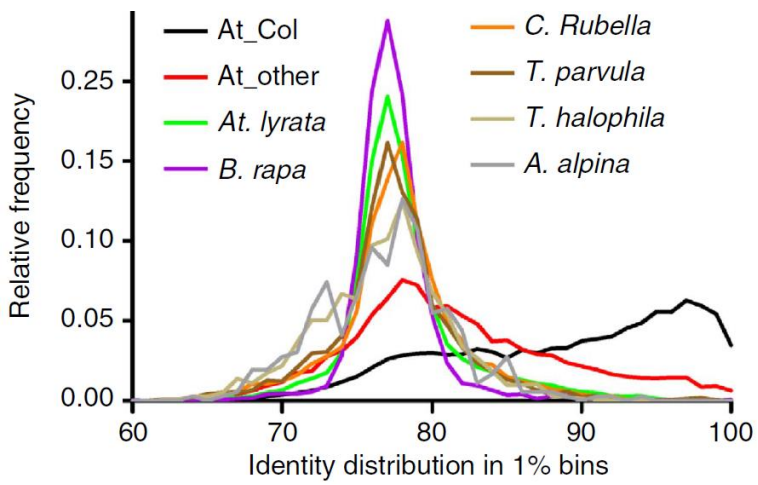

Figure 4: Competitive annotation of the A. thaliana genome with the Brassicaceae library. For each species, we plot the distribution of identity values between genomic copies and consensus sequences in 1% bins. Image taken from publication (Maumus and Quesneville, 2016).

Learning from the results obtained from different exploratory strategies, I have designed a workflow to generate deep but conservative annotation of repetitive elements in *A. thaliana*. Remarkably, this deep repeatome annotation enables the detection of regions of high repeat content besides those commonly found in plant centromeres and pericentromeres; especially one region in the right arm of chromosome 1 that I named At1R2. I have then framed this distribution in the context of karyotype evolution. Importantly, *A. thaliana* has five chromosomes while the *Brassicaceae* ancestor has eight. The *A. thaliana* karyotype thus results of several recent chromosome fusions. By contrast, reconstruction of the ancestral *Brassicaeae* karyotype in collaboration with the teams of Jérôme Salse (INRA) and Hugues Roest-Crollius (ENS) enabled determining that the karyotype of the close relative *Capsella rubella* is highly similar to that of the *Brassicaceae* ancestor. I therefore run a deep repeatome annotation on *C. rubella* and casted orthologous genes between the two genomes (Figure 5). Remarkably, I could establish that the most prominent non-(peri)centromeric repeat density peaks observed on *A. thaliana* actually correspond to the positions of the centromere and pericentromere of chromosome 2 in *C. rubella*. This suggests that relatively high repeat density in At1R2 corresponds the remains of a (peri-)centromere that is currently disappearing in *A. thaliana* following the fusion of ancestral chromosomes 1 & 2. Remarkably, when analyzing some properties of this region using re-sequencing data from a panel of 80 *A. thaliana* accessions (Cao et al., 2011) (Choi et al., 2013), I found that it has elevated deletion and recombination rates as compared to the remainder of chromosome arms, and that intergenic spaces therein are short as compared to homologs from other six other *Brassicaceae* genomes responsible of significant size reduction of this region in *A. thaliana* (Murat et al., 2015) (Maumus and Quesneville, 2016).

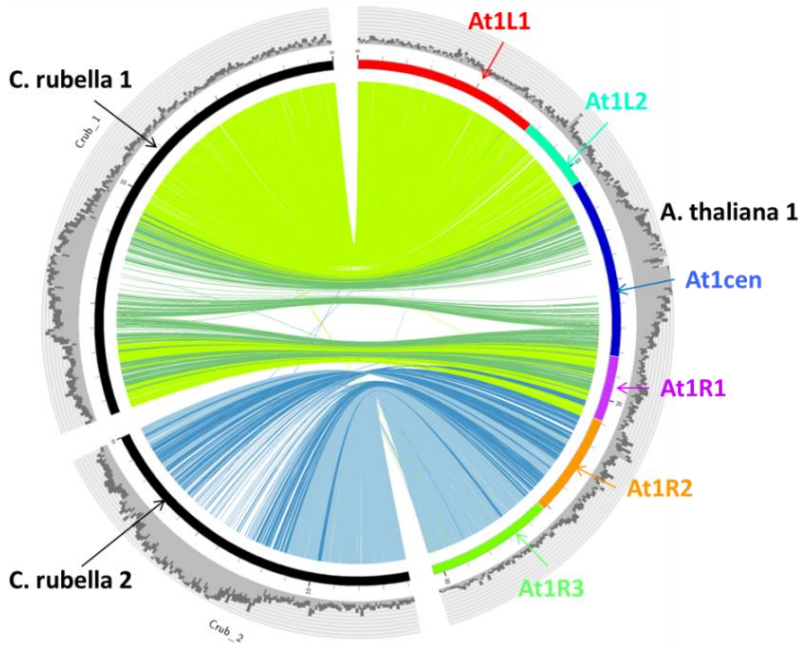

Figure 5: Regular (gray) and deep (dark gray) repeatome annotations on outmost track are positioned along the *A. thaliana* chromosome 1 (right) and the *C. rubella* chromosomes 1 (top left) and 2 (bottom left). At1R2 region is orange on the inner band. The central ribbons connect orthologous genes identified by best reciprocal hits and synteny conservation. Ribbons are bold when deep repeatome density in C. rubella exceeds 50%. Image taken from publication (Maumus and Quesneville, 2016).

These studies enable to better understand the origin of some genomic dark matter, which nature and function remains largely cryptic. Indeed, they show that part of it forms a continuum with repetitive DNA and suggests that another part that remains beyond detection possibilities is likely of similar origin (albeit more ancient). They also suggest that besides the detectable fraction, repeats have played a major role in the evolution of plant genome size and composition and probably in the emergence of genes and regulatory elements as well. We also started illustrating the dynamics of decaying centromeres appearing in monocentric species after chromosome fusions.

Performing deep repeat annotations allows identifying old, fragmented repeat-derived sequences, some of which may be functional and beneficial to the host. I then decided to apply such an annotation to the tomato (Solanum Lycopersicum) genome that I found being an appealing model of crop of agro-economic interest owing to its medium sized genome (approx. 900 Mb) and the large panel of publicly available epigenetic, genome re-sequencing, and transcriptomic data. It was also the opportunity to test a patch I designed for TEannot, the annotation pipeline of REPET. Out from my backpack returning from Barbados (remember, the TEAM meeting), I had designed this patch to filter potential false positive more efficiently and I had tested it on *A. thaliana* with promising results. As currently implemented, to filter weak hits in the annotation, TEannot calculates a threshold score on the basis of the 99th percentile of the scores obtained for all consensus against a random sequence that is generated by shuffling dinucleotides from the reference genome. I tried to use another type of random sequence to bait for potential false positive hits. I used the reversed (not complemented) reference sequence as it has the theoretical advantage to hold more low complexity sequences (which by experience are a source of false positives) than a shuffled genome does. Indeed, the number of hits against reversed genome is two orders of magnitude higher than that obtained versus shuffled sequences. In addition, I calculated a threshold score for each consensus in the repeat library that clearly shows that few naughty consensus containing low complexity DNA need high filtering threshold scores while the most present virtually no risk to cause false positives and need no further filtering than those defined by BLAST settings applied. Basically, the patch was working like charm in my hands, as demonstrated by extreme stability of results when using different levels of “sensitivity” with TEannot, that otherwise can produce substantially different annotations.

This patch should be further benchmarked by the developers of REPET and should be implemented in future releases. This annotation of the tomato genome with REPET has been complemented with k-mer based annotations to produce a repeatome map that covers 96 % of the initial repeat annotation (reciprocally 82 %). This track is now used as reference on the genome browser of the Sol Genomics Network. It provided good starting material to characterize the composition and distribution of the tomato repeatome and its overlap with methylome as well as to address the potential effects of TEs on gene expression. This work has been continued by internship student (then PhD student) Ophélie Jouffroy under my supervision.

ADDRESSING THE LONG-TERM IMPACTS OF REPEATS

Back to Arabidopsis... Having determined that most *A. thaliana* repeats are evolutionary old, this genome offers a good model to analyze the evolution of repeated sequences, their regulation, and their impact over time. To allow comparisons, I arbitrarily defined two categories of repeats based on the copy-vs-consensus identity values: recent (>85% identity) and ancient (<85% identity). It was for instance unknown whether *A. thaliana* repeats show a distribution bias related to their age. Interestingly, I observed that young repeats are found almost exclusively in centromeric and pericentromeric regions, a distribution that overlaps very well the repeat density along the chromosomes. By contrast, older repeats are found all along the chromosomes (Maumus and Quesneville, 2014a).

I next addressed the impact of such a distribution on genome composition and the regulation of gene expression. Recent analysis of the *A. thaliana* genome evolution over 30 generations has shown that most mutations occur in repeats and are C=>T transitions (Ossowski et al., 2010). Because the spontaneous deamination of methylated cytosines (mC) - which are frequent in repeated sequences – resulting in thymines is the most common mutation in DNA, mC deamination is thought to be a driving force in repeat decay. As introduced above, repetitive elements are commonly methylated. But for how long does this last? To answer this question, I have compared the presence or absence of 24-nt small RNA that map to ancient versus recent repeat copies. In line with the known methylated status of repeats, I observed that virtually all recent repeats are targeted by 24nt small RNAs. Interestingly, half the ancient repeats also are and present DNA methylation levels comparable to those found in recent repeats. Surprisingly, the sRNA mapping density was even higher in ancient repeats than in recent ones. Together, it suggests that repeat methylation lasts for prolonged evolutionary times, way more than it takes to functionally inactivate a TE: short TE fragments loaded with mutations keep being methylated. Once repetitive elements are rapidly silenced upon insertion (Mari-Ordonez et al., 2013), they apparently enter a self-feeding loop that lasts for millions of years and that is sustained by an increasing variety of sRNA molecules which is needed to cover independently diverging copies. This process then ends at some point, as illustrated by the amount of ancient repeats that are devoid of DNA methylation (Maumus and Quesneville, 2014a).

If DNA methylation can continue over prolonged periods (i.e. tens of millions of years), so could the deamination process. I addressed how this would translate at the level of genome composition and found that repetitive DNA is G+C-poor in the regions where only ancient repeats are found (i.e. the chromosome arms) compared to the (peri)centromeric regions. In corollary, repeats present a negative correlation between identity with their cognate consensus and G+C content, suggesting that they lose G+C residues over time, as expected through the cumulative effect of

deamination. Incidentally though, the most ancient repeats (70-75% identity with consensus) present a relatively high G+C content. A similar distribution of G+C content over time (i.e. bins of copy-vs-consensus identity) has then been observed individually in different super-families of TEs, showing that this trend is not owed to the behavior of specific and predominant elements (Maumus and Quesneville, 2014a). Altogether, these results suggest a bimodal evolution of G+C content in repeats (Figure 6), which is in fact expected by inertia given the dynamic and products of those conflicting forces - DNA methylation and deamination. DNA methylation causes mC deamination, which itself leads to a reduction of the G+C content of a given repetitive element. Overtime, cytosine targets become scarce so that DNA methylation levels decrease and hence deamination becomes less frequent. The basal transitions and transversions would then take over as the main mutations in repeats, thereby leading to increasing G+C content.

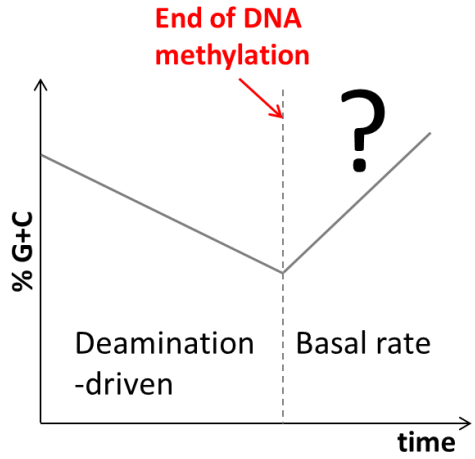

Figure 6: Theoretical evolution of G+C levels in repetitive elements over time: first decreasing owing to deamination events, and then increasing following methylation-free mutational forces.

I then investigated whether the age of a repeat would affect its impact on the expression levels of nearby genes. As expected following the chromosomal repartition of young repeats in gene-poor regions versus old repeats scattered all along, the latter were found to be on average closer to genes. Genes having a repeat flanking or within their sequence were then found to be expressed at lower levels than repeat-free genes. My first interpretation of this result was that repeats would cause low expression levels of nearby genes over prolonged evolutionary periods. A well-advised reviewer asked to try confirming this causal relationship. Fortunately, the data would allow doing so. Using sequences and expression data from two *A. thaliana* accessions (Wang et al., 2013) independently, I found, on average, no difference between the expression levels of repeat-proximal genes in Col-0 and their repeat-free orthologs in Bur-0 and C24 (Maumus and Quesneville, 2014a). Hence, the most common explanation for low gene expression levels in the vicinity of repeats is likely that repeated elements are more frequent in the vicinity of low expressed genes. This may be due to a lower selective pressure against repeated elements nearby weakly expressed genes than nearby highly expressed ones, then indicating that the repeats still present in the *A. thaliana* genome have modest impact on gene expression. The genome-wide data averaged in this analysis is whatsoever likely to hide individual genes which expression is indeed regulated by the presence of repeats, including cases in which this process is adaptive and conserved in different accessions.

In summary, I could establish that the ancient proliferation of repeat families has long-term consequences on plant biology and genome composition. These results are highly significant in the context of the identification of epigenome-associated QTLs and translational research, and will help addressing the epigenetic impact of repetitive elements on plant adaptation and domestication in an evolutionary perspective.

REPEATS AND EVOLUTIONARY GENOMICS (NOT PUBLISHED – KEEP CONFIDENTIAL)

In the last years, I had the opportunity to join the the Brassicales Map Alignment Program (BMAP) which is coordinated by Stephen Wright (University of Toronto). Nineteen *Brassicaceae* species selected to represent a high taxonomic diversity have been sequenced and assembled. In addition to 13 publicly available *Brassicaceae* genomes, the BMAP consortium goals are, among other things, to analyse the sequence conservation and to identify conserved non-coding elements (CNEs) along the *Brassicaceae* chromosomes, at different phylogenetic depths. This project represents a unique opportunity to address TE evolution at the plant family level.

The microevolution of TE families over prolonged periods of time is little documented. Actually, current data argues in favor of a general lack of conservation of TE families between plant species, which is interpreted as the consequence of rapid births and deaths. The taxonomic density represented by *Brassicaceae* genomes allows addressing TE conservation within a plant family at unprecedented resolution. Using a *de novo* approach, we have constructed consensus sequences that are representative of repetitive elements found in each of the 32 Brassica genomes. We have then assessed the conservation of repeat families across different *Brassica* at a level of 80% identity between nucleotide sequences over 50% of their length. Out of 34,424 consensus sequences, only 148 (0.4%) could be grouped in clusters comprising sequences from more than one species. Remarkably, five clusters comprise consensus sequences from at least ten species.

The largest cluster, corresponding to Copia-type LTR retrotransposons, contains 47 consensus sequences from 17 *Brassicaceae* species spanning all lineages (I, II and III, (Huang et al., 2016)). Because it appears to inhabit many *Brassicaceae* genomes, this family of Copia-type elements was named Habitans (latin for “resident”). Importantly, phylogenetic reconstruction is highly consistent with that of the host species hence supporting vertical transmission from a common ancestor. This exceptional conservation across a substantial number of species offers exemplary data set that allows investigating the adaptation and microevolution of a TE family.

We detected several characteristics of LTR retrotransposons in Habitans sequences including a 5’ LTR, followed by a primer-binding site (PBS), a 5’ spacer, a sequence encoding the Gag and Pol polyproteins, a polypurine tract (PPT), and a 3’ LTR (FIG). For most LTR retrotransposons, Gag and Pol open reading frames (ORFs) overlap and are separated by a stop codon or by a frameshift. From a single transcription start site located in the 5’ LTR, occasional translational recoding mechanisms such as readthrough or ribosomal wobbling enable to regulate the ratio of Gag/Pol protein products and to maximize virion production. Surprisingly, in Habitans, we found that Gag and Pol ORFs are separated by 130-220 bp non-coding sequences. Investigating the mechanisms that could mediate the translation of Pol ORFs, we found that secondary structure prediction of RNA corresponding to this intervening region commonly shows thermodynamically stable hairpin or multi-stem shapes (SUPFIG). Such structures are reminiscent of those described in some viruses such as members of *Caulimoviridae* and *Retroviridae* that were reported to induce ribosomal shunt (Futterer et al., 1993); i.e. a takeoff of ribosomes at mRNA stem and reassembly at downstream transcription start site. Remarkably, an ATG codon at the start of the Pol ORF is invariant across all the Habitans consensus sequences, strongly supporting possible ribosomal shunt as a conserved strategy to regulate relative Gag and Pol stoichiometry.

From the alignment of Habitans consensus sequences from different species (Figure 7), we could determine that the external regions, including the LTRs and the 5' spacer, can only be loosely aligned. Although a few sites are highly conserved such as PBS, their alignment varies widely depending on the cost of gap opening and closing. Remarkably, within the internal sequence, the poorest conservation scores cover the putative shunt-inducing region that separates the Gag and Pol polyproteins. Hence, the high variability of this regulatory region is likely to reflect positive selection towards optimal protein and DNA stoichiometry that is crucial to maximize TE mRNA productivity. Finally, within protein-coding regions, the Gag polyprotein shows about half the conservation observed in Pol. With this remarkable example, we demonstrate that, while it appears to be quite rare, a TE family can be conserved across tens of millions of years in plant genomes, thereby revealing some TE evolutionary trends and proposing an alternative hypothesis to horizontal transfer when highly similar TEs are found in relatively distant species from a given plant family. These results will be published in the context of the BMAP consortium paper.

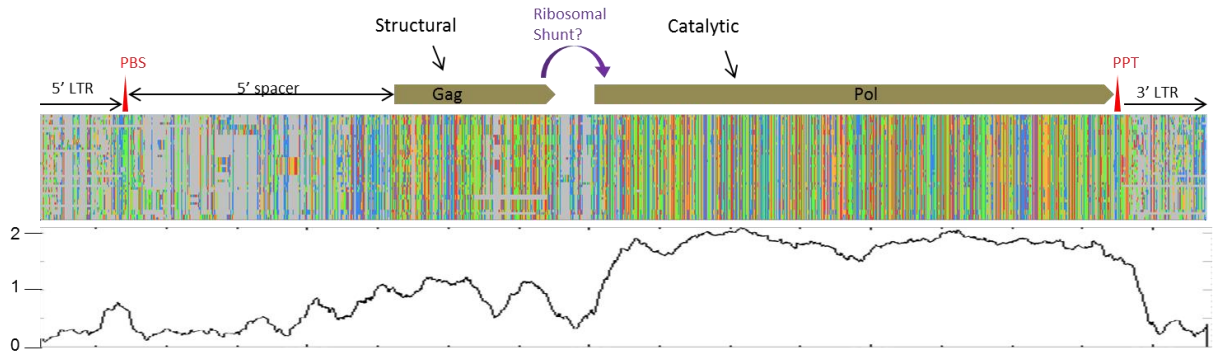

Figure 7: From top to bottom: structural annotation of Habitans elements, overview of multiple sequence alignment of 45 consensus from 17 *Brassicaceae* species, and inferred conservation plot.

ENDOGENOUS VIRUSES

There is a specific type of SGEs, the viruses, which I am especially curious to know more about, particularly in the context of plant evolution. Besides their control over plant populations during episodes of infection, the DNA of environmental viruses can occasionally become integrated in plant genomes and transmitted vertically. Because the mutation rates in eukaryotic genomes are orders of magnitude lower than those of viruses, these endogenous virus elements (EVEs) can conserve information from ancient and extinct viruses. The study of EVEs in the context of virus evolution and host-virus co-evolution is called "Paleovirology'. Much like a fossil record, paleovirology does allow the evolution of viruses to be traced. For example, the study of endogenous retroviruses has enabled to uncover their hidden diversity and host range, and has provided evidence that retroviruses have a marine origin and that they developed in parallel with their vertebrate hosts more than 450 million years ago (Hayward et al., 2015) (Aiewsakun and Katzourakis, 2017). The extent, impact, and importance of EVEs have been extensively described in animals but are much less well understood in plants.

CAULIMOVIRIDAE

The *Caulimoviridae* is one of the five families of reverse-transcribing viruses or virus-like retrotransposons that occur in eukaryotes (Pringle, 1998), and is the only family of viruses with a double-stranded DNA genome that infects plants. Unlike retroviruses, *Caulimoviridae* do not integrate their DNA in the genome of their host to complete their replication cycle. Nevertheless, caulimovirid DNA can occasionally integrate their host genome passively.

When analyzing repetitive elements in grape, I discovered a new genus of *Caulimoviridae* called 'Florendovirus'. Fortuitously, I teamed with Andrew Geering (CSIRO, Australia) and Pierre-Yves Teycheney (CIRAD, France) who simultaneously made a similar observation. Searching for closely related sequences in a series of plant genomes, we found that Florendovirus have colonized the genomes of a large diversity of flowering plants (spanning from ANITA grade angiosperms to Dicotyledons), sometimes at very high copy numbers. Interestingly, the structural and phylogenetic analysis of Florendovirus helps understanding the evolutionary history of the *Caulimoviridae*. For instance, we revealed that some Florendovirus are apparently defined by two complementary genomes (called bi-partite genomes), which is a unique feature among *Caulimoviridae*. Such partitioning is thought to support fine-tuned regulation of the transcription of different ORFs towards optimal stoichiometry of the different viral proteins. Furthermore, by comparing several closely related Oryza genomes, we were able to detect a Florendovirus insertion dating back to an estimated 1.8-2.3 Million years ago, providing proof of long-term retention of EVEs in plant genomes (Geering et al., 2014). Interestingly, we found that the borders of Florendovirus loci consisting of stretches of TA dinucleotides are significantly more frequent than expected by chance (Figure 8) and comparative genomics in Oryza revealed the stretches to predate viral intergation. This observation may point to the mechanism of integration as TA dinucleotide-rich areas of sequence are more likely to form highly stable secondary structures that perturb DNA replication, thereby causing chromosome fragility (Zlotorynski et al., 2003). Florendovirus DNA could then be coopted to act as filler DNA to repair the double-stranded DNA breaks by either non-homologous end joining or microhomology-mediated end joining.

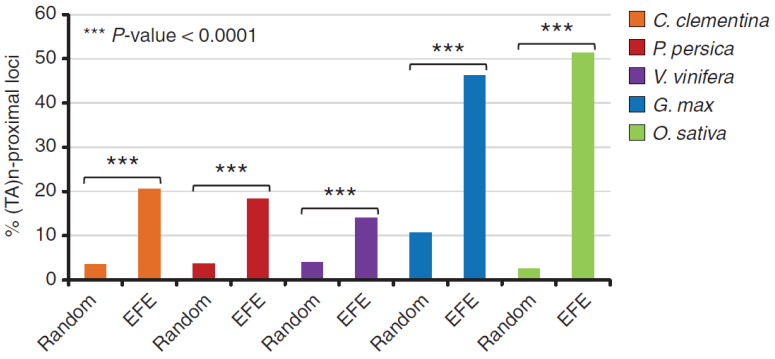

Figure 8: Physical concomitance of endogenous florendovirus elements (EFEs) and TA dinucleotide ((TA)n) repeats. The percentages of EFE and equal numbers of random loci that are located at less than 1kbp from (TA)n repeats are shown. Image taken from publication (Geering et al., 2014).

Plant genomes would probably hold much more information that could help understanding the diversity and host range of *Caulimoviridae*. I have thus aimed at mining a large collection of plant genomes to perform a comprehensive analysis of related EVEs. Setting a framework for macroevolutionary studies, we have screened genomes representative of the breadth of Viridiplantae, from green algae to flowering plants. Because it is most conserved across *Caulimoviridae* and used for classification, we used the reverse transcriptase domain as digital probe

for mining a set of plant genomes. Interestingly, while *Caulimoviridae* were known to infect flowering plants, we found corresponding EVEs in most vascular plants, including the most primitive plants such as ferns, conifers and lycopods, thereby illustrating the capacity of *Caulimoviridae* to infect a wide, previously underestimated range of hosts. Phylogenetic analysis revealed that most *Caulimoviridae* EVEs identified belong to the genera *Florendovirus* and *Petuvirus* (Figure 9). It also led to the discovery of previously unknown evolutionary branches of *Caulimoviridae*: four new genera of the family *Caulimoviridae* were detected in conifers *(Gymnendovirus* 1 to 4), two in ferns *(Fernendovirus* 1 and 2) and five in flowering plants (eg *Xendovirus* or *Yendovirus)* (Diop et al., 2017).

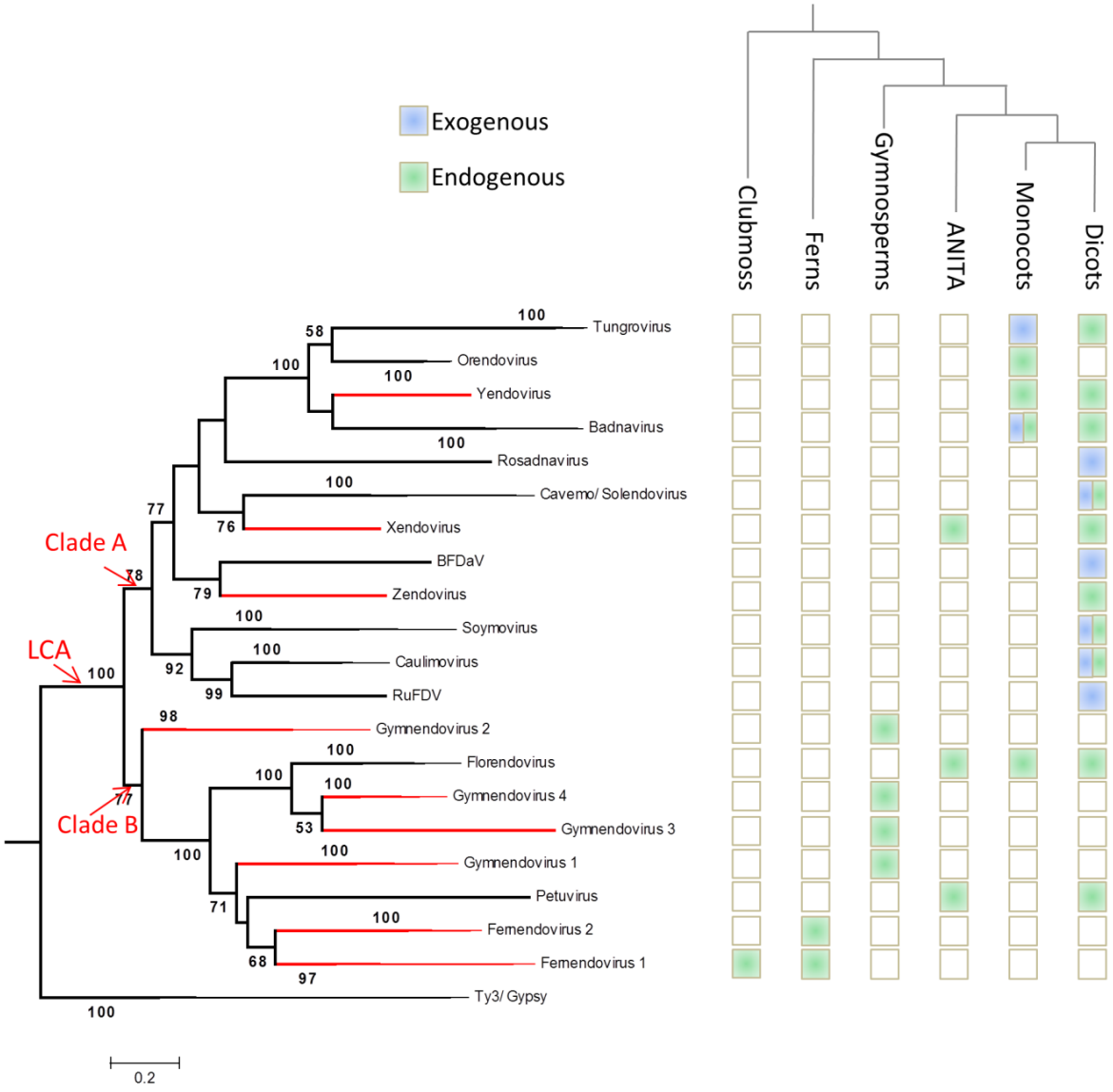

Figure 9: Phylogeny of the *Caulimoviridae*, as inferred using maximum likelihood criteria and a multiple sequence alignment of protease, reverse transcriptase and ribonuclease H1 domain sequences from recognized (black) and putative (red) genera. The upper cladogram indicates the evolutionary relationships between major classes of vascular plants. At the intersection between both trees, colored boxes indicate the presence of either endogenous (green) or exogenous (blue) representatives of the *Caulimoviridae*. Image taken from publication (Diop et al., 2017).

Based on an analysis of the distribution of different *Caulimoviridae* genera across Viridiplantae taxonomy, we proposed an evolutionary scenario in which this virus family emerged in an ancestor shared by all vascular plants, dating back to the Devonian period (- 320 million years). There were apparently subsequent exchanges between hosts belonging to different plant divisions. Remarkably, the large host spectrum of *Caulimoviridae* may correlate with the presence in these viruses of a movement protein (MP) that facilitates the cell-to-cell transport and circulation of viral particles through plasmodesmata, whose structure is characteristic in vascular plants. Incidentally, during the revision process of this work, we found out that a recent article had also described the presence of MP homologues in a variety of non-flowering plants (Mushegian and Elena, 2015).

Altogether, these studies show that *Caulimoviridae* not only had a major impact on plant populations since the emergence of vascular plants, but also that *Caulimoviridae* EVEs likely impacted genome evolution over a similar range of time.

GIANT VIRUSES, VIROPHAGES AND TRANSPOVIRONS

While *Caulimoviridae* are known plant pathogens, I wanted to know whether paleovirology could help revealing the existence of yet unknown plant viruses. Nucleocytoplasmic large DNA viruses (NCLDVs) are eukaryotic viruses with big genomes (100 kb–2.5 Mb). NCLDVs are known to infect animals, protists and phytoplankton but have never been described as pathogens of land plants. Teaming with my colleague Guillaume Blanc (CNRS, France), we have screened the genomes from thirteen land plants for the presence of footprints of genetic elements that potentially originate from NCLDV genomes. Using BLAST, we have compared for each plant protein the best score obtained against NCLDV proteins versus the best score obtained against eukaryotic or prokaryotic proteins. Unexpectedly, we found that the genome of the bryophyte (moss) *Physcomitrella patens* contains clusters of genes with high phylogenetic affinities to NCLDV homologues. This surprising finding suggests that extent NCLDVs are/were capable to prey upon plants, thereby expanding both the known host range for this virus family and the diversity of the plant virosphere. Interestingly, the moss NCLDV-like elements are much smaller than known NCLDVs, revealing a less complex, perhaps ancestral form of NCLDV genomes (Maumus et al., 2014). Overall, the reconstructed region spans over 13kb and contains 20 original ORFs. None of the extant NCLDV-like regions contains a full complement of 20 ORFs.

Altogether, these results strengthen the relevance of applying paleovirology approaches to plant genomes. Similar studies looking for traces of other types of viruses will help characterizing the current and past plant virosphere and may eventually help establishing connections between the selective pressure applied by viruses over plants and important evolutionary transitions. For instance, the prolonged haploid stage in the life cycle of mosses may constitute a period of vulnerability to viral insertions; whereas the microspore haploid stage is relatively short during the reproduction cycle of seed plants. Together, genomics-enabled paleovirology studies stress the need for a profound investigation of EVE types and potential biological functions in plants.

NCLDVs are known to prey upon a wide variety of algae. Owing to their control over algal populations, NCLDVs have significant impacts on biogeochemical cycles. Virophages are recently discovered virus satellites that prey on NCLDVs and the infection of NCLDVs by virophages has been shown to limit the production of NCLDV particles, accompanied by greater survival of the eukaryotic

hosts (La Scola et al., 2008). Although these entities are probably widely distributed in the microbial world, the underlying nested parasitic relationships remain poorly understood.

I was curious to get more insights into these microbial wars and the possible modes of adaptation of such tri-partite (NCLDV – virophage – eukaryote) relationships. Using paleovirology approaches, we revealed that the genome of the single cell alga *Bigelowiella natans* encloses the footprints of molecular battles between viruses and their molecular parasites (Blanc et al., 2015). Indeed, we found that this algal genome contains several copies of integrated virophages showing up as large drops of G+C content in the *B. natans* genomic assembly contigs (Figure 10).

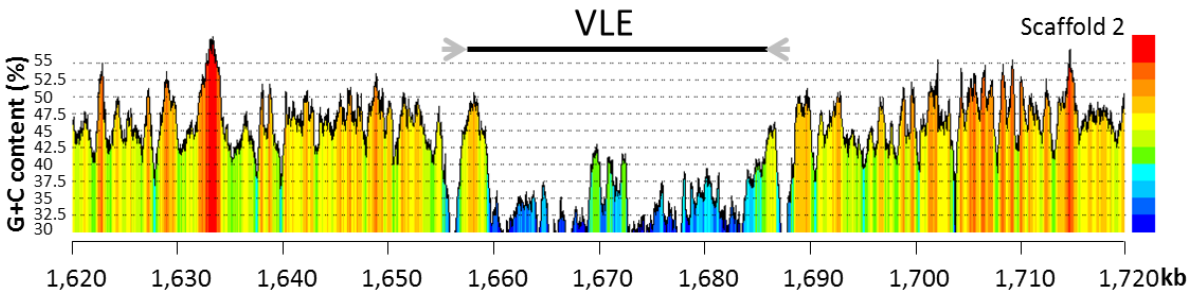

Figure 10: G+C content in the region of the largest virophage-like element (VLE) in the *B. natans* genome.

Interestingly, the endogenous virophage copies appear to be functional as evidenced by the conservation of open reading frames and significant transcription levels. Because virophages hamper NCLDV multiplication, we proposed these endogenized virophages might be beneficial to both partners by increasing the chances of coinfection between virophages and NCDLV preys and by defending eukaryotic hosts from fatal NCLDV infections.

In addition, although no infectious NCLDV is known for *B. natans*, we detected large additional inserts of NCDLV origin in this genome. These traces testify of the passage of NCLDV particles within the intracellular space of the alga, thereby strongly suggesting that the alga is a prey for marine NCLDV. Furthermore, we discovered that the algal genome also contains SGEs showing significant synapomorphy and phylogenetic affinities with a novel family of transposable elements (*i.e.* the Transpovirons) that parasitize NCLDVs (Desnues et al., 2012). Altogether our findings show that the genome of this alga provides a unique genetic record testifying of the complex, entangled parasitic and symbiotic relationships that regulate microbial communities.

These findings bring significant novelty to at least two areas of intense research. Firstly, it introduces a new dimension in the understanding of the interactions within microbial communities by suggesting the complex roles that SGEs can play in regulating algal and viral populations. Secondly, it brings substantial insights to our comprehension of the evolutionary links between virophages and another type of SGEs, called Polintons that are present in the genomes of various eukaryotes and that show structural and compositional similitudes to virophages (Fischer and Suttle, 2011).

SELF-ASSESSMENT

Sometimes I like to think of genomes like attic sales: it takes your curiosity; there is so much stuff in there. Many objects you are familiar with, some others need a closer look to be identified, and a few are invaluable. I have conducted a variety of projects in a large number of species. Too many? Not in terms of publications, but perhaps in terms of expertise. After addressing this and this and that, it is only in the last years that I am finding my deep interests and becoming capable to ask more fundamental questions; which appear so obvious and general in the end!

How do TEs evolve? That sounds like a naïve question Google would answer in a few seconds. And then finding out there is no definitive answer, bits of results here and there. I realize this question is as complex as “How did species evolve”? After collecting some pictures of human and some of chimps, it is really not clear why human rule most of the world.

I have collected several pictures. And now I am asking myself more general questions that allow framing my research into perspective. The keywords from my past research are easy to spot: Transposable elements, endogenous viruses, and epigenetics. These remain my main interests but they now get the tint of Evolutionary Biology.

I have been taken into a vortex of discoveries and new ideas. My objective is now to focus on a limited number of more holistic, possibly interconnected, projects and to address these with robust experimental design. Now I need to put efforts into the construction of a solid line of research that would federate in the team. Together, we can combine our skills so as to aim at collecting desired pieces to build a gallery that will illuminate the natural history of SGEs.

MENTORING

I have supervised two M2 students and I am currently supervising one postdoc and co-supervising a PhD student.

Ophélie Jouffroy, in 2015, joined our team in the context of an M2 internship to study an emerging model species with fleshy fruits, tomato. I had performed a comprehensive annotation of the tomato genome and Ophélie's project was to investigate the organization of the tomato genomes and the potential impact of genomic repeats on the regulation of gene expression, especially in the context of fruit ripening. Under my supervision, Ophélie has performed a series of analyses leading to several remarkable observations. She found, for instance, that bins of the tomato genome can be sorted into three compartments on the basis of repeat density, which is inversely correlated with gene density. Interestingly, she found that the different compartments present distinct enrichment of different types of TEs. For instance, the regions of intermediate repeat density (typically found at the edges of pericentromeres) are enriched for Copia-type retrotransposons, therefore establishing that repeat density also correlates with repeat content in this species. She also found the genes within the different compartments to have contrasting expression levels that are inversely correlated with repeat density and to be enriched for distinct gene ontologies.

Ophélie has then successfully applied for a PhD fellowship (Oct. 2015 - Sept. 2018) to continue her work on tomato under my supervision, our team leader – Hadi Quesneville – being co-supervisor. She first continued along her M2 project and she published her work in BMC Genomics as first author (Jouffroy et al., 2016).

Issa Seydina Diop joined URGI in 2017 for an M2 internship under my supervision. He specifically worked on the *Caulimoviridae* macroevolution project. He performed the phylogenetic analysis of *Caulimoviridae* genomes and EVEs. He has published this study in Scientific Reports as first author (Diop et al., 2017). After that, Issa has obtained a PhD scholarship in Switzerland.

Vikas Sharma recently joined (Feb. 2018) for a postdoc that is funded by ANR project EVENTS coordinated by Pierre-Yves Teycheney (CIRAD, France).His goal during his 18 months contract will be to set up pipelines for the detection of EVEs and to use these to address the complexity of the *Caulimoviridae* genetic network (see more details in project below).

These first mentoring activities have been very instructive. It requires some adaptation and time to understand someone's objectives and capacities, to measure a level of autonomy. Probably most of the time, one cannot truly catch the adequacy of the selected candidate with the project. The best candidate is welcome. It is then important, sometime after project starts, to evaluate and to readjust the project to the profile. It has been so far a pleasure to interact with each of them and a chance to get to know them more personally. My objective as a mentor is that people find great interest in their research project and that they acquire a biology-driven approach in bioinformatics.

RESEARCH PROJECTS

MACRO- AND MICRO-EVOLUTION OF TRANSPOSABLE ELEMENTS IN PLANTS

The evolutionary trajectories of TEs remain poorly understood. How is a TE superfamily able to co-evolve with its host species? This is the kind of question I would like to investigate in the next years.

Microevolution – It has been very challenging to address the microevolution of TEs in the past, mostly because TEs are rapidly evolving sequences and the plants which genomes are available often belong to branches that are too divergent to identify conserved TE families across two or more species. As a result, when identifying a TE sequence, it is difficult to understand what regions are constrained, novel, or lost besides protein-coding ones. Hence, conservation can only be revealed in the light of evolution so that it is necessary to compare several sequences from one TE family to allow addressing evolutionary trends. Tightening the evolutionary distances between sequenced plant genomes, the Brassicales Map Alignment Program (BMAP) enables addressing TE evolution at unprecedented resolution. As described above, analysis of the BMAP repeatome led to the discovery of few TE families that are conserved in a significant number of species and transmitted vertically as inferred using phylogenetic reconstruction. *Habitans*, for instance, is providing a first case of conserved TE family across a plant family. It offers the opportunity to study various facets of TE evolution at fine scale. I aim at taking advantage of the BMAP data to address the microevolution of TEs, specifically LTR retrotransposons. This data will allow addressing both the evolution of coding and non-coding sequences. For instance, the LTRs of Habitans show limited conservation between species. LTRs are the regions allowing transcriptional regulation of LTR retrotransposons and retroviruses. It is tempting to speculate that accelerated evolution of LTRs reflects changes in host biology and/or environment. It comprises the promoter which contains transcription factor binding sites (TFBS) and that will largely determine when transcriptional activation takes place. It is possible that such regulatory regions could gradually co-evolve with the host TFBS to maintain some regulatory traits. The changes between LTRs of different Copia-type families can however be way more severe than would be the case by simply tuning or replacing existing TFBS. Extreme variation of size and sequence can be found among LTRs within TE superfamilies which might reflect conversion to new regulatory modes and this kind of profound sequence variation requires substantial genetic novelty. One can hypothesize that sequence capture, for instance from viruses or bacteria, are a source of novel genetic sequences serving the turnover of LTRs in a process that may be in part supported by the capacity of LTR retrotransposons of shuttling between cytoplasm and nucleus along their replicative cycle and hence to interact with free RNA and DNA from a range of endogenous entities. A first example of LTRs sharing similarity with bacterial proteins was presented in 2017 with the Hodor element in apple tree genome (Daccord et al., 2017).

The analysis of Habitans elements and two other conserved families of Copia-type elements identified in *Brassicaceae* will likely provide an assortment of insights regarding TE evolution. The evolution of non-coding regions can be investigated in a variety of contexts. The significance of highly variable or highly constrained regions and motifs should be investigated to gather knowledge regarding the transcriptional and translational regulation. The evolution of LTRs, for instance, can be further qualified by the analysis of their different regions (i.e. U5, R, and U3). As described above, the alignment of non-coding regions of *Habitans* elements was very variable depending on the alignment program used, and more specifically on the way potential unrelated segments are treated, i.e. by

rather forcing alignment or by permitting gaps. Allowing so, a more thorough and modular characterization of LTRs should be possible enabling to identify conserved regions and motifs. Alignment-free search for conserved motifs can also be run independently. Conserved and variable sequences should be compared to a comprehensive repertoire of transcription factor binding sites (TFBS): the *A. thaliana* cistrome database (O'Malley et al., 2016), to address the evolution of U5 with TFBS.

The impact of DNA methylation can be addressed at the level of the whole TE sequence. TEs are generally G+C-poor which is thought to be the cause of deamination of methylcytosines and to reflect selection for limiting DNA methylation levels allowing balanced fitness. This process hence imposes strong evolutionary constraints on TEs that can be investigated by comparing G+C content and more specifically the distribution of different methylation contexts across conserved TEs.

TE sequences are minted by this complex assortment of evolutionary constraints to subsist (Figure 11). Altogether, I am convinced that the study of conserved TE families can provide insights regarding several levels of TE regulation including epigenetic, transcriptional, post-transcriptional, and translational as well as regarding their replication process. Further resolution in this kind of analysis could be met by sequencing whole genomes from a number of additional *Brassicaceae* species standing at pivotal phylogenetic positions.

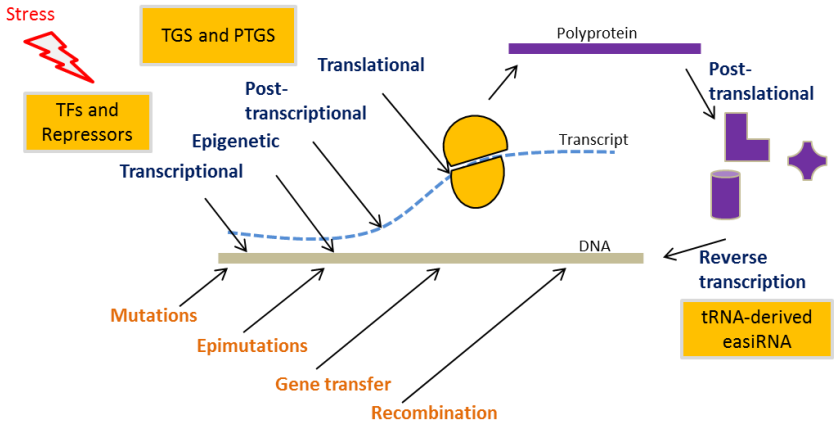

Figure 11: Schematic view of the TE forge. TE sequences are under a range of selective constraints governing all the steps of their own replication and imposed by the host biology and evolution.

Another central question follows: what could explain why few TEs families are conserved across species? Have they evolved superior survival skills as compared to other TEs? Or do *Habitans* elements confer any beneficial function to their hosts? They are relatively low-copy number elements and a detailed analysis of the functional annotation and gene ontology of nearby genes could provide some hypotheses to be further investigated. A hypothetical function could also be searched by addressing the assortment of *Habitans*-derived RNAs (long or small) and its potential function of trans-regulation of gene expression.

Such a microevolution project has been the subject of a preliminary analysis which is very encouraging. The perimeter has been determined and the approaches to be employed are sufficiently defined to propose a PhD project in the next years. Funding to generate more sequences could be obtained through internal INRA BAP division call or through “France génomique” proposal.

Macroevolution - The taxonomic distribution of TEs across eukaryotes suggests that the main superfamilies (for instance LINEs, SINEs, Ty1/Copia or Ty3/Gypsy) were present in the genome of the last ancestor common of eukaryotic supergroups.

What lineages are most successful, when did they emerge? What are the reasons and consequences of linage success? Current classification in genome analysis is commonly limited to the Superfamily level, i.e. the results indicate if Gypsy versus Copia is most abundant or documents their relative activity over time. By contrast, the contribution of specific TE families has been mostly addressed at the level of single species, providing insufficient evolutionary perspective.

I wish to address the TE assortment in plant genomes in an evolutionary context, supported by phylogenetic and sequence composition analyses.

The number of species with publicly available plant genomes nears a hundred and fifty or so; ranging from Chlorophytes (green algae) to Angiosperms (flowering plants) and comprising members from a variety of groups including Bryophytes (e.g. mosses), Pteridophytes (e.g. ferns) and seed plants. Owing to an array of economic and scientific motivations, sampling density is however severely biased towards Angiosperms with few families such as *Brassicaceae*, *Poaceae*, and *Solanaceae* being relatively densely sampled. Furthermore, I was recently named to be leader PI for the TE analysis of a new project named Open Green Genomes (OGG) funded by Joint Genome Institute (JGI) and coordinated by Jim Leebens-Mack (University of Georgia, USA). OGG project aims at sequencing 35 land plant genomes that were carefully selected to fill several phylogenetic sampling gaps across Viridiplantae. Thus, the number and phylogenetic breadth of sequenced plant genomes now allows beginning to investigate the macroevolution of TEs along the evolution of Viridiplantae.

To observe TE macroevolution, one needs to address a significant number of plant genomes from a number of families, including representatives of deep plant lineages. Using REPET on the robust computer cluster available at URGI, I aim at generating repeat libraries from at least a hundred plant genomes encompassing members from green algae to dicotyledons so as to collect an atlas for Viridiplantae TEs. Direct search for conserved TE domains should also be employed. All transposable elements found will be classified into superfamilies, families, and subfamilies. For each TE superfamily, conserved domains will be retrieved from each plant genome. Using similarity-based clustering, redundancy will be eliminated and families will be established on the basis of phylogenetic reconstruction. Each copy from each plant genome will then be assigned to a TE family using phylogenetic placement. As for gene families, gain and loss of TEs in specific branches will be inferred from their taxonomic distribution. Pilot analyses are being run on the set of *Brassicaceae* genomes. The ups and downs of TEs will be analyzed in the context of their sequence specificities and in regard of the repertoire of host genes with known impact on TE regulation. The Viridiplantae TE library should provide valuable ground for a variety of evolutionary analyses such as addressing the diversity of SGEs in basal plant lineages and the frequency of horizontal TE transfer, and will provide reference sequences for genome annotation for instance from the generation of high quality HMM profiles. This Atlas should be lodged into RepetDB, a specific database recently developed at URGI.

Such project would require a diversity of competencies such as database management & system administration, large scale phylogenetic analysis, and HTML coding to go from the core search to the distribution of data and results. Such a large scale project would probably develop into a roadmap for the team and enroll several members with different expertise and interests. To be fully implemented,

such a project should be submitted in the coming years to French and/or European research agencies for funding so as to involve external collaborators and students or postdocs .

ENDOGENOUS VIRUSES

A major conclusion stemming out from previous work is that, instead of representing a single element, endogenous *Caulimoviridae* often appear to define complex networks of related sequences with structural and genetic variants. For instance, *Caulimoviridae* EVEs can be re-organized in plant genomes into sequences capable to generate episomal infection (Gayral et al., 2008). In addition, several *Caulimoviridae* genera have bipartite genomes (Geering et al., 2014) - each viral genome is defined by two "chromosomes" - but it is unclear whether this genomic organization has emerged once or repeatedly. Furthermore, when annotating EVEs in the oak (*Quercus robur*) genome, we found a very high number of short *Caulimoviridae* segments with movement protein domain relative to the number of full length viral genomes (Plomion et al., accepted in Nature Plants), suggesting that they could represent by-products or satellite sequences which duplication would rely on proteins from autonomous entities, either exogenous *Caulimoviridae*, or endogenous elements.

Combining information from several endogenous copies enables reconstructing consensus sequences that closely determine native elements. To disentangle the genetic relationships and functional interactions defining these networks, we will perform an accurate reconstruction of *Caulimoviridae* and related EVE sequences in several plant genomes, starting with species such as oak, vine, and citrus, for which we have collected exemplary data. The acquired methodology will be automatized and transposed to dozens of plant genomes to identify, characterize and compare *Caulimoviridae* sequence networks and we will distribute a pipeline named Caulifinder (Figure 12) allowing the identification and annotations of *Caulimoviridae* EVEs in plant genomes. Comparing *Caulimoviridae* networks across plant families with high EVE content and for which a substantial set of genomes are publicly available (e.g. *Solanaceae*) or to which we have access in the context of collaborations (e.g. *Rutaceae*) will allow determining times of emergence and performing detailed co-evolution analyses to gain further insights into the evolution of this complex plant pathosystem.

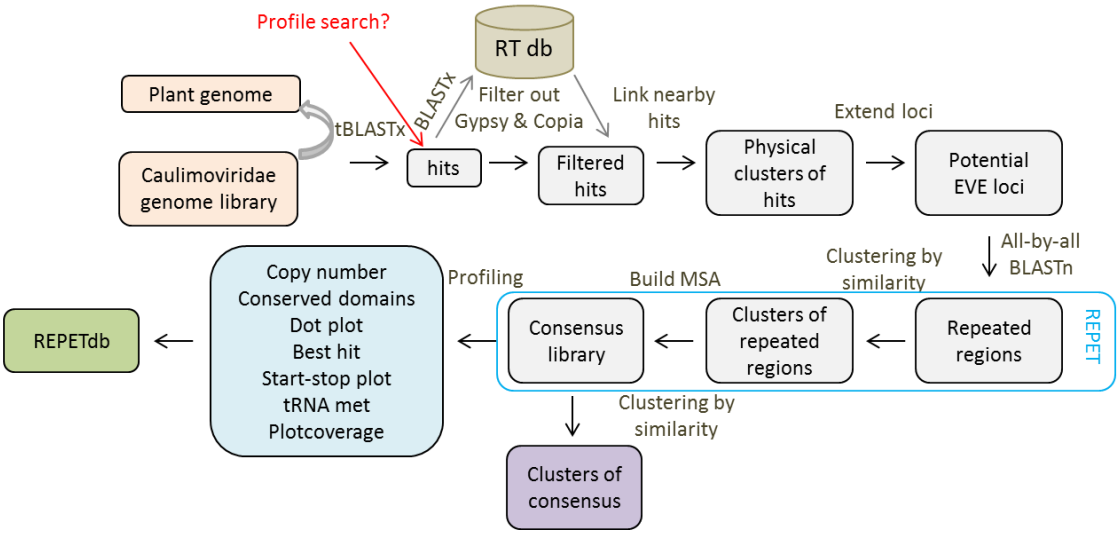

Figure 12: Graphical illustration of the main steps of the Caulifinder workflow

EVEs can be of different types for instance full length or concatemers, rearranged or intact, with conserved ends between copies, or not. The detailed analysis of EVE copies related to specific genera is likely to provide information regarding virology and impacts on host biology. Considering Florendovirus for instance, previous analysis of their EVEs has suggested that their genomes are bipartite, with evocative evidence that both genomes interact together at some point of their cycle being trapped in the vine genome in the form of a number of loci presenting clusters containing a complex arrangement of both parts. Footprints of this physical interaction should be assessed in several plant genomes and this kind of study should be extended to discover more about host-virus interactions. As example, the replication of viral DNA commonly uses several origins and, as a result, the dsDNA molecules contain a small number of single-strand nicks. These represent potential fragile sites that could be preferential points of genome break causing genome linearization. Following this hypothesis, some EVEs could present conserved ends that would reflect specific stages of the replication of viral DNA. Furthermore, *Caulimoviridae* can produce several transcripts, typically one which is reverse-transcribed representing the full genome, and a few others corresponding to specific ORFs called subgenomic RNA (sgRNA). As a result, it is conceivable that highly repeated EVEs that correspond to fragments of viral genomes may represent the product of reverse-transcribed sgRNA that integrated plant genomes at high frequency relative to full genomes. The analysis of EVE flanks could also provide evidence regarding integration mechanisms of different types of EVEs, e.g. active or passive. For instance, (TA)n repeats could be the hallmark of full genome integration while satellites sequences may not, which could mirror distinct paths into host genome.

I am now work package leader of an ANR-funded project coordinated by Pierre-Yves Teycheney (CIRAD, France) called EVENTS which aims at studying EVEs towards characterizing virus evolution and potential EVE functions in plant biology. In this framework, I am supervising a postdoctoral researcher for 18 months who will work on benchmarking Caulifinder and address EVEs in a virus and plant biology perspective.

! Because this document will become public, a more detailed research project will be presented privately during the viva voce.

CONCLUSION

One may wonder why I have not received any grant as principal investigator so far. I tried hard, though. For the last three years I have been submitting and re-submitting a fascinating biotechnology project. Shall I say that I almost got it funded? It has even been selected for the last step (the Brussels interview) of the European Research Council in the “starting grant” category. Not being funded is somehow the norm. But I think I also failed at the strategical level because pushing this project did not really help me moving forward in my daily research nor did it foster team spirit.

The project was aiming to develop drugs and protocols for plant mutagenesis by inducing the transposition of endogenous TEs. It was hence a project that technically implies a lot of bench work, plant culture, high throughput screening of chemical compounds, etc. I had this desire of going back to the bench, to molecular biology, to “live” plants. And that meant one step out of my team (which is purely bioinformatics). But I do have this double competency so it would not stop me. However, it obviously failed at federating within the team so I was almost on my own on a project with so many components.

That project required very strong expertise in epigenetics and the literature in this area of research is difficult to follow, even when considering Arabidopsis only. So it was very much related to my concrete work but at the same time disconnected from it as requiring different expertise than that I was actually acquiring in bioinformatics and evolutionary biology. Hence it was a lot of efforts writing down these proposals and very challenging to have it up to date. New information was also not always so relevant to my actual studies so there was in the end substantial loss of time in this effort. At this stage this project has recently lost some novelty and I don’t know whether I should invest anymore in this topic.

I probably should discard this and revise my strategy. My main interests have also slightly shifted in the last years. What is awesome with bioinformatics is that you can do a lot without receiving external funding. Public databases are gigantic and our computing cluster in place offers a lot of resources. Projects related to the evolution of plant SGEs fit very well with the landscape of our team, both at the level of objectives and of human competencies. This area of research would combine our collective skills and could be integrated in our roadmap to become a structuring project in the team and unit for the next years. It is at the core of what I am most curious about in Biology and it represents long term goals that can be followed by the team and possibly boosted by grants, fellows and students.

Finally, aside from managing and writing proposals, I wish I can also go on doing what I like most; exploring, browsing, mining… new ideas come up from down there!

REFERENCES

AIEWSAKUN, P. & KATZOURAKIS, A. 2017. Marine origin of retroviruses in the early Palaeozoic Era. *Nat Commun,* 8**,** 13954.

BLANC, G., GALLOT-LAVALLÉE, L. & MAUMUS, F. 2015. Provirophages in the Bigelowiella genome bear testimony to past encounters with giant viruses. *Proceedings of the National Academy of Sciences,* 112**,** E5318-E5326.

CAO, J., SCHNEEBERGER, K., OSSOWSKI, S., GUNTHER, T., BENDER, S., FITZ, J., KOENIG, D., LANZ, C., STEGLE, O., LIPPERT, C., WANG, X., OTT, F., MULLER, J., ALONSO-BLANCO, C., BORGWARDT, K., SCHMID, K. J. & WEIGEL, D. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet,* 43**,** 956-63.

CHOI, K., ZHAO, X., KELLY, K. A., VENN, O., HIGGINS, J. D., YELINA, N. E., HARDCASTLE, T. J., ZIOLKOWSKI, P. A., COPENHAVER, G. P., FRANKLIN, F. C., MCVEAN, G. & HENDERSON, I. R. 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat Genet,* 45**,** 1327-36.

DACCORD, N., CELTON, J. M., LINSMITH, G., BECKER, C., CHOISNE, N., SCHIJLEN, E., VAN DE GEEST, H., BIANCO, L., MICHELETTI, D., VELASCO, R., DI PIERRO, E. A., GOUZY, J., REES, D. J. G., GUERIF, P., MURANTY, H., DUREL, C. E., LAURENS, F., LESPINASSE, Y., GAILLARD, S., AUBOURG, S., QUESNEVILLE, H., WEIGEL, D., VAN DE WEG, E., TROGGIO, M. & BUCHER, E. 2017. High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. *Nat Genet,* 49**,** 1099-1106.

DESNUES, C., LA SCOLA, B., YUTIN, N., FOURNOUS, G., ROBERT, C., AZZA, S., JARDOT, P., MONTEIL, S., CAMPOCASSO, A., KOONIN, E. V. & RAOULT, D. 2012. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci U S A,* 109**,** 18078-83.

DIOP, S., GEERING, A. D., ALFAMA-DEPAUW, F., LOAEC, M., TEYCHENEY, P.-Y. & MAUMUS, F. 2017. Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *bioRxiv***,** 158972.

FISCHER, M. G. & SUTTLE, C. A. 2011. A virophage at the origin of large DNA transposons. *Science,* 332**,** 231-4.

FUTTERER, J., KISS-LASZLO, Z. & HOHN, T. 1993. Nonlinear ribosome migration on cauliflower mosaic virus 35S RNA. *Cell,* 73**,** 789-802.

GEERING, A. D., MAUMUS, F., COPETTI, D., CHOISNE, N., ZWICKL, D. J., ZYTNICKI, M., MCTAGGART, A. R., SCALABRIN, S., VEZZULLI, S. & WING, R. A. 2014. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nature communications,* 5.

HAYWARD, A., CORNWALLIS, C. K. & JERN, P. 2015. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc Natl Acad Sci U S A,* 112**,** 464-9.

HOEN, D. R., HICKEY, G., BOURQUE, G., CASACUBERTA, J., CORDAUX, R., FESCHOTTE, C., FISTON-LAVIER, A.-S., HUA-VAN, A., HUBLEY, R. & KAPUSTA, A. 2015. A call for benchmarking transposable element annotation methods. *Mobile DNA,* 6**,** 13.

HU, T. T., PATTYN, P., BAKKER, E. G., CAO, J., CHENG, J. F., CLARK, R. M., FAHLGREN, N., FAWCETT, J. A., GRIMWOOD, J., GUNDLACH, H., HABERER, G., HOLLISTER, J. D., OSSOWSKI, S., OTTILAR, R. P., SALAMOV, A. A., SCHNEEBERGER, K., SPANNAGL, M., WANG, X., YANG, L., NASRALLAH, M.

E., BERGELSON, J., CARRINGTON, J. C., GAUT, B. S., SCHMUTZ, J., MAYER, K. F., VAN DE PEER, Y., GRIGORIEV, I. V., NORDBORG, M., WEIGEL, D. & GUO, Y. L. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat Genet,* 43**,** 476-81.

HUANG, C. H., SUN, R., HU, Y., ZENG, L., ZHANG, N., CAI, L., ZHANG, Q., KOCH, M. A., AL-SHEHBAZ, I., EDGER, P. P., PIRES, J. C., TAN, D. Y., ZHONG, Y. & MA, H. 2016. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. *Mol Biol Evol,* 33**,** 394-412.

JARVIS, E. E., DUNAHAY, T. G. & BROWN, L. M. 1992. DNA NUCLEOSIDE COMPOSITION AND METHYLATION IN SEVERAL SPECIES OF MICROALGAE. *Journal of Phycology,* 28**,** 356-362.

JOUFFROY, O., SAHA, S., MUELLER, L., QUESNEVILLE, H. & MAUMUS, F. 2016. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics,* 17**,** 624.

LA SCOLA, B., DESNUES, C., PAGNIER, I., ROBERT, C., BARRASSI, L., FOURNOUS, G., MERCHAT, M., SUZAN-MONTI, M., FORTERRE, P., KOONIN, E. & RAOULT, D. 2008. The virophage as a unique parasite of the giant mimivirus. *Nature,* 455**,** 100-4.

LANG, D., ULLRICH, K. K., MURAT, F., FUCHS, J., JENKINS, J., HAAS, F. B., PIEDNOEL, M., GUNDLACH, H., VAN BEL, M., MEYBERG, R., VIVES, C., MORATA, J., SYMEONIDI, A., HISS, M., MUCHERO, W., KAMISUGI, Y., SALEH, O., BLANC, G., DECKER, E. L., VAN GESSEL, N., GRIMWOOD, J., HAYES, R. D., GRAHAM, S. W., GUNTER, L. E., MCDANIEL, S. F., HOERNSTEIN, S. N. W., LARSSON, A., LI, F. W., PERROUD, P. F., PHILLIPS, J., RANJAN, P., ROKSHAR, D. S., ROTHFELS, C. J., SCHNEIDER, L., SHU, S., STEVENSON, D. W., THUMMLER, F., TILLICH, M., VILLARREAL AGUILAR, J. C., WIDIEZ, T., WONG, G. K., WYMORE, A., ZHANG, Y., ZIMMER, A. D., QUATRANO, R. S., MAYER, K. F. X., GOODSTEIN, D., CASACUBERTA, J. M., VANDEPOELE, K., RESKI, R., CUMING, A. C., TUSKAN, G. A., MAUMUS, F., SALSE, J., SCHMUTZ, J. & RENSING, S. A. 2018. The Physcomitrella patens chromosome-scale assembly reveals moss genome structure and evolution. *Plant J,* 93**,** 515-533.

LIPPMAN, Z., GENDREL, A. V., COLOT, V. & MARTIENSSEN, R. 2005. Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods,* 2**,** 219-24.

LLORENS, C., FUTAMI, R., COVELLI, L., DOMÍNGUEZ-ESCRIBÁ, L., VIU, J. M., TAMARIT, D., AGUILAR-RODRÍGUEZ, J., VICENTE-RIPOLLES, M., FUSTER, G. & BERNET, G. P. 2010. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic acids research,* 39**,** D70-D74.

MALLORY, A. C. & VAUCHERET, H. 2009. ARGONAUTE 1 homeostasis invokes the coordinate action of the microRNA and siRNA pathways. *EMBO Rep,* 10**,** 521-6.

MARI-ORDONEZ, A., MARCHAIS, A., ETCHEVERRY, M., MARTIN, A., COLOT, V. & VOINNET, O. 2013. Reconstructing de novo silencing of an active plant retrotransposon. *Nat Genet,* 45**,** 1029-39.

MAUMUS, F., ALLEN, A. E., MHIRI, C., HU, H., JABBARI, K., VARDI, A., GRANDBASTIEN, M.-A. & BOWLER, C. 2009. Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC genomics,* 10**,** 624.

MAUMUS, F., EPERT, A., NOGUÉ, F. & BLANC, G. 2014. Plant genomes enclose footprints of past infections by giant virus relatives. *Nature communications,* 5.

MAUMUS, F. & QUESNEVILLE, H. 2014a. Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. *Nat Commun,* 5**,** 4104.

MAUMUS, F. & QUESNEVILLE, H. 2014b. Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PloS one,* 9**,** e94101.

MAUMUS, F. & QUESNEVILLE, H. 2016. Impact and insights from ancient repetitive elements in plant genomes. *Current opinion in plant biology,* 30**,** 41-46.

MURAT, F., LOUIS, A., MAUMUS, F., ARMERO, A., COOKE, R., QUESNEVILLE, H., CROLLIUS, H. R. & SALSE, J. 2015. Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome biology,* 16**,** 262.

MUSHEGIAN, A. R. & ELENA, S. F. 2015. Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. *Virology,* 476**,** 304-15.

O'MALLEY, R. C., HUANG, S. C., SONG, L., LEWSEY, M. G., BARTLETT, A., NERY, J. R., GALLI, M., GALLAVOTTI, A. & ECKER, J. R. 2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell,* 166**,** 1598.

OSSOWSKI, S., SCHNEEBERGER, K., LUCAS-LLEDO, J. I., WARTHMANN, N., CLARK, R. M., SHAW, R. G., WEIGEL, D. & LYNCH, M. 2010. The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science,* 327**,** 92-4.

OYAMA, R. K., CLAUSS, M. J., FORMANOVA, N., KROYMANN, J., SCHMID, K. J., VOGEL, H., WENIGER, K., WINDSOR, A. J. & MITCHELL-OLDS, T. 2008. The shrunken genome of Arabidopsis thaliana. *Plant Systematics and Evolution*

273**,** 257-271.

PRINGLE, C. R. 1998. The universal system of virus taxonomy of the International Committee on Virus Taxonomy (ICTV), including new proposals ratified since publication of the Sixth ICTV Report in 1995. *Arch Virol,* 143**,** 203-10.

RENSING, S. A., LANG, D., ZIMMER, A. D., TERRY, A., SALAMOV, A., SHAPIRO, H., NISHIYAMA, T., PERROUD, P. F., LINDQUIST, E. A., KAMISUGI, Y., TANAHASHI, T., SAKAKIBARA, K., FUJITA, T., OISHI, K., SHIN, I. T., KUROKI, Y., TOYODA, A., SUZUKI, Y., HASHIMOTO, S., YAMAGUCHI, K., SUGANO, S., KOHARA, Y., FUJIYAMA, A., ANTEROLA, A., AOKI, S., ASHTON, N., BARBAZUK, W. B., BARKER, E., BENNETZEN, J. L., BLANKENSHIP, R., CHO, S. H., DUTCHER, S. K., ESTELLE, M., FAWCETT, J. A., GUNDLACH, H., HANADA, K., HEYL, A., HICKS, K. A., HUGHES, J., LOHR, M., MAYER, K., MELKOZERNOV, A., MURATA, T., NELSON, D. R., PILS, B., PRIGGE, M., REISS, B., RENNER, T., ROMBAUTS, S., RUSHTON, P. J., SANDERFOOT, A., SCHWEEN, G., SHIU, S. H., STUEBER, K., THEODOULOU, F. L., TU, H., VAN DE PEER, Y., VERRIER, P. J., WATERS, E., WOOD, A., YANG, L., COVE, D., CUMING, A. C., HASEBE, M., LUCAS, S., MISHLER, B. D., RESKI, R., GRIGORIEV, I. V., QUATRANO, R. S. & BOORE, J. L. 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science,* 319**,** 64-9.

VELUCHAMY, A., LIN, X., MAUMUS, F., RIVAROLA, M., BHAVSAR, J., CREASY, T., O'BRIEN, K., SENGAMALAY, N. A., TALLON, L. J. & SMITH, A. D. 2013. Insights into the role of DNA methylation in diatoms by genome-wide profiling in Phaeodactylum tricornutum. *Nature Communications,* 4**,** ncomms3091.

WANG, X., WEIGEL, D. & SMITH, L. M. 2013. Transposon variants and their effects on gene expression in Arabidopsis. *PLoS Genet,* 9**,** e1003255.

ZLOTORYNSKI, E., RAHAT, A., SKAUG, J., BEN-PORAT, N., OZERI, E., HERSHBERG, R., LEVI, A., SCHERER, S. W., MARGALIT, H. & KEREM, B. 2003. Molecular basis for expression of common and rare fragile sites. *Mol Cell Biol,* 23**,** 7143-51.