HAL open science

Genetic and metabolic signatures of Salmonella enterica subsp. enterica associated with animal sources at the pangenomic scale

Méryl Vila Nova, Kévin Durimel, Kévin La, Arnaud Felten, Philippe Bessieres, Michel-Yves Mistou, Mahendra Mariadassou, Nicolas Radomski

To cite this version:

Méryl Vila Nova, Kévin Durimel, Kévin La, Arnaud Felten, Philippe Bessieres, et al.. Genetic and metabolic signatures of Salmonella enterica subsp. enterica associated with animal sources at the pangenomic scale. BMC Genomics, 2019, 20 (1), 10.1186/s12864-019-6188-x . hal-02392205

HAL Id: hal-02392205

https://hal.science/hal-02392205

Submitted on 3 Dec 2019

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

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

Distributed under a Creative Commons Attribution 4.0 International License

Vila Nova *et al. BMC Genomics* (2019) 20:814
https://doi.org/10.1186/s12864-019-6188-x

BMC Genomics

RESEARCH ARTICLE Open Access

Genetic and metabolic signatures of *Salmonella enterica* subsp. *enterica* associated with animal sources at the pangenomic scale

Meryl Vila Nova[1,2], Kévin Durimel[1], Kévin La[1], Arnaud Felten[1], Philippe Bessières[2], Michel-Yves Mistou[1], Mahendra Mariadassou[2] and Nicolas Radomski[1*]

Abstract

Background: *Salmonella enterica* subsp. *enterica* is a public health issue related to food safety, and its adaptation to animal sources remains poorly described at the pangenome scale. Firstly, serovars presenting potential mono- and multi-animal sources were selected from a curated and synthetized subset of Enterobase. The corresponding sequencing reads were downloaded from the European Nucleotide Archive (ENA) providing a balanced dataset of 440 *Salmonella* genomes in terms of serovars and sources (i). Secondly, the coregenome variants and accessory genes were detected (ii). Thirdly, single nucleotide polymorphisms and small insertions/deletions from the coregenome, as well as the accessory genes were associated to animal sources based on a microbial Genome Wide Association Study (GWAS) integrating an advanced correction of the population structure (iii). Lastly, a Gene Ontology Enrichment Analysis (GOEA) was applied to emphasize metabolic pathways mainly impacted by the pangenomic mutations associated to animal sources (iv).

Results: Based on a genome dataset including *Salmonella* serovars from mono- and multi-animal sources (i), 19,130 accessory genes and 178,351 coregenome variants were identified (ii). Among these pangenomic mutations, 52 genomic signatures (iii) and 9 over-enriched metabolic signatures (iv) were associated to avian, bovine, swine and fish sources by GWAS and GOEA, respectively.

Conclusions: Our results suggest that the genetic and metabolic determinants of *Salmonella* adaptation to animal sources may have been driven by the natural feeding environment of the animal, distinct livestock diets modified by human, environmental stimuli, physiological properties of the animal itself, and work habits for health protection of livestock.

Keywords: Microbial genomics, *Salmonella* adaptation, Genome wide association study, Gene ontology enrichment analysis

* Correspondence: nicolas.radomski@anses.fr
[1]French Agency for Food, Environmental and Occupational Health and Safety (Anses), Laboratory for Food Safety (LSAL), Paris-Est University, Maisons-Alfort, France
Full list of author information is available at the end of the article

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Salmonella is one of the main agents of foodborne bacterial infections in human. In particular, *Salmonella enterica subsp. enterica* serovars are responsible for around 80 million foodborne cases annually in developed countries [1, 2]. The 2600 known *S. enterica* subsp. *enterica* serovars exhibit a broad diversity in phenotypes including infectious patterns, lifestyle, reservoirs, vectors and host spectrum [3]. The genomic determinants of these phenotypes remain however partially characterized [4–11]. The present work tackles the genomic and metabolic signatures highlighting the poorly understood mechanisms of adaptation to animal sources at the pangenome scale of *Salmonella enterica subsp. enterica*.

From extremely clonal to the freely recombinant, bacterial evolution is mainly governed by stochastic point mutations induced by replication errors or damage of DNA (i.e. single nucleotide polymorphisms SNPs and small insertions/deletions InDels), and Horizontal Gene Transfers (HGT) promoted by homologous and non-homologous recombination events [12]. The homologous recombination events correspond to the replacement or inversion of identical or similar sequences [13], while the non-homologous recombination refers to the incorporation of new genetic material between distinct genomes [12]. The HGT whose large fragments are also named Mobile Genetic Elements (MGEs), can occur in bacterial genomes during transformation (i.e. transfer of pathogenicity islands, transposons or insertion sequences between two bacterial chromosomes), conjugation (i.e. transfer of plasmids between two bacterial genomes) and transduction (i.e. transfer and/or chromosomal incorporation of phages into bacterial genomes) [12].

The molecular mechanisms of host adaptation driven by the evolution were revealed by conventional molecular biology highlighting that *S. enterica* subsp. *enterica* extended over a wide range of hosts including birds, fishes, reptiles, amphibians, bovines, pigs and others [14]. Since the divergence from the most recent common ancestor (MRCA) with *Escherichia coli* approximately 100–160 million years ago [15], the coevolution of *Salmonella* and animal hosts during millions of years, has led to the acquisition of genes required for intestinal infection (i.e. *S. bongori* species), colonization of deeper tissues (i.e. other *S. enterica* subspp.), and expansion toward warm-blooded vertebrates (i.e. *S. enterica* subsp. *enterica*) [16]. The adaptation to warm-blooded animals started by generalist host associations related to gastrointestinal infections and transmission induced by the short-term proliferation in the intestine, or independently of the replication in the intestine by dissemination and persistence in systemic niches that are devoid of competing microbiota and can last for the lifetime of the hosts [17].

Without exhaustive data for all known serovars of *S. enterica* subsp. *enterica*, some are considered to be more adapted to mono-hosts, like Gallinarum in avian [4, 7, 10] or Dublin in bovine [4, 6]. The evolution of *S. enterica* subsp. *enterica* within hosts may have led some serovars to specialize to their host. This adaptation is accompanied by loss of bacterial fitness for inter-host transmission and apparent convergence in pathogenesis [17]. For instance, Typhi and Paratyphi A cause typhoid and paratyphoid in human, Gallinarum is associated with fowl typhoid, Abortusovis induces abortion in sheep, and Dublin and Choleraesuis are involved in bacteraemia of cattle and pigs, respectively [17]. Even if most of studies focusing on transformed seafood products [18, 19] do not provide prevalence of infected fish *in natura* [20], the serovar Bareilly is also supposed to be adapted to fish. Causing gastroenteritis, other serovars are also considered as adapted to multiple hosts like Typhimurium [9, 21] or Enteritidis [11].

Most of studies based on conventional molecular biology demonstrated that acquisition by HGT of *Salmonella* Pathogenicity Islands (SPIs) that contain genes coding for invasion, survival, and extraintestinal spread is among the prominent molecular mechanisms explaining the host adaptation of *S. enterica* subsp. *enterica* [22]. The 23 known SPIs are mainly involved in adhesion to epithelial cells (i.e. SPI-3, 4 and 5), invasion in their *Salmonella* containing vacuoles (SCV) (i.e. SPI-1 and 14), resistance to overcoming colonization of the intestinal mucus layer (i.e. SPI-6), induction of inflammation and neutrophil recruitment (i.e. SPI-1), as well as survival (SPI-11, 12 and 16) and outer membrane remodeling (SPI-2, 5 and 13) when they are in macrophages [23–25]. More precisely, two type III secretion systems (i.e. T3SS-1 and T3SS-2) encoded on SPI-1 and SPI-2 allow invasion of host epithelium and intracellular survival, respectively [17]. It must also be noted that the prophages Gifsy-2 and Fels-1 are involved in resistance to oxidative stress from neutrophils during infection, while the prophages Gifsy-1 and sopEΦ induce downregulation of inflammation in SCV and robust inflammation of the epithelial cells, respectively [25].

Albeit host adaptation of *S. enterica* subsp. *enterica* is poorly described at the genomic scale [4–11], the studies focusing on its accessory genome, confirmed that SPIs play a major role in the adaptation of few serovars to avian (e.g. SPI19 in Gallinarum and Pullorum [7, 10]) and bovine (e.g. SPI6 and SPI7 in Dublin [4, 7]) hosts. These studies emphasized that plasmids are also a major determinant explaining adaptation to avian (e.g. resistance-virulence plasmid of Kentucky [5]) and bovine (e.g. plasmid pSDV of Dublin [6]). The unique study focusing on the coregenome demonstrated that the divergence, probably induced by animal diet, between mammalian-host adapted Dublin and multi-host adapted Enteritidis was due to fixed variants targeting

regions involved in metabolic pathways of amino acids linked to glutamate [11]. This study also showed that limited ion supply in avian tract and L-arginine used for growth of laying hens, implied modifications of ion transport (i.e. potassium-efflux system in Gallinarum) and L-arginine catabolism (i.e. alanine racemase in Pullorum) of avian-adapted serovars [11].

The Genome Wide Association Study (GWAS) aims to identify the genetic variations associated with particular phenotypic traits within a population [26]. Following the first tool computing GWAS with a correction of Eukaryotic population structure based on SNPs (PLINK) [27], combinations of different methods have been implemented in the recently developed microbial GWAS. Over the last 10 years, microbial GWAS was implemented to explore a diversity of biological problems: genetic backgrounds of microbial origin [28], persistence [29], host preference [30], virulence [31, 32], and antibiotic resistance [33–42]. In comparison to human GWAS, the confounding factors of the microbial GWAS include genome selection, homologous recombination events, population structure, as well as genome wide significance [43]. Microbial GWAS takes into account these confounding factors and tests for associations between mutations and phenotypes of interest [40, 43–50]. In a context of source tracking for food safety [1, 2], microbial GWAS seems a promising tool to identify mutations associated to animal sources in order to improve models of source attribution [51].

Compared to the 10 years of developments focusing on microbial GWAS, Gene Ontology Enrichment Analysis (GOEA) has been undergoing constant improvements since the beginning of the twenty-first century and recently reached maturity for bacteria. GOEA is indeed rarely applied to bacterial genomes in spite of successful studies applying this approach to decipher host adaptation of *S. enterica* at the coregenome level [11], compare transcriptome expression profiles of minimally and highly pathogenic *S. enterica* [52], or cluster orthologous groups among differentially expressed microbial genes [53]. The GOEA proposes to test the hypergeometric distributions of GO-terms from a list of interest (i.e. tested sample) with regards to a broader set of GO-terms (i.e. universe) based on the assumption of dependencies between the GO-terms implemented through a parent-child approach [54]. GOEA was historically proposed by the Gene Ontology Consortium [55] and is today centralized in the universal protein knowledgebase commonly known as UniProt [56]. More precisely, the GO-terms link the genes and/or variants to the metabolic pathways [57] and are synthetized through a directed acyclic graph (DAG) of GO-terms into three independent ontologies called biological process (BP), molecular function (MF) and cellular component (CC) [55].

Taking into account confounding factors (i.e. genome selection, homologous recombination events, population structure and genome wide significance), the present study proposes to decipher *Salmonella* adaptation to animal sources (i.e. avian, bovine, swine and fish) based on microbial GWAS implementing accessory genes and coregenome variants (i.e. SNPs and InDels), as well as an advanced population structure correction [40]. The mutations (i.e. genes and variants) associated to traits of interest (i.e. avian, bovine, swine and fish sources) were also linked to metabolic pathways by GOEA implementing a parent-child approach [11]. To our knowledge, the present study is the first to apply successively microbial GWAS and GOEA at the pangenome scale.

Results

Distributions of serovars from potential mono-and multi-animal sources

The composition of *Salmonella* serovars from EnteroBase [58] were investigated in order to build a genome dataset taking into account the confounding factors of microbial GWAS (Additional file 1), namely genome selection [43, 44], recombination [43, 45–47], population structure [33, 40, 43, 48] and genome wide significance [43, 50]. Out of 13,635 records from a curated and synthetic subset of Enterobase, *Salmonella* isolates were mainly distributed in avian, bovine, fish, plant, shellfish and swine sources, enabling the selection of multiple strains for each studied serovar and source when building our dataset (Additional file 2). Because the detailed records from Enterobase were not enough detailed to determine if the strains from plants and shellfishes were isolated inside or outside tissues, the present study focuses on adaption to the following sources: avian, bovine, swine and fish. Among strains isolated from these sources ($n = 11{,}450$), most (22 out of 35) serovars (Fig. 1) had single animal sources ($p < 4.5 \times 10^{-1}$, Chi-square tests of uniformity to find serovars associated with some sources). Respecting high levels of diversity in terms of phylogenomic relationships in agreement with previous studies [59], geographical origins, dates of isolation and BioProject accession numbers, a balanced dataset of serovars from putative mono- and multi-animal sources (Fig. 1) were selected. This dataset was used to detect mutations and metabolic pathways associated with the adaptation of *Salmonella* serovars to their animal sources. More precisely, isolates of the *Salmonella* serovars Newport, Typhimurium and Anatum were selected as multi-animal sources, whereas other serovars were selected as mono-animal sources related to avian (i.e. Heidelberg, Kentucky, Hadar), bovine (i.e. Dublin, Cerro, Meleagridis), swine (i.e. Chloraesuis, Rissen, Derby) or fish (i.e. Brunei, Lexington, Bareilly) (Additional file 3).

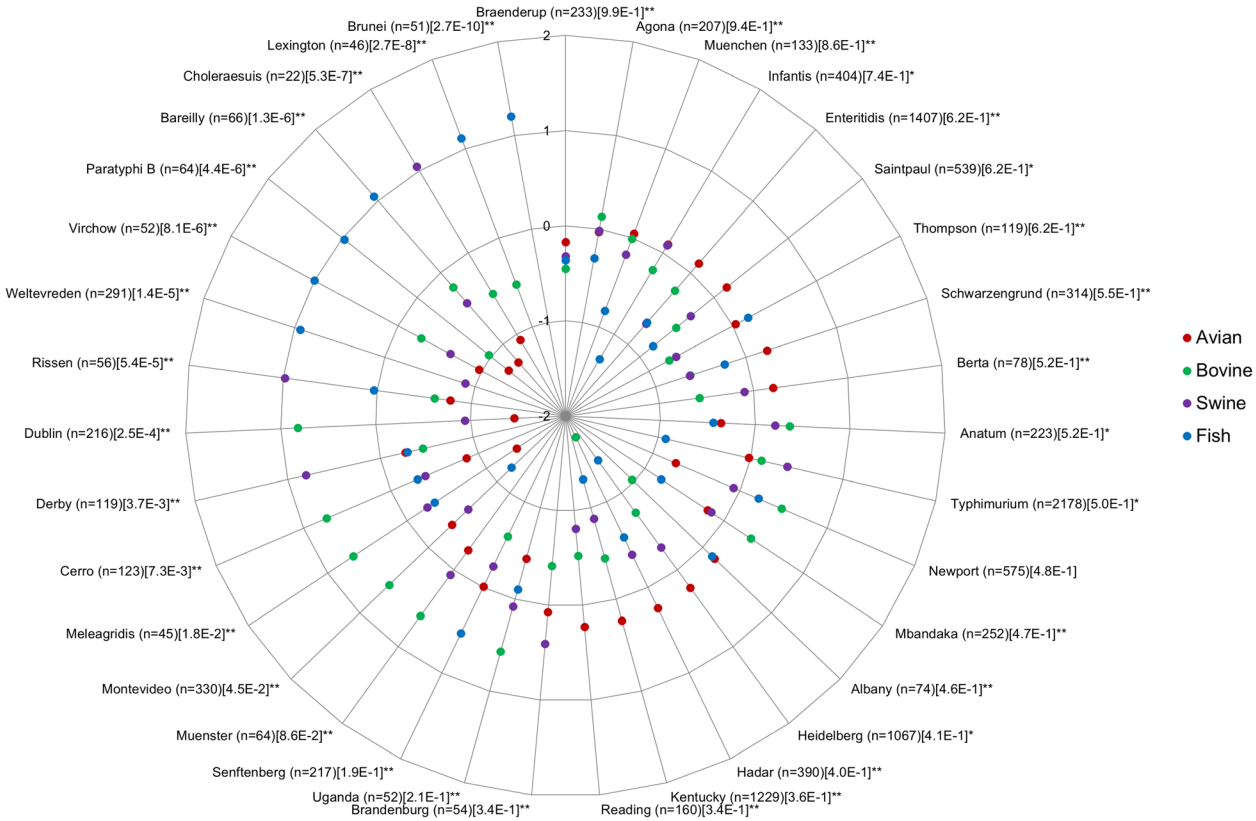

Fig. 1 Relative proportions of serovars of *Salmonella enterica* subsp. *enterica* found in each animal source (i.e. avian, bovine, fish and swine) in log-scale and corrected by the baseline proportions in the curated subset of Enterobase (see text for details). The present study focusing on adaptation to animal sources ($n = 13{,}635$) does not include isolates from environment, composite foods of the retail market and humans, which are considered as vectors of pathogen expositions and exposed susceptible consumers, respectively. The indexes higher and lower than zero represent sources in which serovars are over- and under-represented, respectively. The total effectives and *p*-values of Chi-square tests of uniformity applied to indexes are in brackets and square brackets, respectively. The serovars are sorted from the lowest (i.e. potentially mono-animal source) to highest (i.e. potentially multi-animal source) *p*-values. An asterisk stands for less than 20 samples from fish. A double asterisk stands for less than 20 samples from avian, bovine, swine and fish sources

Authenticity and completeness of detected mutations

Among the 440 selected isolates, we replaced 25 strains for which paired-end reads presenting signs of exogenous DNA and inconsistencies between in vitro (i.e. sero-agglutination register in Enterobase) [60] and in silico (i.e. SISTR program) identifications of serovars [61]. The absence of exogenous DNA was checked based on the distribution of GC% (i.e. 52.12 ± 0.09) and total sizes of studied draft genomes (i.e. Additional file 4) in comparison with the complete circular genomes selected as references during the scaffolding steps (i.e. $4.73 \pm 0.16 \times 10^{-6}$; $n = 74$).

The sizes of these 440 draft genomes (Fig. 2) agreed with the literature and ranged from 3.39 to 5.59 Mbp (i.e. between 3969 and 9898 genes) [62]. In line with studies emphasizing that host adaptation and increased pathogenicity of *Salmonella* serovars are not necessarily reflected in smaller genome sizes [5], we did not detect significant differences in terms of median values and distributions of total genomes sizes (Fig. 2) between strains from mono- and multi-animal sources (Fig. 1).

NG50 values close to the sizes of the reference circular genomes, low number of long scaffolds (i.e. between 1 and 83 higher than 1000 bp), and almost complete genome fractions (i.e. ≈ 100%) (Additional file 4), were considered as evidences of assembly quality sufficiently high to perform pangenome extraction [63]. The pangenome extraction revealed logarithmic and hyperbolic forms of curves representing the new and conserved genes according to the sizes of genome dataset, respectively (Additional file 4). According to previous studies that estimated strict coregenome sizes of *Salmonella* between 1500 [64] and 2800 [65] genes, the present open pangenome of *Salmonella enterica* consists in 2705 core genes and 19,130 accessory genes. Given the high breadth (i.e. ≈ 100%) and depth coverages (i.e. > 30X)

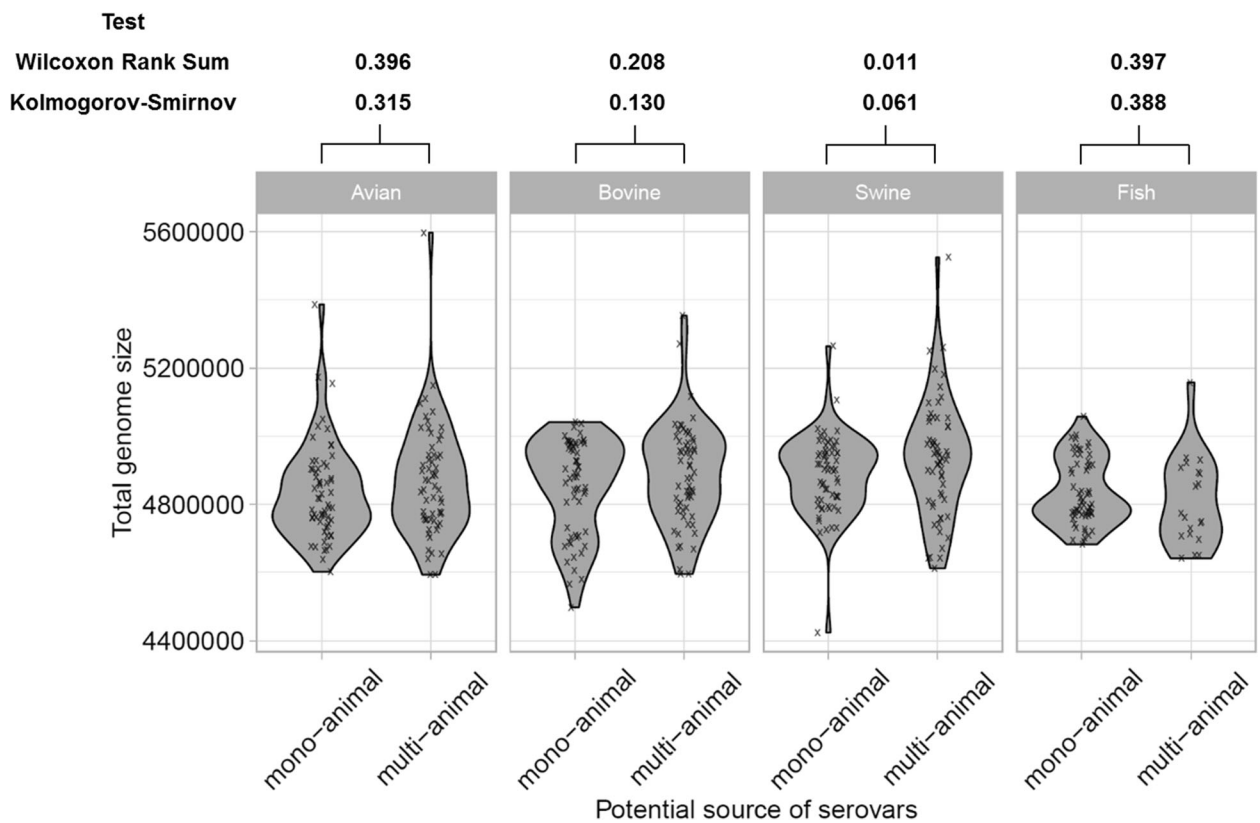

Fig. 2 Total genome sizes of *Salmonella enterica* subsp. *enterica* serovars isolated from potential mono- and multi-animal sources related to avian ($n = 120$), bovine ($n = 120$), swine ($n = 120$) and fish ($n = 80$). Based on a curated and synthetic dataset of Enterobase, the *Salmonella* serovars Newport, Typhimurium and Anatum were selected and considered as serovars from potential multi-animal sources. The other selected serovars were considered as serovars from potential mono-animal sources related to avian (i.e. Heidelberg, Kentucky, Hadar), bovine (i.e. Dublin, Cerro, Meleagridis), swine (i.e. Chloraesuis, Rissen, Derby) and fish (i.e. Brunei, Lexington, Bareilly). Normality of the data was checked using Shapiro-Wilk test ($p < 1.0 \times 10^{-2}$). The statistical differences in terms of median and distribution were assessed by non-parametric Wilcoxon rank sum and Kolmogorov-Smirnov tests, respectively

(Additional file 4), we performed variant calling analysis based on reference mapping [66]. Overall, 178, 351 variants (98% of SNPs and 2% of InDels) were detected in the coregenome, including 139,514 variants from 3030 homologous recombination events. These accessory genes and coregenome variants were considered as genuine mutations, as the analysis followed best practices for genome assembly [63] and variant calling [66].

Congruencies of phylogenomic reconstructions

Visual inspections of the few incongruencies between the phylogenomic trees obtained from 3 different approaches, namely 'variants including homologous recombination events' (called A), 'variants excluding homologous recombination events' (called B) and 'concatenated orthologous genes' (called C) (Additional file 5), are in accordance with the high congruencies of pairwise distances emphasized by the corresponding cophenetic correlation coefficients (Table 1). Even though the trees have some branches in conflicts (see Robinson-Foulds indexes in Table 1), the few incongruencies result from a Subtree Prune Regrafting move and the topologies are globally congruent (see Fowlkes-Mallows indexes in Table 1). Swapped nodes are present comparing the serovars Typhimurim and Heidelberg to Anatum (A versus C), Bareilly (B versus C), or Anatum and Bareilly (A versus B) (Additional file 5). Considering the high level of agreement between the phylogenies, (Table 1 and Additional file 5) and following the recommendations of Hedge and Wilson [67], the present study will discuss the adaptation to animal sources mainly based on the tree retaining most of genetic information (i.e. reconstructed from the approach 'A'). The phylogenomic reconstruction from the approach 'A' (i.e. iVarCall2) was indeed inferred based on coregenome SNPs from intra- and intergenic regions, as well as homologous recombination events, contrary to the approaches 'B' (i.e. 'variants excluding homologous recombination events' from iVarCall2 and ClonalFrameML) and 'C' (i.e. 'concatenated orthologous genes' from Roary).

Table 1 Congruency parameters between phylogenomic reconstructions of strains belonging to different serovars of *Salmonella enterica* subsp. *enterica* ($n = 440$) in terms of distance and topology. The phylogenomic reconstructions were performed by maximum likelihood selecting the most appropriate models of evolution and checking ultrafast bootstrap convergences (i.e. IQ-Tree). The compared approaches 'variants' and 'genes' correspond to phylogenomic trees reconstructed using pseudogenomes from variant calling analysis (i.e. iVARCall2) including (A) or excluding (B) variants from recombination events (i.e. ClonalFrameML), and concatenated orthologous genes (C) from pangenome analysis (i.e. Roary), respectively. The cophenetic function of the 'dendextend' R package was used to compute the cophenetic correlations. The dendrogram function of the 'dendextend' R package was used to compute the Fowlkes-Mallows indexes. The treedist function of the 'phangorn' R package was used to compute the Robinson-Foulds indexes

Tree parameters [a]	Congruency parameters	Compared approaches of phylogenomic reconstructions		
		'A' vs 'B'	'C' vs 'A'	'C' vs 'B'
Distance	Cophenetic correlation (Pearson)	0.989	0.993	0.981
	Cophenetic correlation (Kendall)	0.766	0.828	0.742
	Cophenetic correlation (Spearman)	0.924	0.954	0.911
Topology	Fowlkes-Mallows index	0.650	0.600	0.600
	Robinson-Foulds index	370	264	410

[a] distance refers to similarity between trees in terms of correlation between the cophenetic distance matrices. Topology refers to differences between two trees in terms of node clustering, respectively

Phylogenomic relationships between serovars from potential mono- and multi-animal sources

With the exception of serovars Newport and Cerro, all other serovars were monophyletic (Fig. 3) in all trees (Additional file 5). While the genomes of serovars from multi-animal sources were clustered into three distinct phylogenomic clusters (i.e. first lineage of Newport versus second lineage of Newport and Typhimurium versus Anatum), those from mono-animal sources were grouped by serovar (Fig. 3). The coexistence of purely clonal (i.e. mono-animal sources) and nearly panmictic (i.e. multi-animal sources) serovars (Fig. 3), emphasizes the necessity to correct the population structure when performing a microbial GWAS (Additional file 1) to find mutations associated to animal sources (i.e. avian, bovine, swine and fish).

Consideration of confounding factors during microbial GWAS

With the objective to take into account the confounding factors during microbial GWAS (Additional file 1), we compared different dataset of genomes to assess the correction of population structure and estimated the impact of the homologous recombination events [43]. More precisely, 9 microbial GWAS were performed for each animal sources (i.e. 36 analyses) considering different datasets of genomes from multi- (i.e. panmictic expansion) and/or mono- (i.e. clonal expansion) animal sources in the cluster presenting the phenotype of interest, as well as the cluster without this latter one (Additional file 6). Excluding the variants from homologous recombination events, 9 other microbial GWAS (i.e. 36 analyses) were performed with these different datasets of genomes (Additional file 7). Probably due to the coexistence of purely clonal to nearly panmictic lineages in the dataset of 440 genomes (Additional file 1), the datasets of genomes and variants from homologous recombination events affected the population structure corrections (Additional files 6 and 7). Expected shapes of quantile-quantile (QQ) plots referring to suitable population structure corrections (i.e. inflation for only highly significant observed *p*-values) were systematically checked including genomes from mono- and multi-animal sources in both studied strains and compared strains for the avian, bovine, swine and fish sources (Additional files 6 and 7). Concerning these expected shapes of QQ plots presenting inflations for only highly significant observed *p*-values, much more stratification of causal mutations were observed including variants from homologous recombination events (Additional file 6), compared to microbial GWAS excluding them (Additional file 7). All the 440 genomes included, we observed that most of the associated mutations were different comparing microbial GWAS performed with and without variants from recombination events (Table 2). According to this observation and the authors suspecting the homologous recombination events to conceal the detection of causal variants by microbial GWAS [43, 45–47], we decided to exclude the coregenome variants from these regions during microbial GWAS (i.e. 139,514 variants from 3030 homologous recombination events). Taking into account all the known confounding factors (Additional file 1), and even if common genome wide significance of human GWAS is around $p \leq 1 \times 10^{-6}$, the polygenicity was estimated at $p \leq 1 \times 10^{-2}$ according to the QQ plots of the present study focusing on microbial GWAS (Additional file 7). Without consensus concerning the genome wide significance of

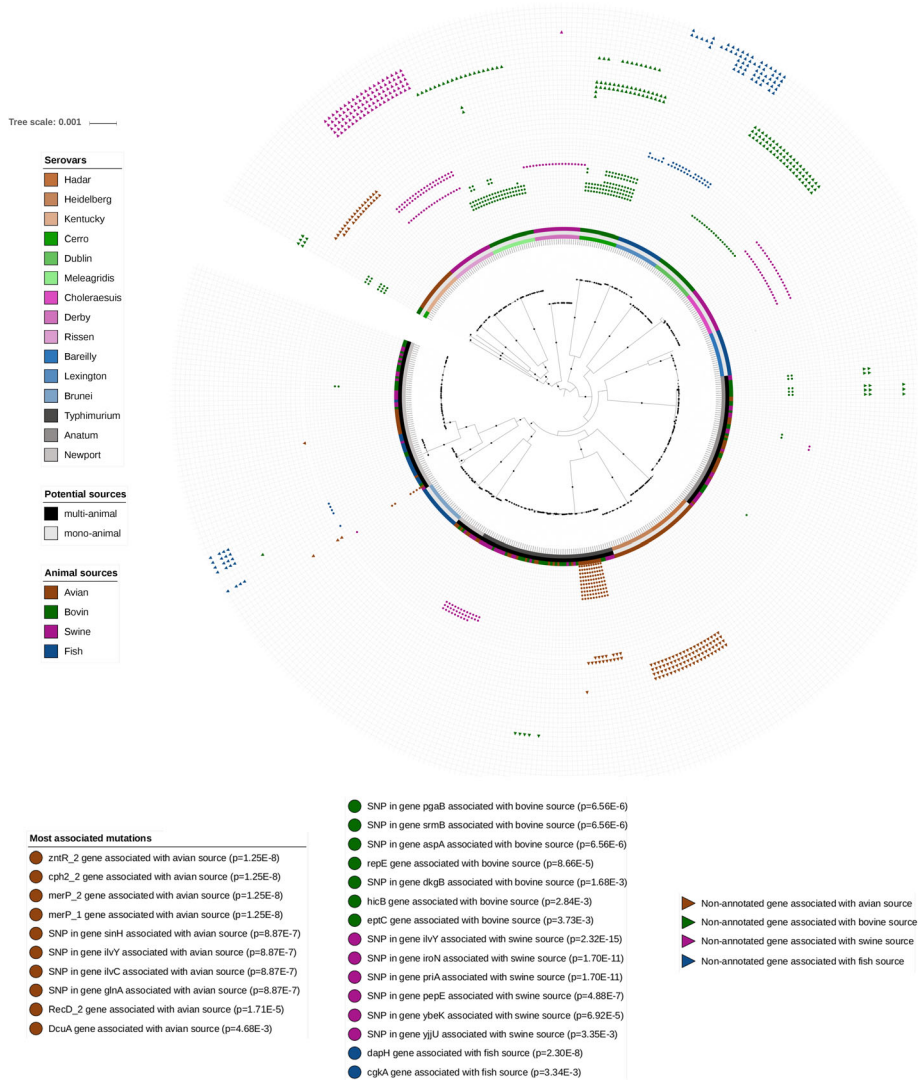

Fig. 3 Maximum likelihood phylogenomic tree of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) from potential mono- and multi-animal sources. Based on pseudogenomes inferred with the variant calling workflow iVARCall2, the workflow IQ-Tree selected the most appropriate model of evolution (GTR + I + G4) according to Akaike Information Criteria (AIC) and reconstructed the tree with an ultrafast approximation of phylogenomic bootstrap. The present phylogenomic tree was inferred including SNPs from recombination events and was rooted using the most closely related *indica* subspecies as an outgroup. The potential mono- and multi-animal sources were assigned based on Chi-square tests of uniformity applied on a curated and synthetic subset of Enterobase. Examples of mutations associated with animal sources by microbial GWAS are presented (i.e. Wald tests). These associated mutations refer to polygenicity with regard to Quantile-Quantile (QQ) plots from microbial GWAS (i.e. $p < 1 \times 10^{-2}$) and present high (i.e. > 5%) and low (i.e. < 5‰) frequencies of presence (i.e. genes and alternative variants) in the studied and compared genomes, respectively. The serovars (i.e. colored squares), potential sources (i.e. black and grew squares), animal sources (i.e. colored squares), as well as annotated (i.e. colored circles) and non-annotated (i.e. colored triangles) mutations associated to animal sources, are represented from the internal to external rings. The colored circles and triangles represent present genes or alternative variants, whereas missing data refers to absente genes or reference variants, respectively. Most of the branches of the tree (i.e. 85%) are supported by bootstrap values higher than 90% (i.e. black circles) and the corresponding newick file is accessible under request

microbial GWAS [43], and with regards to frequencies of presence and absence of genes and alternative variants (Additional file 8), we estimated and checked visually that associated mutations present *p*-values of association between $p = 8.78 \times 10^{-3}$ and $p = 2.32 \times 10^{-15}$ (Fig. 3 and Additional file 8). These mutations associated by microbial GWAS have been retained to apply downstream GOEA.

Mutation associated with animal sources (i.e. microbial GWAS)

No matter the phenotype of interest, only partial associated mutations were detected by microbial GWAS (Fig. 3). While the presence of genes and presence of alternative variants were associated with animal sources, the absence of genes and presence of reference variants were not

Table 2 Mutations of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) associated with animal sources (i.e. avian bovine, swine and fish) by microbial GWAS including or excluding variants from recombination events. The accessory genes and coregenome variants (i.e. SNPs and InDels) were annotated with Prokka (1.12) and SNPeff (4.1 g), respectively. After potential exclusion of variants from recombination events based on iVARCall2 and ClonalFrameML, the workflow 'microbial-GWAS' corrects the population structure based on Linear Mixed Model (LMM), then perform associations with Wald tests implemented in GEMMA. The associated mutations (i.e. Wald tests) refer to polygenicity with regard to Quantile-Quantile (QQ) plots from microbial GWAS (i.e. $p < 1 \times 10^{-3}$ and $p < 1 \times 10^{-2}$, with or without recombination events) and present high (i.e. > 5%) and low (i.e. < 5‰) frequencies of presence (i.e. genes and alternative variants) in the studied and compared genomes, respectively

Animal source	Comparison of associated mutations from microbial GWAS			
	Including homologous recombination		Excluding homologous recombination	
	All	Unique	All	Unique
avian	41	36	18	13
bovine	21	18	16	13
swine	35	30	11	6
fish	5	4	7	6

associated with animal sources. This observation is in accordance with the fact that losses of unessential functions do not necessarily refer to the adaptation to animal sources, as previously reported [12], or unconfirmed [5], concerning the host adaptation and restricted host transmission. As suspected with regard to higher functional impacts of accessory genes compared to coregenome variants, 38 genes were detected as associated with animal sources, whereas only 3 intergenic, 3 synonymous and 8 non-synonymous variants (SNPs and InDels) were associated to these traits of interest (Table 3). Due to the fact that synonymous variants associated to traits of interest (Table 3) may emphasize elements of regulation [68] or phenotypical impacts [69], we decided to retain them in GOEA. To summarize, 38, 34, 26 and 14 associated mutations were detected as signatures of avian, bovine, swine and fish sources, respectively (Additional file 8). Among the latter, annotations are available for only 10, 7, 6 and 2 mutations

Table 3 Mutations before and after microbial GWAS aiming to associate animal sources (i.e. avian bovine, swine and fish) with mutations from accessory (i.e. genes) and coregenome (i.e. SNPs and InDels) of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$). The accessory genes and coregenome variants (i.e. SNPs and InDels) were annotated with Prokka (1.12) and SNPeff (4.1 g), respectively. After exclusion of variants from recombination events based on iVARCall2 and ClonalFrameML, the workflow 'microbial-GWAS' corrects the population structure based on Linear Mixed Model (LMM), then perform associations with Wald tests implemented in GEMMA. The associated mutations (i.e. Wald tests) refer to polygenicity with regard to Quantile-Quantile (QQ) plots from microbial GWAS (i.e. $p < 1 \times 10^{-2}$) and present high (i.e. > 5%) and low (i.e. < 5‰) frequencies of presence (i.e. genes and alternative variants) in the studied and compared genomes, respectively

Mutations	Annotations			Before GWAS		After GWAS			
				Including homologous recombination	Excluding homologous recombination	Avian source	Bovine source	Swine source	Fish source
accessory genes and variants	annotated and hypothetical			178,351	38,837	38	34	26	14
accessory genes	annotated			6387	6387	6	3	0	2
	hypothetical			12,743	12,743	8	9	5	5
coregenome variants	intergenic			17,362	2288	1	1	1	0
	intragenic	synonymous		68,157	8365	1	1	1	0
		non synonymous	missenses	65,044	8017	2	2	4	0
			start lost	144	19	0	0	0	0
			stop gained	4202	525	0	0	0	0
			frameshift	1019	136	0	0	0	0
			disruptive inframe insertions	122	14	0	0	0	0
			disruptive inframe deletions	204	31	0	0	0	0
	multiple annotations			2967	312	0	0	0	0

associated with avian, bovine, swine and fish sources, respectively (Tables 3 and 4).

Metabolic pathways mainly impacted by mutations associated with animal sources (i.e. GOEA)

Based on the mutations associated by microbial GWAS (Table 3 and Additional file 8), the GO-terms retrieved by GOEA (Additional file 9) were parsed to retain the most accurate (i.e. GO-levels ≥5) and the most enriched (i.e. Bonferroni corrected p-values < 5.0×10^{-2}), as previously described [11]. This resulted in 6, 1, 0 and 2 GO-terms of interest for the avian, bovine, swine and fish sources, respectively (Table 5). These GO-terms (Table 5) were mainly related to

Table 4 Functionally annotated mutations (i.e. excluding genes coding hypothetical proteins) of *Salmonella enterica* subsp. *enterica* serovars (i.e. SNPs, InDels and genes) associated by microbial GWAS with animal sources (i.e. avian bovine, swine and fish). The accessory genes and coregenome variants (i.e. SNPs and InDels) were annotated with Prokka (1.12) and SNPeff (4.1 g), respectively. After exclusion of variants from recombination events based on iVARCall2 and ClonalFrameML, the workflow 'microbial-GWAS' corrects the population structure based on Linear Mixed Model (LMM), then perform associations with Wald tests implemented in GEMMA. The associated mutations (i.e. Wald tests) refer to polygenicity with regard to Quantile-Quantile (QQ) plots from microbial GWAS (i.e. $p < 1 \times 10^{-2}$) and present high (i.e. > 5%) and low (i.e. < 5‰) frequencies of presence (i.e. genes and alternative variants) in the studied and compared genomes, respectively. The genes with undefined names are assigned to STM identifiers with regard to the reference genome of *Salmonella* Typhimurium LT2 (NCBI NC_003197.1). HGVS stands for Human Genome Variation Society. N/A and ND stand for not applicable and not determined. N/A refers to intergenic regions. The term 'gene' refers to the gene presence

Studied animal source	Mutation	p-value (Wald test)	Gene name	Annotation	Variant position	HGVS notation (DNA)	HGVS notation (protein)	UniprotKB
Avian	Gene	1.2×10^{-8}	*zntR2*	HTH-type transcriptional regulator ZntR	N/A	N/A	N/A	P0ACS5
Avian	Gene	1.2×10^{-8}	*cph2_2*	Phytochrome-like protein cph2	N/A	N/A	N/A	Q55434
Avian	Gene	1.2×10^{-8}	*merP2*	Mercuric transport protein periplasmic component	N/A	N/A	N/A	P13113
Avian	Gene	1.2×10^{-8}	*merP1*	Mercuric transport protein periplasmic component	N/A	N/A	N/A	P13113
Avian	Gene	1.7×10^{-5}	*recD2*	ATP-dependent RecD-like DNA helicase	N/A	N/A	N/A	Q9RT63
Avian	Gene	4.6×10^{-3}	*dcuA*	Anaerobic C4-dicarboxylate transporter DcuA	N/A	N/A	N/A	P0ABN5
Avian	SNP	8.8×10^{-7}	*sinH*	Intimin-like inverse autotransporter protein SinH	2,650,403	c.399C > T	p.Pro133Pro	E8XGK6
Avian	SNP	8.8×10^{-7}	*ilvY*	HTH-type transcriptional activator IlvY	4,116,598	c.616G > A	p.Glu206Lys	P0A2Q2
Avian	SNP	8.8×10^{-7}	*ilvC*	Ketol-acid reductoisomerase (NADP(+))	4,117,833	c.457C > T	p.Ala153Ser	P05989
Avian	SNP	8.8×10^{-7}	*N/A*	N/A	4,217,302	N/A	N/A	N/A
Bovine	Gene	8.6×10^{-5}	*repE*	Replication initiation protein	N/A	N/A	N/A	P03856
Bovine	Gene	2.8×10^{-3}	*hicB*	Antitoxin HicB	N/A	N/A	N/A	P67697
Bovine	Gene	3.7×10^{-3}	*eptC*	Phosphoethanolamine transferase EptC	N/A	N/A	N/A	P0CB40
Bovine	SNP	1.6×10^{-3}	N/A	N/A	294,951	N/A	N/A	N/A
Bovine	SNP	6.5×10^{-6}	*arnD*	4-deoxy-4-formamido-L-arabinose phosphoundecaprenol deformylase ArnD	2,408,955	c.884A > C	p.Ala295Ala	O52326
Bovine	SNP	6.5×10^{-6}	*srmB*	ATP-dependent RNA helicase SrmB	2,783,562	c.660 T > C	p.Lys220Asn	Q8ZMX7
Bovine	SNP	6.5×10^{-6}	*aspA*	Aspartate ammonia-lyase	4,572,050	c.332C > T	p.Asn111Ile	Q7CPA1
Swine	Indel	3.3×10^{-3}	N/A	N/A	4,816,900	N/A	N/A	N/A
Swine	SNP	4.8×10^{-7}	*pepE*	Dipeptidase E	4,414,198	c.488G > T	p.Pro163Leu	P36936
Swine	SNP	1.7×10^{-11}	*iroN*	TonB-dependent siderophore receptor protein	2,924,248	c.1516G > C	p.Gly506Arg	Q8ZMN0
Swine	SNP	1.7×10^{-11}	*priA*	Primosomal protein N	4,304,871	c.689 T > C	p.Lys230Thr	Q8ZKN8
Swine	SNP	6.9×10^{-05}	*ybeK or rihA*	Pyrimidine-specific ribonucleoside hydrolase RihA	725,582	c.912A > G	p.Ala304Ala	Q8ZQY4
Swine	SNP	2.3×10^{-15}	*ilvY*	HTH-type transcriptional activator IlvY	4,116,897	c.317C > A	p.Leu106Gln	P0A2Q2
Fish	Gene	2.3×10^{-8}	*dapH*	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase	N/A	N/A	N/A	Q7A2S0
Fish	Gene	3.3×10^{-3}	*cgkA*	Kappa-carrageenase	N/A	N/A	N/A	P43478

Table 5 GO-terms mainly enriched by GOEA applied on accessory genes and coregenome variants of *Salmonella enterica* subsp. *enterica* serovars associated by microbial GWAS with animal sources (i.e. avian bovine, swine and fish). The GOEA was performed with the workflow 'fastGSEA' based on the parent-child approach integrating hypergeometric tests and Bonferroni corrections. The GOEA input sample is a list of corresponding RefSeq identifiers of accessory genes (i.e. RefSeq from Roary) and coregenome variants (i.e. NP from SNPeff 4.1 g) associated by microbial GWAS. The input universe is a list of RefSeq identifiers of all accessory genes (i.e. RefSeq from Roary) and all core genes (i.e. NP from SNPeff 4.1 g). The highest GO-levels presenting the most accurate GO-terms (i.e. ≥ 5) and the lowest Bonferroni corrected *p*-values representing highly enriched GO-terms (i.e. $< 5.0 \times 10^{-2}$), are presented. BP, MF and CC stand for biological process, molecular function and cellular component, respectively

Animal source	Uniprotkb	Associated Mutations	GO-term identifier	GO-term	Hits	Exp. hits	GO level	Corr. *p*-value	Ontology
avian	Q55434	gene *cph2_2*	GO:0009585	red, far-red light phototransduction	1	0.01	7	1×10^{-7}	BP
avian	Q55434	gene *cph2_2*	GO:0009584	detection of visible light	1	0.01	7	1×10^{-7}	BP
avian	Q55434	gene *cph2_2*	GO:0009883	red or far-red light photoreceptor activity	1	0.01	5	1×10^{-7}	MF
avian	Q9RT63	gene *recD2*	GO:0043141	ATP-dependent 5′-3′ DNA helicase activity	1	0.01	11	1×10^{-7}	MF
avian	Q9RT63	gene *recD2*	GO:0008094	DNA-dependent ATPase activity	5	0.28	10	1×10^{-3}	MF
avian	P0ABN5	gene *dcuA*	GO:0015740	C4-dicarboxylate transport	3	0.13	10	1×10^{-2}	BP
bovine	Q7CPA1	SNP in *aspA*	GO:0008797	aspartate ammonia-lyase activity	1	0.01	6	1×10^{-7}	MF
fish	Q7A2S0	gene *dapH*	GO:0047200	tetrahydrodipicolinate N-acetyltransferase activity	1	0.01	8	1×10^{-7}	MF
fish	P43478	gene *cgkA*	GO:0033918	kappa-carrageenase activity	1	0.01	6	1×10^{-7}	MF

molecular functions (i.e. 66%) and biological processes (i.e. 33%).

Discussion

Restricted and unrestricted animal sources across *Salmonella*

Salmonella serovars might be considered as having restricted (mono-) or broad (multi-) animal sources. Here we used the Enterobase resource providing both genomic data and metadata to build a dataset to explore the relationships between genotype and adaptation to the animal sources (Fig. 1). As exemplified with *Escherichia* (only host-unrestricted lineages), *Campylobacter* (both host-restricted and -unrestricted lineages) and *Staphylococcus* (only host-restricted lineages), the lineages resulting of phylogenomic reconstructions reflect the genetic structure (i.e. patterns of mutations) established through either host-adapted lineages, physical barriers to colonization, or local clonal spreading induced by selection or genetic drift [12]. The restricted and unrestricted-host lineages can be the result of a diversity of genetic processes: neutral diversification, acquisition of a host-adaptive trait causing a genome-wide purge within the population, large recombination between strains creating a hybrid lineage or negative frequency-dependent selection induced by decreasing of fitness [12]. Our segmentation distinguishing mono- and multi-animal sources should consequently reflect a representation of clonal and panmictic serovars (Additional file 1) [43] rather than a phenomenon of adaptation to single or multiple niches. This hypothesis is supported by our ability to correct population structure considering both serovars from potential mono- and multi-animal sources as genomes of interest during microbial GWAS (Additional files 6 and 7).

Genetic signatures of *Salmonella* adaptation to animal sources

Especially in highly recombinant bacterial genomes, phylogeographic signatures can be weakened due to dissemination around the world and genomic changes occurring within the reservoir hosts [70]. Even with a dataset of genomes highly diversified in terms of serovars (i.e. 12 clonal and 3 panmictic serovars including 13 monophyletic and 2 polyphyletic serovars), geographical origin (i.e. 26 countries, 68% from United States) and time of isolation (i.e. 25th and 75th percentiles: 2005–2013) origins (Additional file 3), we were able to identify genetic signatures of animal sources (Table 2, Table 4 and Additional file 8) by microbial GWAS (Fig. . 4 and Additional file 7). Host-associated genetic signatures

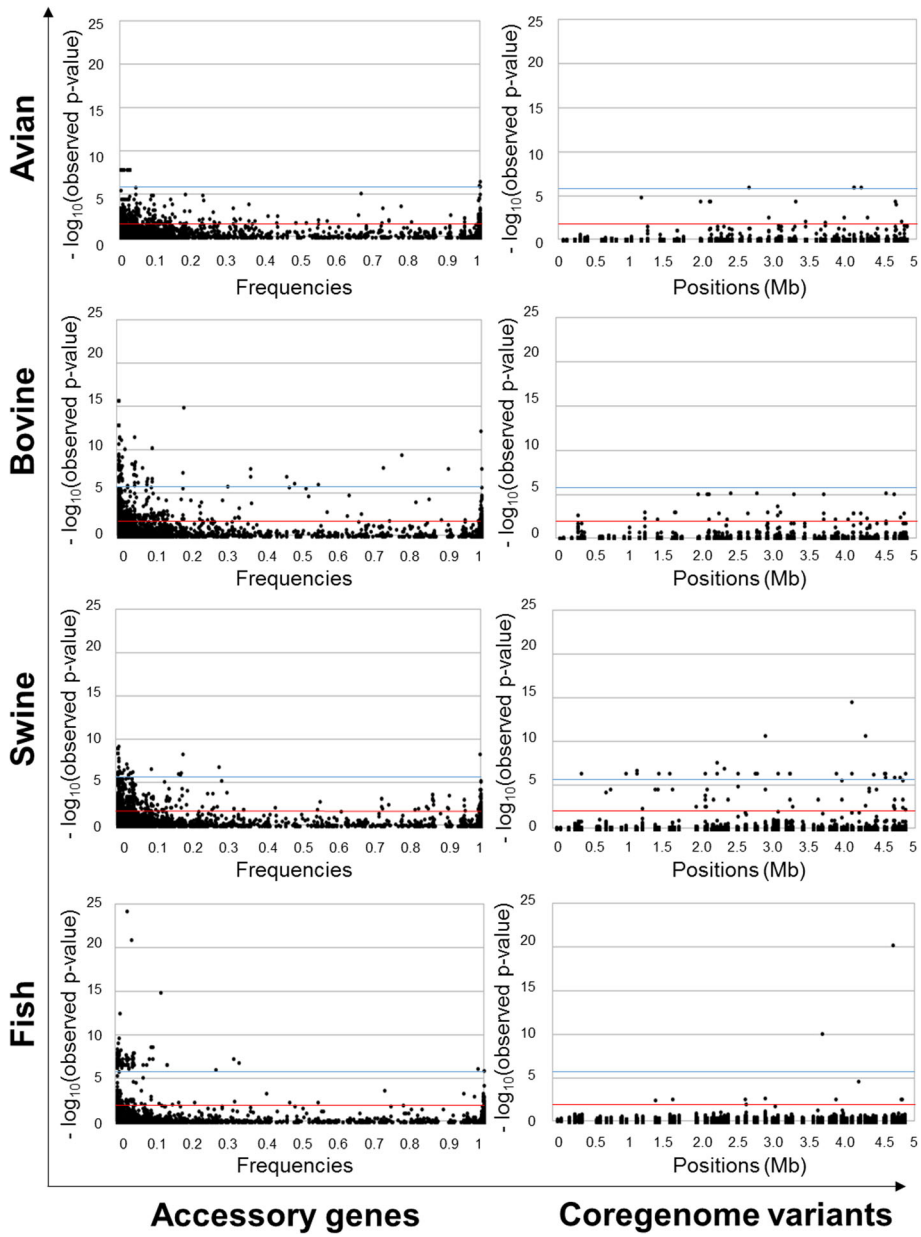

Fig. 4 Manhattan plots from microbial GWAS aiming to detect accessory genes and coregenome variants of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) associated with avian, bovine, swine and fish sources. Strains from both potential mono- and multi-animal sources were considered as having phenotype 1 and others strains (i.e. mono-source with a different animal) as having phenotype 0. The accessory genes were sorted by frequency and the positions of coregenome variants are in accordance with the reference genome of *Salmonella* Typhimurium LT2 (NCBI NC_003197.1). The blue and red lines ($-\log_{10}$(observed *p*-values)) correspond to the genome wide significance from human GWAS and polygenicity with regard to Quantile-Quantile (QQ) plots from microbial GWAS (i.e. $p < 1 \times 10^{-2}$)

have been previously detected for *Staphylococcus aureus* [71] and *Campylobacter* [72] which expanded into vast open livestock niches from humans [73] or pre-agriculture wild animal [73, 74]. Probably because *Salmonella enterica* subsps. *enterica* did not evolve as an obligate intracellular pathogens, we did not observe evidence of accumulation of deleterious mutations and losses of unessential functions (Fig. 2 and Table 2), that have been associated [12], or not [5], with host adaptation and restricted host transmission in other organisms. *Salmonella* genomes from human source were not included in the dataset of genomes, because it would conceal the mutations associated with the avian, bovine, swine and fish sources, but the mutations identified in the present study as associated with animal sources (Table 2 and Additional file 8) could be used as in silico or in vitro markers to identify them from

human isolates in a context of source tracking for food safety [1, 2]. In this framework, we plan to develop in a near future a workflow to attribute animal sources from human samples based on the markers identified in the present study to improve the models of source attribution at the genomic scale [51], as recently proposed to prediction sources of *S.* Typhimurium by machine learning Random Forest classifier [75]. Even if the annotations of hypothetical proteins associated with animal sources have to be improved in the future (Table 2 and Additional file 8), we will discuss about the annotated mutations which have been associated with animal sources (Fig. . 3).

Signatures of adaptation to the avian source

The studied *Salmonella* genomes from avian sources (Additional file 3) principally come from samples found in the food chain (i.e. 92% of poultry) and may consequently be impacted by the environment and work practices of this sector. The *S.* Typhimurium mutations associated with avian source (Fig. . 3) impacted mainly the metabolism processes related to detection of visible light (GO: 0009584 and GO:0009883), red and far-red light phototransduction (GO:0009585), DNA dependent activities (GO:0043141 and GO:0008094) and C4-dicarboxylate transport (GO:0015740) (Table 5).

Following the same pattern (i.e. a list of mutations observed in the same genomes), the genes *zntR2*, *cph2_2*, *merP_1* and *merP_2* are associated with avian source (Fig. 3, Table 4 and Additional file 8). The zing dependent helix-turn-helix domain (i.e. gene *zntR2*) allows binding to DNA cognate sites [76]. Never described in *Salmonella*, the cyanobacterial phytochrome Cph2 activate mobility capacity (i.e. phototaxis) towards red light (i.e. gene *cph2_2*) [77] and may be related to conditions of poultry growth were red light his frequently used to improve liveweight [78]. The mercuric transport protein periplasmic components (i.e. genes *merP_1* and *merP_2*) may also be an adaptation related to mercury exposure. DNA damage has been reported in rat and mouse fibroblasts as well as cells from Chinese hamster ovary and human cells [79]. Due to anthropogenic activities, the changes in the human chromatin is indeed known to be induced by mercury exposure of the biosphere during 500 years [80]. As recently observed in a large subantarctic avian community, the diet (i.e. crustacean, fish, squid and carrion-consumers), rather than taxonomy, is an important driver of avian mercury exposure [81]. Even if the mercury is considered as undesirable substance in animal feed by European Food Safety Authority [82], and in a context of prevention and control of contagious poultry diseases [83], the domestic avian community have been exposed to mercury through vaccination [84]. The ethyl mercury is very toxic water-soluble form of mercury developed in the 1920s to preserve vaccines, variously called Thimerosal, Merthiolate and Thiomersa. For instance, the Thimerosal contains 49% of mercury mass [85] and is a preservative used in vaccines preventing human flu (A/H1N1) [86], as well as infections of domestic poultry and other bird species with virulent Newcastle disease virus [84].

Following also the same pattern, the SNPs in genes *sinH*, *ilvY* and *ilvC* are associated with avian source (Fig. 3, Table 4 and Additional file 8). Among other delivery devices encoded in type V protein secretion systems (T5SS), the intimin-like inverse autotransporter protein SinH (i.e. synonymous SNP in *sinH*), also known as SivH, is a virulence factor involved in internal colonization of *Salmonella* [87]. Organized as a LysR protein-regulated system, the HTH-type transcriptional activator IlvY (i.e. missense SNP in *ilvC*) is the transcriptional regulator of the ketol-acid reductoisomerase NADP+ (i.e. missense SNP in *ilvC*) involved in the parallel pathway for the biosynthesis of L-isoleucine and L-valine [88], and was associated by different mutations to avian (p.Glu206Lys) and swine (p.Leu106Gln) sources (Table 4). As detailed latter concerning the diet of weaned piglets [89], the isoleucine and valine are also controlled in practical broiler formulas because these amino acids are limited in nutrition based on corn and soybean meals [90], and may consequently explain this missense SNP in *ilvC* (p.Glu206Lys) (Table 4).

Associated to different genomes of avian source, the genes *recD2* and *dcuA* are also mutations explaining this animal source. The ATP-dependent RecD-like DNA helicase (i.e. gene *recD2*) inhibits stress-induced mutations independently of effects on SOS induction in *Escherichia coli* [91]. Mediated by an anaerobic C4-dicarboxylate transporter DcuA (i.e. gene *dcuA*), *S.* Typhimurium performs a complete tricarboxylic acid cycle during colonization of the intestinal lumen to uptake and use poorly fermentable dicarboxylic acids, such as succinate, conferring a fitness advantage in competition with the native gut microbiota [92].

Signatures of adaptation to the bovine source

All the studied *Salmonella* genomes from bovine sources are related to livestock, and like their avian counterpart, are exposed to the related environment and work practices in the food industry (Additional file 3). The mutations associated with bovine source in *S.* Cerro, *S.* Dublin and/or *S.* Meleagridis (Fig. . 3) affected the metabolism process related to aspartate ammonia-lyase activity (GO: 0008797) (Table 5).

With an identical pattern, the SNP in *arnD*, as well as, the genes *arnD*, *srmB* and *aspA*, are associated with the bovine source (Fig. 3, Table 4 and Additional file 8). The 4-deoxy-4-formamido-L-arabinose phosphoundecaprenol deformylase ArnD (i.e. synonymous SNP in *arnD*) is involved in modification of LPS with arabinose and required for resistance to polymyxin and cationic antimicrobial

peptides [93]. This adaptation signal may be explained by the bovine exposures to polymyxin treatments. The polymyxin and colistin (i.e. polymyxins E2 and E1) are currently last-line therapeutic options to treat infections caused by multidrug-resistant Gram-negative bacteria [94], whose residues can be detected in bovine milk and tissues [95].

The ATP-dependent RNA helicase SrmB (i.e. gene *srmB*) is a dead-box family of helicase proteins involved in ribosomal biogenesis, but his function in *Salmonella* remains to be determined [96]. The aspartate ammonia-lyase encoded by *aspA*, converts aspartate to fumarate which is reduced by a fumarate reductase into succinate [97]. As observed with enterohemorrhagic *Escherichia coli*, aspartate deamination and anaerobic fumarate respiration, may be important pathways favoring *Salmonella* adaptation to bovine gut [98].

Present in different genomes from bovine source, the genes *repE*, *hicB* and *eptC* are also associated with bovine source. The replication initiation proteins (i.e. gene *repE*) relate to incompatibility of plasmids and compete each other, with potential linked accessory genes, for replication in the bacterial host [99]. As demonstrated with *E. coli*, the antitoxin HicB (i.e. gene *hicB*) forms probably a complex with the mRNA interferase HicA which becomes active after dissociation induced by nutrient starvation and produces bacteriostatic conditions for growth of other bacterial cells [100]. The phosphoethanolamine transferase EptC (i.e. gene *eptC*) adds a phosphoethanolamine to the inner core lipooligosaccharide of *C. jejuni*, promoting recognition by a human Toll-like receptor and providing resistance to relevant mammalian and avian antimicrobial peptides [101].

Signatures of adaptation to the swine source

The studied *Salmonella* genomes from swine sources are mainly representative of livestock of this sector (i.e. 88% of livestock) and potentially exposed to the environment and work habits of this food chain (Additional file 3). The mutations associated by microbial GWAS to different genomes of *S.* Choleraesuis, *S.* Derby and/or *S.* Rissen from swine source (Fig. 3 and Additional file 8), are not over-enriched by GOEA (Table 5) and are only constituted of core variants including an InDel in the intergenic region STM4562-*yjjU* and SNPs, as well as genes *pepE*, *iroN*, *priA*, *ybeK* and *ilvY* (Table 4). The dipeptidase E of *S.* Typhimurium (i.e. missense SNP in *pepE*) is hypothetically involved in sequestration of peptide aspartate used in synthesis of the aspartate family of amino acids [102], and the aspartate may be added in diets of piglets to improve growth performance and protect them against oxidative stress and mycotoxin infection [103]. Keeping in mind that iron availability increases the pathogenic potential of *S.* Typhimurium [104], the TonB-dependent siderophore receptor protein (i.e. missense SNP in *iroN*) is involved in iron acquisition in *S. enterica* [105]. The primosomal protein N (i.e. missence SNP in *priA*) allows restarting of stalled replication forks via its helicase activity [106] and the pyrimidine-specific ribonucleoside hydrolase RihA (i.e. synonymous SNP in *rihA*, also called *ybeK*) is involved in conversion of cytidine into cytosine [107]. Requiring more elements to consolidate the following hypothesis, this mutation associated to swine (i.e. synonymous SNP in *rihA*) may be linked to the pig specific pathway including the cytidine-5′-monophospho-N-acetylneuraminic acid hydroxylase (CMAH). This CMAH is implicated in production of carbohydrates on the surface of intestinal epithelial cells, which are considered as the primary elements interacting with microbes and viruses during foreign parasitic infection [108]. As previously emphasized, the HTH-type transcriptional activator IlvY (i.e. missense SNPs in gene *ilvY*: p.Glu206Lys in avian and p.Leu106Gln in swine) is involved in the parallel pathway for the biosynthesis of L-isoleucine and L-valine [88]. Just as the practical broiler formulas (Corzo et al. 2009), the isoleucine and valine are limited and added in the diet of weaned piglets (i.e. barley, wheat, maize and soya) [89], impacting expression of metabolisms involved in branched-chain amino acid, as well as amino acid composition of tissues [109].

Signatures of adaptation to the fish source

The fish sources of the studied genomes may be related to environment and work habits of this food chain because the corresponding *Salmonella* samples were isolated from fresh (i.e. 28%), frozen (42%) and processed fresh (27%) fish (Additional file 3). Without annotated mutations associated with *S.* Bareilly, the mutations associated with fish source in *S.* Lexington and *S.* Brunei impacted mainly the metabolic processes involved in kappa-carrageenase (GO:0033918) and tetrahydrodipicolinate N-acetyltransferase activities (GO:0047200), respectively (Table 5). Never studied in *Salmonella*, the kappa-carrageenase (i.e. gene *cgkA*) has been described the first time in a marine bacterium *Alteromonas carrageenovora* [110] and is involved in degradation of k-carrageenan, a linear sulfated polysaccharides extracted from red edible seaweeds [111]. The 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase (i.e. gene *dapH*) is known as the first step of the L-lysine biosynthesis via diaminopimelate pathway [112] and the fish diets based on plant ingredients are deficient in lysine which is added in fish feed to improve growth [113] and liveweight [114]. These mutations associated with fish may consequently refer to adaptation induced by natural (e.g. gene *cgkA*) and artificial (e.g. gene *dapH*) diets.

Conclusions

The strains of different serovars of the recombinant taxa *Salmonella enterica* subsp. *enterica*, evolved through clonal and panmictic lineages and adapted their

genomic contents to animal sources of food chains at the accessory and coregenome scales. The major genetic and metabolic determinants of *Salmonella* adaptation to animal sources may have been driven by the natural feeding environment of the animal (e.g. k-carrageenan from red edible seaweeds for fish) and distinct livestock diets modified by human (e.g. isoleucine and valine for poultry and pig, aspartate for piglets, and lysine for fish). Environmental stimuli (e.g. red light exposure of poultry), physiological properties of the animal itself (e.g. aspartate deamination related to bovine gut adaptation), and work habits for health protection of livestock (e.g. exposure of poultry to mercury-based vaccines and exposure of bovine to polymyxin) may have also contributed to *Salmonella* adaptation underpinned by genetic and metabolic mutations associated with animal sources through the food chain.

Methods

Approach

We propose to decipher the adaptation to animal sources of *Salmonella* serovars. Our approach aimed at selecting 440 isolates, representative of most animal sources, sequenced using paired-end reads and recorded in a curated and synthetized subset of Enterobase (i). Secondly, accessory genes and coregenome variants (i.e. SNPs and InDels) were detected (ii). Thirdly, accessory genes and coregenome variants (i.e. SNPs and InDels) were associated with the animal sources of interest (i.e. avian, bovine, swine and fish) based on an implementation of microbial GWAS correcting for strong population structure (iii). Finally, GOEA were performed in order to decipher metabolic pathways mainly impacted by the pangenomic mutations associated with the animal sources (i.e. accessory genes and coregenome variants) (iv).

Selection of a genome dataset (i)

With regard to metadata from Enterobase (i.e. December 2016: 83618 records), we selected 440 isolates in order to depict a high level of genomic diversity of *Salmonella enterica* subsp. *enterica* serovars, potentially related to mono- or multi-animal sources [58]. The corresponding reads were downloaded from the ENA [115]. With a homemade python script (version 2.7), the metadata from Enterobase was curated retaining complete records (i.e. BioProject, ENA ID, Host, sample Matrix, serovar, source niche, source origin, source details, country and collection years) and standardizing typos. Based on this curated subset of Enterobase (i.e. 37,747 records), the samples from environment, composite foods of the retail market and humans were not retained because they are considered as vectors of pathogen expositions and exposed susceptible consumers in the present study focusing on adaptation to animal sources (i.e. 13,635 records of considered sources). Taking into account the unbalanced distributions of serovars and sources in this curated database, indexes representative of the association levels of animal sources were calculated for each serovar and each source. These indexes ($i = log\frac{\%(\frac{serovar}{}source)}{\%(\frac{strains}{}source)}$) represent the common logarithm of the number of strains per source for each specific serovar ($\%(\frac{serovar}{}source)$) divided by the number of strains per source in the full curated database ($\%(\frac{strains}{}source)$). Deviations from 0 correspond to over- or under-representation of the serovar in the source. Chi-square tests of uniformity of these indexes for each serovar allowed sorting of serovars from potential multi-animal sources ($p > 0.02$ with $i \approx 0$) to potential mono-animal sources ($p < 0.02$ with $i > 0$ for over-represented serovars or $i < 0$ for under-represented serovars). Based on the curated and synthetic subset, we built a collection of 440 genomes so that its composition was genetically diversified (i.e. 15 serovars) and roughly balanced considering animal sources (i.e. mono- and multi-animal sources from 4 animal sources). More precisely, 20 genomes from each of 3 serovars from potential mono-animal sources were selected for each of the studied sources: avian, bovine, swine and fish (i.e. 240 genomes). Between 60 and 80 genomes from each of 3 serovars from potential multi-animal sources (i.e. 200 genomes) were added in order to get a roughly balanced dataset of potential mono- and multi-animal sources. The balance between mono- and multi-animal sources was deliberate and used to evaluate the impact on several confounding factors during microbial GWAS. The manual selection of isolates was performed checking the animal sources and respecting high levels of diversity concerning the geographical origins, isolation dates and BioProject accession numbers.

Coregenome variants (ii)

The coregenome SNPs and small InDels were detected based on the variant caller HaplotypeCaller implemented in the iVARCall2 workflow [11], using *Salmonella* Typhimurium LT2 (NCBI NC_003197.1) as a reference genome, and following the best practices proposed by the Genome Analysis ToolKit [116]. More precisely, secondary alignments around small InDels were performed and duplications were excluded before variant calling analysis via local de novo assembly of haplotypes in active regions. The variants (i.e. SNPs and InDels) were flagged with SnpSift (version 4.1 g) [117] and the functional annotations of these variants were obtained with SNPeff (version 4.1 g without variants from intron, UTR-5′, UTR-3′, upstream regions, and downstream regions) [118]. As previously described, variants from homologous recombination events were detected with ClonalFrameML [13] and excluded to build set of 38,837 variants, or not to build set of 178,351 variants, with the script 'Clonal_VCFilter' [11]. The pseudogenomes were produced with the script 'VCFtoPseudoGenome' and correspond to the reference genome where the genotypes of detected variants were replaced in each genome [11].

Accessory genome (ii)

With an in-house workflow called ARTwork, the assembly was performed based on coverage control (i.e. > 100X) with Bbmap [119], read normalization (i.e. 100X) with Bbnorm [120], control of read quality with FastQC [121], read trimming (i.e. > 20 of Quality Control) with Trimmomatic [122], de novo assembly with SPAdes [123], selection of the closely related reference genomes with MinHash among 74 reference circular genomes [124], scaffolding with MeDuSa [125], gap filling with GMcloser [126], trimming of small scaffolds (i.e. < 200 bases) with Biopython [127], as well as control of assembly quality with QUAST [128], MultiQC [129] and ggplot2 [130] graphics. Based on these draft genomes, pangenome was constructed with Roary [131] setting 95% of identity for blastp and a strict definition of the coregenome (i.e. 100% of isolates with core genes).

Population structure (iii)

The phylogenomic reconstructions were performed based on the coregenome variants including or excluding variants from homologous recombination events (i.e. pseudogenomes from iVARCall2 [11]), as well as core genes (i.e. concatenated orthologous genes from Roary [131]). IQ-Tree [132] was applied on our datasets made up of millions of aligned sites to perform fast selections of the models of evolution based on Akaike Information Criteria (AIC) [133] and efficient tree reconstructions by maximum likelihood based on the most appropriated model of evolution [132]. More precisely, the consensus trees were produced considering all possible Nearest-Neighbor-Interchanges (NNIs) instead of only surrounded computed NNIs [132]. The search in the tree space started from a BIONJ tree [132] and an improved version [134] of the ultrafast bootstrap [135] was applied with 1000 iterations to compute boostrap support values. UFBoot convergences were checked after the IQ-Tree computation [132]. As stated in the literature, *Salmonella enterica* subsp. *indica* is the subspecies closest to subsp. *enterica* and was consequently used as an outgroup to root the tree of the subspecies *enterica* [59, 64]. Practically, this subspecies *enterica* root was identify using three isolates of the subsp. *indica* (SRR1840570, SRR1060719 and SRR1060512) and three isolates of each studied serovar. The tree distances were compared numerically with the cophenetic function of the 'dendextend' R package based on the Pearson, Kendall and Spearman correlations (i.e. between −1 and + 1, referring to anti-correlated and correlated distances) [136]. The tree topologies were compared visually with the cophylo function of the 'phytools' R package [137].

Table 6 Summary of microbial GWAS. Microbial GWAS developed until now are listed comparing their workflows, mutations of interest, studied phenotypes and genome dataset

Workflow	Explicative mutations	Population structure correction	Trait	Species	Sample	Reference [a]
bespoke	Phenotype + kmer	YES	Preferential host	*Campylobacter jejuni*	192	[30]
PhyC	Phenotype + SNP	YES	Antibiotic resistance	*Mycobacterium tuberculosis*	123	[33]
N/A	Phenotype + SNP	NO	Virulence	*Staphylococcus aureus*	90	[32]
Scoary	Gene	YES	Antibiotic resistance	*Streptococcus pneumoniae*	3085	[38]
Gemma	SNP	YES	Antimicrobial resistance	*M. tuberculosis, S. aureus, E coli, K. pneumoniae*	3144	[40]
Treewas	SNP + Gene + kmer	YES	Antimicrobial resistance	*Neisseria meningitidis*	ND	[41]
PLINK	SNP	NO	Drug resistance	*Mycobacterium tuberculosis*	123	[36]
PhyC	SNP	ND	Drug resistance	*Mycobacterium tuberculosis*	498	[39]
RoadTrips	SNP	NO	Drug Resistance	*Staphylococcus aureus*	75	[34]
PLINK	SNP	NO	Drug resistance	*Streptococcus pneumoniae*	3701	[35]
Scoary	Gene	YES	Geographical origin	*Salmonella enterica*	1327	[28]
DBGWAS	kmer	YES	Antibiotic resistance	*M. tuberculosis, S. aureus, P. aeruginosa*	1302, 992, 282	[42]
Scoary + GEMMA	Gene + SNPs	YES	Cold persistence	*Listeria monocytogenes*	51	[29]
PLINK	SNP	NO	Drug resistance	HIV	343	[140]
PLINK	SNP	NO	Viral load	HIV	1071	[31]
FaST-LMM	SNP	YES	Drug resistance	*Plasmodium falciparum*	1063	[37]

[a] references completed from Power et al. [43]. ND stands for not determined

The tree distances were also numerically compared computing the Fowlkes-Mallows index (i.e. between 0 and 1, referring to dissimilar and similar topologies, respectively) with the dendrogram function of the 'dendextend' R package [138], and the Robinson-Foulds index (i.e. number of different nodes between both tree) with the treedist function of the 'phangorm' R package [139].

Genome wide association study (iii)

Within a range from 51 *Listeria monocytogenes* [29] to 3701 *Streptococcus pneumoniae* strains [35] and without consensus on the appropriated size of genome dataset, most of the microbial GWAS includes around 500 samples under clonal and/or panmictic status (Table 6) [43]. Contrary to human GWAS focusing on the effects of individual SNPs, microbial GWAS has also to access phenotype associations based on presence/absence of genes alongside SNPs [43]. In addition, microbial GWAS has to take into account confounding factors such as genome selection, homologous recombination events, population structure related to Linkage Disequilibrium (LD), and genome wide significance, because they can induce false positive identifications of seemingly causal mutations [43, 141]. With regard to the confounding factors (Additional file 1), we applied the developed microbial GWAS (Fig. 5) based on GEMMA [40]. This workflow was applied to 440 genomes, comparing different sizes of genome dataset, taking into account variants from homologous recombination events and checking population structure corrections. The associated mutations (i.e. Wald tests) refer to polygenicity with regard to QQ plots from microbial GWAS (i.e. $p < 1 \times 10^{-2}$), and present high (i.e. > 5%) and low (i.e. < 5‰) frequencies of presence or absence (i.e. genes and alternative variants) in the studied and compared genomes, respectively.

Gene ontology enrichment analysis (iv)

Based on our recently published workflows called 'GetGOxML' and 'EveryGO' aiming at retrieving GO-terms online from coregenome variants and perform GOEA at any

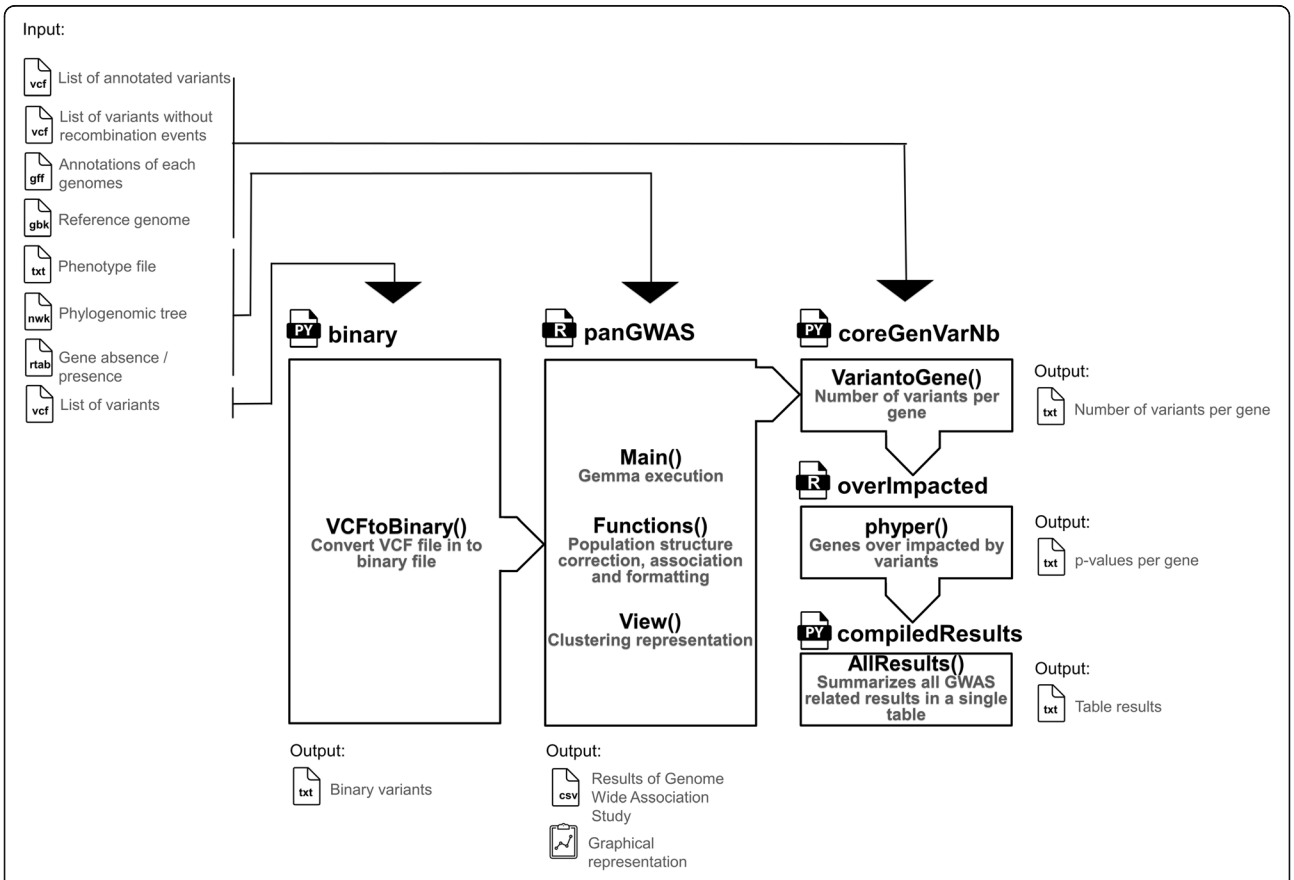

Fig. 5 Developed scripts and published programs (i.e. black letters) with their corresponding effects (i.e. green letters) implemented in the driving script 'microbial-GWAS' performing microbial GWAS integrating Linear Mixed Model (LMM) for population structure correction. Based on the LMM integrated in GEMMA, the sequential workflow called 'microbial-GWAS' is written in R and Python 2.7. It runs successively scripts called 'binary', 'panGWAS', 'coreGenVarNb', 'overImpacted' and 'AllResults' in order to standardize SNPs, InDels and genes as binary data, compute Kinship matrix, fit a LMM and perform Wald tests, as well as detect coregenome variants presenting high gene densities (i.e. hotspots of variants) and high functional impacts (i.e. non-synonymous variants)

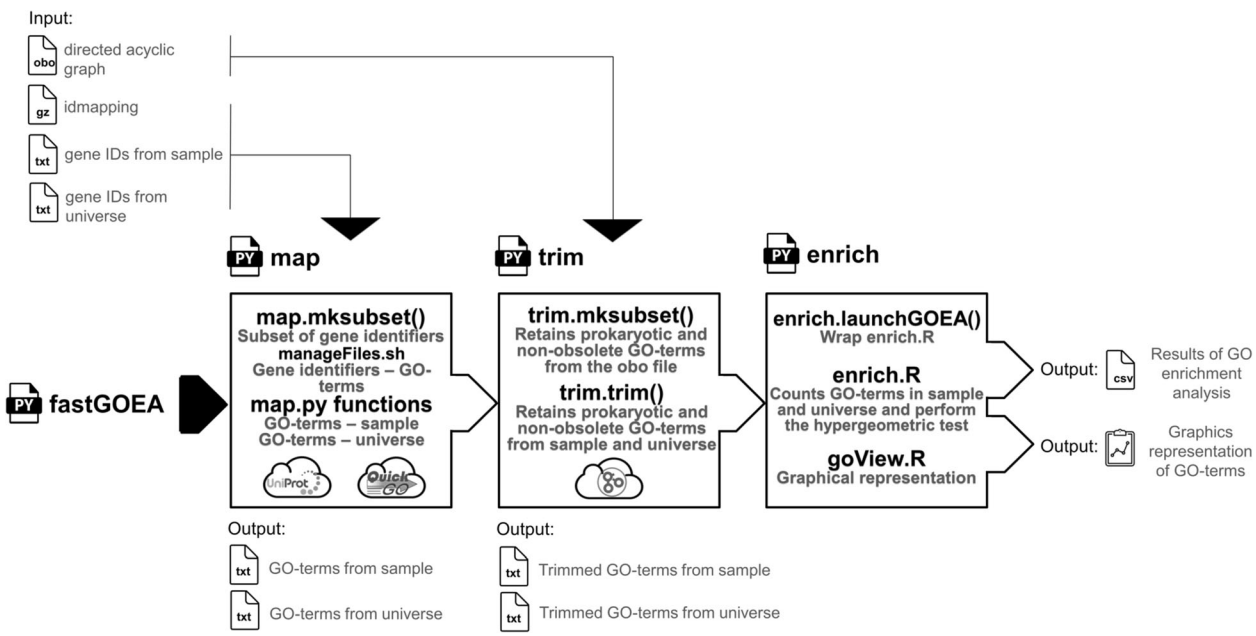

Fig. 6 Developed scripts and published programs (i.e. black letters) with their corresponding effects (i.e. grew letters) implemented in the driving script 'fastGOEA' performing GOEA based on the parent-child approach integrating hypergeometric tests and Bonferroni corrections. This workflow examines the Uniprot API (i.e. slow mode) or by default a local dataset of gene identifiers (fast mode) from Uniprot (i.e. idmapping.selected.table.gz) in order to associate GO-terms to gene identifiers provided by the user as two lists referring to the sample and universe of hypergeometric tests. With regard to a directed acyclic graph (DAG) of GO-terms (i.e. go-basic.obo including eukaryotic and prokaryotic GO-terms or gosubset_prok.obo including prokaryotic GO-terms), this workflow retains prokaryotic GO-terms and avoids obsolete GO-terms before to perform GOEA. These scripts were written with Python 2.7 and implement R libraries 'p.ajust', 'phyper' and 'ggplot2'. The whole workflow is automated and the scripts 'trim' and 'enrich' has to be performed for each GO-terms during trimming and enrichment steps, respectively

node of a phylogenomic reconstruction [11], we developed an improved workflow called 'fastGSEA' (Fig. 6). This workflow 'fastGSEA' produces a fast GOEA dependently of a local Uniprot dataset of GO-terms to decrease the execution duration, and provide a complete automatic workflow applicable to many kinds of gene identifiers (i.e. 15 different gene identifiers) [56]. This workflow can also produce a slower GOEA dependently of the current version of an application programming interface provided by QuickGO (i.e. 'Annotations' https://www.ebi.ac.uk/QuickGO/api/index.html). The driver script 'fastGSEA' is written in Python (version 2.7) and uses as input a dataset of gene identifiers (i.e. idmapping.selected.table.gz; current release from Uniprot: ftp://ftp.uniprot.org/). It requires also two lists of gene identifiers from the sample of interest and universe, as well as a file representing the DAG of GO-terms (i.e. go-basic.obo including eukaryotic and prokaryotic GO-terms or gosubset_prok.obo including only prokaryotic GO-terms). The first step of the workflow 'fastGSEA' aims at selecting from the dataset 'idmapping', a subset of gene identifiers linking the gene identifiers provided by the user, corresponding Uniprot identifiers and associated GO-terms from the sample and universe lists locally (i.e. based on GO-terms from the subset) or online (i.e. based on Uniprot identifiers from the subset). Secondly, the workflow uses the DAG of GO-terms to retain prokaryotic GO-terms and avoid obsolete GO-terms. The third step of the workflow tests the hypergeometric distributions of GO-terms (i.e. 'phyper' R function) [142] and corrects the produced *p*-values based on the Bonferroni correction ('p.ajust' R function) [143]. In parallel to a file centralizing the results (i.e. GO-terms, number of hits, GO levels, *p*-values, ontology), the workflow 'fastGSEA' produces finally a graphical representation of the GOEA with the plotting system ggplot2 [130].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6188-x.

Additional file 1. Confounding factors of microbial GWAS. The confounding factors of microbial GWAS include the selection of genomes, homologous recombination events, population structure, as well as genome wide significance.

Additional file 2. Distribution of source niches and source types of *Salmonella enterica* subsp. *enterica* serovars from Enterobase among full records in terms of read identifier, serovar, source niche, source niche, source detail, BioProject accession number of the European Nucleotide

Archive (ENA), date of isolation and country of isolation. Due to typos and missing information, the metadata of Enterobase was downloaded on December 2016 (83,618 records), then curated and synthesized (i.e. 37,747 records) with a homemade python script (version 2.7). The present study focusing on adaptation to animal sources ($n = 13,635$ records) does not include isolates from composite foods of the retail market and humans, which are considered as vectors of pathogen expositions and exposed susceptible consumers, respectively.

Additional file 3. Samples of *Salmonella enterica* subsp. *enterica* serovars studied in the present study ($n = 440$). The read identifier, serovar, source niche, source detail, BioProject accession number of the European Nucleotide Archive (ENA), date of isolation and country of isolation were retrieved from Enterobase on December 2016. ND stands for not determined and corresponds to three samples added to reach a balanced dataset of genomes.

Additional file 4. Boxplots (median, 25th percentile, 75th percentile, minimum and maximum) of depth (A) and breadth (B) coverages, numbers of scaffolds higher than 1000 bp (C), NG50 (D), genome fractions (E), and number of genes resulting of pangenome extraction (F) of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$). Assembly, variant calling analysis, computing of metrics, and pangenome analysis were performed with ARTWork, iVARCall2, Quast-MultiQC and Roary, respectively. *Salmonella* Typhimurium LT2 (NCBI NC_003197.1) was used as the reference genome for mapping during variant calling analysis. Black dots represent the sizes of the closely related reference genomes selected among 74 reference-circularized genomes based on MinHash distances.

Additional file 5. Topology differences of phylogenomic trees of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$). The phylogenomic trees were reconstructed by maximum likelihood selecting the most appropriated models of evolution and checking ultrafast bootstrap convergences (i.e. IQ-Tree). The compared approaches 'variants' and 'genes' correspond to phylogenomic reconstructions based on pseudogenomes from variant calling analysis (i.e. iVARCall2) including (A) or excluding (B) recombination events (i.e. ClonalFrameML), and concatenated orthologous genes (C) from pangenome analysis (i.e. Roary), respectively. These graphical representations were produced with the cophylo function of the 'phytools' R package. Most of the branches of the trees (i.e. 85, 55 and 77% for approaches A, B and C, respectively) are supported by bootstrap values higher than 90% and the corresponding newick files are accessible under request.

Additional file 6. Quantile-Quantile (QQ) plots from microbial GWAS aiming to identify polygenicity during associations of accessory genes and coregenome variants including homologous recombination events of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) with avian (A), bovine (B), swine (C) and fish (D) sources. The samples were assigned to potential mono- and multi-animal sources based on a curated and synthetic version of Enterobase. The absence of GEMMA convergence is represented by a cross. The red line (i.e. - $\log_{10}$(observed *p*-values) = - $\log_{10}$(expected *p*-values)) corresponds to the reference line reflecting the level of population structure correction.

Additional file 7. Quantile-Quantile (QQ) plots from microbial GWAS aiming to identify polygenicity during associations of accessory genes and coregenome variants excluding homologous recombination events of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) with avian (A), bovine (B), swine (C) and fish (D) sources. The samples were assigned to potential mono- and multi-animal sources based on a curated and synthetic version of Enterobase. The absence of GEMMA convergence is represented by a cross. The red line (i.e. - $\log_{10}$(observed *p*-values) = - $\log_{10}$(expected *p*-values)) corresponds to the reference line reflecting the level of population structure correction.

Additional file 8. Microbial GWAS results aiming to associate accessory genes and coregenome variants of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) with animal sources (i.e. avian bovine, swine and fish). The microbial GWAS was performed with the workflow 'microbial-GWAS' based on Linear Mixed Model (LMM) for population structure correction and Wald tests for association. The genome dataset includes both genomes assigned to potential mono- and multi-animal sources based on a curated and synthesized version of Enterobase. The associated mutations (i.e. Wald tests) refer to polygenicity with regard to Quantile-Quantile (QQ) plots from microbial GWAS (i.e. $p < 1 \times 10^{-2}$) and present high (i.e. > 5%) and low (i.e. < 5‰) frequencies of presence or absence (i.e. genes and alternative variants) in the studied and compared genomes, respectively.

Additional file 9. GO-terms enriched by GOEA applied on accessory genes and coregenome variants of *Salmonella enterica* subsp. *enterica* serovars ($n = 440$) associated with animal sources (i.e. avian bovine, swine and fish). The GOEA was performed with the workflow 'fastGSEA' based on the parent-child approach integrating hypergeometric tests and Bonferroni corrections. The GOEA input sample is a list of corresponding RefSeq identifiers of accessory genes (i.e. RefSeq from Roary) and coregenome variants (i.e. NP from SNPeff 4.1 g) associated by microbial GWAS. The input universe is a list of RefSeq identifiers of all accessory genes (i.e. RefSeq from Roary) and all core genes (i.e. NP from SNPeff 4.1 g). BP, MF and CC stand for Biological Process, Molecular Function and Cellular Component, respectively.

Abbreviations

AIC: Akaike Information Criteria; BP: Biological Process; CC: Cellular Component; CMAH: Cytidine-5′-monophospho-N-acetylneuraminic acid hydroxylase; EFSA: European Food Safety Authority; ENA: European Nucleotide Archive; GOEA: Gene Ontology Enrichment Analysis; GWAS: Genome Wide Association Study; HGT: Horizontal Gene Transfers; InDels: Small insertions/deletions; LD: Linkage Disequilibrium; MF: Molecular Function; MGEs: Mobile Genetic Elements; NNIs: Nearest-Neighbor-Interchanges; QQ plots: Quantile-Quantile plots; SPIs: *Salmonella* Pathogenicity Islands; T3SS: Type III secretion systems; T5SS: Type V protein secretion systems

Acknowledgements

We thank especially Pierre-Yves Letournel and Thomas Texier for providing high-performance computing resources.

Authors' contributions

NR, MM, MYM, and PB conceived the study. NR, MM and MYM draft the manuscript. NR, MM, MVN, KD, KL and AF performed the analyses. All authors commented and approved the final manuscript, take public responsibility for appropriate portions of the content and agree to be accountable for all aspects of the work in terms of accuracy or integrity.

Funding

The French Agency for Food, Environmental and Occupational Health and Safety (Anses) and the French National Institute for Agricultural Research (INRA) (grant name Typautobac) covered the salary of MVN. The project 'COllaborative Management Platform for detection and Analyses of (Re-) emerging and foodborne outbreaks in Europe' (COMPARE) received funding from the European Union's Horizon 2020 research and innovation programme (grant agreement No. 643476) and supported the participations at congresses of MVN. None of the funding bodies participated to the design of the study, collection, analysis, and interpretation of data, or writing the manuscript.

Availability of data and materials

BioProjets and sequencing reads analysed during this study are included in Additional file 3. All data generated or analysed during this study are included in this published article and its Additional files. The scripts of the developed workflow called 'microbial-GWAS' can be found in the following GitHub repository: https://github.com/VilaNovaMeryl. The scripts of the developed workflow called 'fastGSEA' can be found in the following GitHub repository: https://github.com/KDurimel/DNAlogy/tree/master/FAST_GOEA.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details
[1]French Agency for Food, Environmental and Occupational Health and Safety (Anses), Laboratory for Food Safety (LSAL), Paris-Est University, Maisons-Alfort, France. [2]Applied Mathematics and Computer Science, from Genomes to the Environment (MaIAGE), French National Institute for Agricultural Research (INRA), Université Paris-Saclay, 78350 Jouy-en-Josas, France.

Received: 24 June 2019 Accepted: 15 October 2019
Published online: 06 November 2019

References

1. Majowicz SE, Musto J, Scallan E, Angulo FJ, Kirk M, O'Brien SJ, et al. The global burden of Nontyphoidal *Salmonella* gastroenteritis. Clin Infect Dis. 2010;50:882–9.
2. EFSA-ECDC. The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. Eur Food Saf Auth J. 2016;14:1-231.
3. Guibourdenche M, Roggentin P, Mikoleit M, Fields PI, Bockemühl J, Grimont PAD, et al. Supplement 2003–2007 (no. 47) to the White-Kauffmann-Le minor scheme. Res Microbiol. 2010;161:26–9.
4. Porwollik S, Santiviago CA, Cheng P, Florea L, McClelland M. Differences in gene content between Salmonella enterica Serovar Enteritidis isolates and comparison to closely related Serovars Gallinarum and Dublin. J Bacteriol. 2005;187:6545–55.
5. Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, LeClerc JE, et al. Comparative genomics of 28 Salmonella enterica isolates: evidence for CRISPR-mediated adaptive sublineage evolution. J Bacteriol. 2011;193:3556–68.
6. Chu C, Feng Y, Chien A-C, Hu S, Chu C-H, Chiu C-H. Evolution of genes on the Salmonella Virulence plasmid phylogeny revealed from sequencing of the virulence plasmids of S. enterica serotype Dublin and comparative analysis. Genomics. 2008;92:339–43.
7. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, et al. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. Proc Natl Acad Sci. 2015;112:863–8.
8. Zhou Z, McCann A, Litrup E, Murphy R, Cormican M, Fanning S, et al. Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* Serovar Agona. PLoS Genet. 2013;9:e1003471.
9. Pang S, Octavia S, Feng L, Liu B, Reeves PR, Lan R, et al. Genomic diversity and adaptation of Salmonella enterica serovar Typhimurium from analysis of six genomes of different phage types. BMC Genomics. 2013;14:718.
10. Thomson NR, Clayton DJ, Windhorst D, Vernikos G, Davidson S, Churcher C, et al. Comparative genome analysis of Salmonella Enteritidis PT4 and Salmonella Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways. Genome Res. 2008;18:1624–37.
11. Felten A, Vila Nova M, Durimel K, Guillier L, Mistou M-Y, Radomski N. First gene-ontology enrichment analysis based on bacterial coregenome variants: insights into adaptations of Salmonella serovars to mammalian- and avian-hosts. BMC Microbiol. 2017;17:1–22.
12. Sheppard SK, Guttman DS, Fitzgerald JR. Population genomics of bacterial host adaptation. Nat Rev Genet. 2018;19:549–65.
13. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11:e1004041.
14. Evangelopoulou G, Kritas S, Govaris A, Burriel AR. Animal salmonelloses: a brief review of "host adaptation and host specificity" of Salmonella spp. Vet World. 2013;6:703–8.
15. Meysman P, Sánchez-Rodríguez A, Fu Q, Marchal K, Engelen K. Expression divergence between Escherichia coli and Salmonella enterica serovar Typhimurium reflects their lifestyles. Mol Biol Evol. 2013;30:1302–14.
16. Bäumler AJ, Tsolis RM, Ficht TA, Adams LG. Evolution of host adaptation in Salmonella enterica. Infect Immun. 1998;66:4579–87.
17. Tanner JR, Kingsley RA. Evolution of Salmonella within hosts. Trends Microbiol. 2018;26:986–98.
18. Hoffmann M, Luo Y, Monday SR, Gonzalez-Escalona N, Ottesen AR, Muruvanda T, et al. Tracing origins of the *Salmonella* Bareilly strain causing a food-borne outbreak in the United States. J Infect Dis. 2016;213:502–8.
19. Heinitz ML, Ruble RD, Wagner DE, Tatini SR. Incidence of Salmonella in fish and seafood. J Food Prot. 2000;63:579–92.
20. Bibi F, Qaisrani S, Ahmad A, Akhtar M, Khan B, Ali Z. Occurrence of salmonella in freshwater fishes. Pakistan Agricultural Scientists Forum. 2015; 25:303–10.
21. Rabsch W. Salmonella enterica serotype Typhimurium and its host-adapted variants. Infect Immun. 2002;70:2249–55.
22. Siriken B. Salmonella Pathogenicity Islands. Mikrobiyol Bul. 2013;47:181–8.
23. Fookes M, Schroeder GN, Langridge GC, Blondel CJ, Mammina C, Connor TR, et al. Salmonella bongori provides insights into the evolution of the salmonellae. PLoS Pathog. 2011;7:e1002191.
24. Kérouanton A, Hirchaud E, Rose V, Esnault E, Naquin D, Denis M. First Complete Genome Sequence of a *Salmonella enterica subsp. enterica* Serovar Derby Strain Associated with Pork in France. Genome Announc. 2015;3:e00853–15 /ga/3/4/e00853–15.atom.
25. Ilyas B, Tsai CN, Coombes BK. Evolution of Salmonella-host cell interactions through a dynamic bacterial genome. Front Cell Infect Microbiol. 2017;7:1-15.
26. Lees JA, Bentley SD. Bacterial GWAS: not just gilding the lily. Nat Rev Microbiol. 2016;14:406.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
28. Palma F, Manfreda G, Silva M, Parisi A, Barker DOR, Taboada EN, et al. Genome-wide identification of geographical segregated genetic markers in Salmonella enterica serovar Typhimurium variant 4,[5],12:i:-. Sci Rep. 2018;8:15251.
29. Fritsch L, Felten A, Palma F, Mariet J-F, Radomski N, Mistou M-Y, et al. Insights from genome-wide approaches to identify variants associated to phenotypes at pan-genome scale: application to L. monocytogenes' ability to grow in cold conditions. Int J Food Microbiol. 2019;291:181–8.
30. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in campylobacter. Proc Natl Acad Sci. 2013;110:11923–7.
31. Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, et al. A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. eLife. 2013;2:e01123.
32. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, et al. Predicting the virulence of MRSA from its genome sequence. Genome Res. 2014;24:839–49.
33. Farhat MR, Shapiro BJ, Kieser KJ, Sultana R, Jacobson KR, Victor TC, et al. Genomic analysis identifies targets of convergent positive selection in drug-resistant mycobacterium tuberculosis. Nat Genet. 2013;45:1183–9.
34. Alam MT, Petit RA, Crispell EK, Thornton TA, Conneely KN, Jiang Y, et al. Dissecting Vancomycin-intermediate resistance in Staphylococcus aureus using genome-wide association. Genome Biol Evol. 2014;6:1174–85.
35. Chewapreecha C, Marttinen P, Croucher NJ, Salter SJ, Harris SR, Mather AE, et al. Comprehensive identification of single nucleotide polymorphisms associated with Beta-lactam resistance within pneumococcal mosaic genes. PLoS Genet. 2014;10:e1004547.
36. Chen PE, Shapiro BJ. The advent of genome-wide association studies for bacteria. Curr Opin Microbiol. 2015;25:17–24.
37. Miotto O, Amato R, Ashley EA, MacInnis B, Almagro-Garcia J, Amaratunga C, et al. Genetic architecture of artemisinin-resistant plasmodium falciparum. Nat Genet. 2015;47:226–34.
38. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17:238.
39. Desjardins CA, Cohen KA, Munsamy V, Abeel T, Maharaj K, Walker BJ, et al. Genomic and functional analyses of mycobacterium tuberculosis strains implicate ald in D-cycloserine resistance. Nat Genet. 2016;48:544–51.
40. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1:16041.
41. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018;14:e1005958.
42. Jaillard M, Lima L, Tournoud M, Mahé P, van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: bridging the gap between k-mers and genetic events. PLoS Genet. 2018;14:e1007758.
43. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. Nat Rev Genet. 2017;18:41–50.
44. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, et al. Replicating genotype–phenotype associations. Nature. 2007;447:655–60.
45. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.

46. Read TD, Massey RC. Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. Genome Med. 2014;6:109.
47. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends Microbiol. 2010;18:315–22.
48. Lees JA, Vehkala M, Välimäki N, Harris SR, Chewapreecha C, Croucher NJ, et al. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. Nat Commun. 2016;7:1-8.
49. Farhat M, Moletta-Denat M, Frère J, Onillon S, Trouilhé MC, Robine E. Effects of disinfection on Legionella spp., eukarya, and biofilms in a hot water system. Appl Environ Microbiol. 2012;78:6850–8.
50. Dudbridge F, Gusnanto A. Estimation of significance thresholds for genomewide association scans. Genet Epidemiol. 2008;32:227–34.
51. Wheeler NE. Tracing outbreaks with machine learning. Nat Rev Microbiol. 2019. https://doi.org/10.1038/s41579-019-0153-1.
52. Fruzangohar M, Ebrahimie E, Adelson DL. A novel hypothesis-unbiased method for gene ontology enrichment based on transcriptome data. PLoS One. 2017;12:e0170486.
53. Park J, Taslim C, Lin S. BOG: R-package for bacterium and virus analysis of orthologous groups. Comput Struct Biotechnol J. 2015;13:366–9.
54. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of gene-ontology annotations with parent child analysis. Bioinformatics. 2007;23:3024–31.
55. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
56. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69.
57. Lee I-H, Lee K, Hsing M, Choe Y, Park J-H, Kim SH, et al. Prioritizing disease-linked variants, genes, and pathways with an interactive whole-genome analysis pipeline. Hum Mutat. 2014;35:537–47.
58. Alikhan N-F, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. PLoS Genet. 2018;14:e1007261.
59. Roer L, Hendriksen RS, Leekitcharoenphon P, Lukjancenko O, Kaas RS, Hasman H, et al. Is the Evolution of *Salmonella enterica subsp. enterica* Linked to Restriction-Modification Systems? mSystems. 2016;1:1-15.
60. Grimont P. Antigenic formulae of the Salmonella serovars. WHO Collab Cent Ref Res Salmonella. 9th ed; 2007. p. 1–166.
61. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, et al. The Salmonella in Silico typing resource (SISTR): an open web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. PLoS One. 2016;11:e0147101.
62. Nava MG, Hernandez YE. Comparative Genomics of Salmonella Could Reveal Key Features of Adaptation. J Data Min Genomics Proteomics. 2016;07:1-2.
63. Jünemann S, Prior K, Albersmeier A, Albaum S, Kalinowski J, Goesmann A, et al. GABenchToB: a genome assembly benchmark tuned on Bacteria and Benchtop sequencers. PLoS One. 2014;9:e107014.
64. Laing CR, Whiteside MD, Gannon VPJ. Pan-genome analyses of the species Salmonella enterica, and identification of genomic markers predictive for species, subspecies, and Serovar. Front Microbiol. 2017;8:1-16.
65. Jacobsen A, Hendriksen RS, Aaresturp FM, Ussery DW, Friis C. The Salmonella enterica Pan-genome. Microb Ecol. 2011;62:487–504.
66. Hollander D, Donyo M, Atias N, Mekahel K, Melamed Z, Yannai S, et al. A network-based analysis of colon cancer splicing changes reveals a tumorigenesis-favoring regulatory pathway emanating from ELK1. Genome Res. 2016;26:541–53.
67. Hedge J, Wilson DJ. Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. mBio. 2014;5:1-4. https://doi.org/10.1128/mBio.02158-14.
68. Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013;41:2073–94.
69. Hammarlöf DL, Kröger C, Owen SV, Canals R, Lacharme-Lora L, Wenner N, et al. Role of a single noncoding nucleotide in the evolution of an epidemic African clade of *Salmonella*. Proc Natl Acad Sci. 2018;115:E2614–23.
70. Pascoe B, Méric G, Yahara K, Wimalarathna H, Murray S, Hitchings MD, et al. Local genes for local bacteria: evidence of allopatry in the genomes of transatlantic *Campylobacter* populations. Mol Ecol. 2017;26:4497–508.
71. Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, et al. Recent human-to-poultry host jump, adaptation, and pandemic spread of Staphylococcus aureus. Proc Natl Acad Sci. 2009;106:19545–50.
72. Sheppard SK, Colles F, Richardson J, Cody AJ, Elson R, Lawson A, et al. Host Association of Campylobacter Genotypes Transcends Geographic Variation. Appl Environ Microbiol. 2010;76:5269–77.
73. Fitzgerald JR. Livestock-associated Staphylococcus aureus: origin, evolution and public health threat. Trends Microbiol. 2012;20:192–8.
74. Sheppard SK, Maiden MCJ. The evolution of *Campylobacter jejuni* and *Campylobacter coli*. Cold Spring Harb Perspect Biol. 2015;7:a018119.
75. Zhang S, Li S, Gu W, den Bakker H, Boxrud D, Taylor A, et al. Zoonotic source attribution of *Salmonella enterica* serotype Typhimurium using genomic surveillance data, United States. Emerg Infect Dis. 2019;25. https://doi.org/10.3201/eid2501.180835.
76. Rodriguez-Maillard JM, Arutyunov D, Frost LS. The F plasmid transfer activator TraJ is a dimeric helix-turn-helix DNA-binding protein: F TraJ binds DNA in vivo. FEMS Microbiol Lett. 2010;310:112–9.
77. Wilde A, Fiedler B, Börner T. The cyanobacterial phytochrome Cph2 inhibits phototaxis towards blue light. Mol Microbiol. 2002;44:981–8.
78. Prayitno D, Phillips C, Omed H. The effects of color of lighting on the behavior and production of meat chickens. Poult Sci. 1997;76:452–7.
79. Ghosh AK, Sen S, Sharma A, Talukder G. Effect of chlorophyllin on mercuric chloride-induced clastogenicity in mice. Food Chem Toxicol Int J Publ Br Ind Biol Res Assoc. 1991;29:777–9.
80. Lombardi G, Lanzirotti A, Qualls C, Socola F, Ali A-M, Appenzeller O. Five hundred years of mercury exposure and adaptation. J Biomed Biotechnol. 2012;2012:1–10.
81. Carravieri A, Cherel Y, Blévin P, Brault-Favrou M, Chastel O, Bustamante P. Mercury exposure in a large subantarctic avian community. Environ Pollut. 2014;190:51–7.
82. EFSA. Mercury as undesirable substance in animal feed - Scientific opinion of the Panel on Contaminants in the Food Chain: Mercury as undesirable substance in animal feed - Scientific opinion of the Panel on Contaminants in the Food Chain. Eur Food Saf Auth J. 2008;6:654.
83. Marangon S, Busani L. The use of vaccination in poultry production. Rev Sci Tech Int Off Epizoot. 2007;26:265–74.
84. Stone HD. Effect of thimerosal concentration on the efficacy of inactivated Newcastle disease oil-emulsion vaccines. Avian Dis. 1985;29:1030–5.
85. Santos JCN, da Silva IM, Braga TC, de Fátima Â, Figueiredo IM, Santos JCC. Thimerosal changes protein conformation and increase the rate of fibrillation in physiological conditions: spectroscopic studies using bovine serum albumin (BSA). Int J Biol Macromol. 2018;113:1032–40.
86. Barregard L, Rekić D, Horvat M, Elmberg L, Lundh T, Zachrisson O. Toxicokinetics of mercury after Long-term repeated exposure to Thimerosal-containing vaccine. Toxicol Sci. 2011;120:499–506.
87. Heinz E, Stubenrauch CJ, Grinter R, Croft NP, Purcell AW, Strugnell RA, et al. Conserved features in the structure, mechanism, and biogenesis of the inverse autotransporter protein family. Genome Biol Evol. 2016;8:1690–705.
88. Rhee KY, Opel M, Ito E, Hung S -p, Arfin SM, Hatfield GW. Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. Proc Natl Acad Sci. 1999;96:14294–9.
89. Nørgaard JV, Fernández JA. Isoleucine and valine supplementation of crude protein-reduced diets for pigs aged 5–8 weeks. Anim Feed Sci Technol. 2009;154:248–53.
90. Corzo A, Loar RE, Kidd MT. Limitations of dietary isoleucine and valine in broiler chick diets1. Poult Sci. 2009;88:1934–8.
91. He AS, Rohatgi PR, Hersh MN, Rosenberg SM. Roles of E. coli double-strand-break-repair proteins in stress-induced mutation. DNA Repair. 2006;5:258–73.
92. Spiga L, Winter MG. Furtado de Carvalho T, Zhu W, Hughes ER, Gillis CC, et al. An Oxidative Central Metabolism Enables Salmonella to Utilize Microbiota-Derived Succinate. Cell Host Microbe. 2017;22:291–301.
93. Gunn JS, Ryan SS, Van Velkinburgh JC, Ernst RK, Miller SI. Genetic and functional analysis of a PmrA-PmrB-regulated locus necessary for lipopolysaccharide modification, antimicrobial peptide resistance, and oral virulence of Salmonella enterica serovar typhimurium. Infect Immun. 2000;68:6139–46.
94. Velkov T, Roberts KD, Nation RL, Thompson PE, Li J. Pharmacology of polymyxins: new insights into an 'old' class of antibiotics. Future Microbiol. 2013;8:711–24.
95. Decolin D, Leroy P, Nicolas A, Archimbault P. Hyphenated liquid chromatographic method for the determination of colistin residues in bovine tissues. J Chromatogr Sci. 1997;35:557–64.
96. Khemici V, Linder P. RNA helicases in bacteria. Curr Opin Microbiol. 2016;30:58–66.

97. Lacey M, Agasing A, Lowry R, Green J. Identification of the YfgF MASE1 domain as a modulator of bacterial responses to aspartate. Open Biol. 2013;3:130046.
98. Bertin Y, Segura A, Jubelin G, Dunière L, Durand A, Forano E. Aspartate metabolism is involved in the maintenance of enterohaemorrhagic *Escherichia coli* O157:H7 in bovine intestinal content. Environ Microbiol. 2018; 20:4473–85.
99. Phan M-D, Kidgell C, Nair S, Holt KE, Turner AK, Hinds J, et al. Variation in Salmonella enterica Serovar Typhi IncHI1 plasmids during the global spread of resistant typhoid fever. Antimicrob Agents Chemother. 2009;53:716–27.
100. Jorgensen MG, Pandey DP, Jaskolska M, Gerdes K. HicA of Escherichia coli defines a novel family of translation-independent mRNA Interferases in Bacteria and Archaea. J Bacteriol. 2009;191:1191–9.
101. Cullen TW, O'Brien JP, Hendrixson DR, Giles DK, Hobb RI, Thompson SA, et al. EptC of campylobacter jejuni mediates phenotypes involved in host interactions and virulence. Infect Immun. 2013;81:430–40.
102. Conlin CA, Håkensson K, Liljas A, Miller CG. Cloning and nucleotide sequence of the cyclic AMP receptor protein-regulated Salmonella typhimurium pepE gene and crystallization of its product, an alpha-aspartyl dipeptidase. J Bacteriol. 1994;176:166–72.
103. Li Y, Han H, Yin J, Zheng J, Zhu X, Li T, et al. Effects of glutamate and aspartate on growth performance, serum amino acids, and amino acid transporters in piglets. Food Agric Immunol. 2018;29:675–87.
104. Kortman GAM, Boleij A, Swinkels DW, Tjalsma H. Iron availability increases the pathogenic potential of Salmonella Typhimurium and other enteric pathogens at the intestinal epithelial Interface. PLoS One. 2012;7:e29968.
105. Bäumler AJ, Norris TL, Lasco T, Voight W, Reissbrodt R, Rabsch W, et al. IroN, a novel outer membrane siderophore receptor characteristic of Salmonella enterica. J Bacteriol. 1998;180:1446–53.
106. Manhart CM, McHenry CS. The PriA replication restart protein blocks Replicase access Prior to helicase assembly and directs template specificity through its ATPase activity. J Biol Chem. 2013;288:3989–99.
107. Petersen C, Møller LB. The RihA, RihB, and RihC Ribonucleoside hydrolases of *Escherichia coli*: substrate specificity, gene expression, and regulation. J Biol Chem. 2001;276:884–94.
108. Song K-H, Kwak C-H, Chung T-W, Ha S-H, Park J-Y, Ha K-T, et al. Intestine specific regulation of pig cytidine-5′-monophospho-N-acetylneuraminic acid hydroxylase gene for N-glycolylneuraminic acid biosynthesis. Sci Rep. 2019; 9. https://doi.org/10.1038/s41598-019-40522-9.
109. Xu YT, Ma XK, Wang CL, Yuan MF, Piao XS. Effects of dietary valine:lysine ratio on the performance, amino acid composition of tissues and mRNA expression of genes involved in branched-chain amino acid metabolism of weaned piglets. Asian-Australas J Anim Sci. 2018;31:106–15.
110. Barbeyron T, Henrissat B, Kloareg B. The gene encoding the kappa-carrageenase of Alteromonas carrageenovora is related to β-1,3-1,4-glucanases. Gene. 1994;139:105–9.
111. Manuhara GJ, Praseptiangga D, Riyanto RA. Extraction and characterization of refined K-carrageenan of red algae [Kappaphycus Alvarezii (Doty ex P.C. Silva, 1996)] originated from Karimun Jawa Islands. Aquat Procedia. 2016;7:106–11.
112. Simms SA, Voige WH, Gilvarg C. Purification and characterization of succinyl-CoA: tetrahydrodipicolinate N-succinyltransferase from Escherichia coli. J Biol Chem. 1984;259:2734–41.
113. Ovie SO, Eze SS. Lysine requirement and its effect on the body composition of Oreochromis niloticus fingerlings. J Fish Aquat Sci. 2013;8:94–100.
114. Hauler RC, Carter CG. Reevaluation of the quantitative dietary lysine requirements of fish. Rev Fish Sci. 2001;9:133–63.
115. Harrison PW, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Holt S, et al. The European Nucleotide Archive in 2018. Nucleic Acids Res. 2018;47:D84-88.
116. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
117. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, et al. Using Drosophila melanogaster as a model for Genotoxic chemical mutational studies with a new program, SnpSift. Front Genet. 2012;3:1–35.
118. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Landes Biosci. 2012;6:1–13.
119. Bushnell B. BBMap: A Fast, Accurate, Splice-Aware Aligner. Berkeley Lab. 2014;Report Number: LBNL-7065E.
120. Xu S, Ackerman MS, Long H, Bright L, Spitze K, Ramsdell JS, et al. A male-specific genetic map of the microcrustacean Daphnia pulex based on single-sperm whole-genome sequencing. Genetics. 2015;201:31–8.
121. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinforma. 2018;:http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
122. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.
123. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.
124. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, et al. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016;17:132.
125. Bosi E, Donati B, Galardini M, Brunetti S, Sagot M-F, Lió P, et al. MeDuSa: a multi-draft based scaffolder. Bioinformatics. 2015;31:2443–51.
126. Kosugi S, Hirakawa H, Tabata S. GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. Bioinformatics. 2015;31(23):3733–41 btv465.
127. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25:1422–3.
128. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29:1072–5.
129. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32:3047–8.
130. Wickham H. ggplot2. New York: Springer New York; 2009. https://doi.org/10.1007/978-0-387-98141-3.
131. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31:3691–3.
132. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32:268–74.
133. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14:587–9.
134. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018; 35:518–22.
135. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30:1188–95.
136. Sokal RR, Rohlf FJ. The comparison of Dendrograms by objective methods. Taxon. 1962;11:33.
137. Revell LJ. Phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. Methods Ecol Evol. 2012;3:217–23.
138. Fowlkes EB, Mallows CL. A method for comparing two hierarchical Clusterings. J Am Stat Assoc. 1983;78:553–69.
139. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53:131–47.
140. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, et al. Genome-wide association study of HIV whole genome sequences validated using drug resistance. PLoS One. 2016;11:e0163746.
141. Ioannidis JPA, Thomas G, Daly MJ. Validating, augmenting and refining genome-wide association signals. Nat Rev Genet. 2009;10:318–29.
142. Johnson NL, Kotz S, Kemp AW, Johnson NL. Univariate discrete distributions. 2nd ed. New York: Wiley; 1992.
143. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34:502–8.

Publisher's Note

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