Felten et al. BMC Microbiology (2017) 17:222
DOI 10.1186/s12866-017-1132-1

BMC Microbiology

METHODOLOGY ARTICLE Open Access

First gene-ontology enrichment analysis based on bacterial coregenome variants: insights into adaptations of *Salmonella* serovars to mammalian- and avian-hosts

Arnaud Felten, Meryl Vila Nova, Kevin Durimel, Laurent Guillier, Michel-Yves Mistou and Nicolas Radomski*

Abstract

Background: Many of the bacterial genomic studies exploring evolution processes of the host adaptation focus on the accessory genome describing how the gains and losses of genes can explain the colonization of new habitats. Consequently, we developed a new approach focusing on the coregenome in order to describe the host adaptation of *Salmonella* serovars.

Methods: In the present work, we propose bioinformatic tools allowing (i) robust phylogenetic inference based on SNPs and recombination events, (ii) identification of fixed SNPs and InDels distinguishing homoplastic and non-homoplastic coregenome variants, and (iii) gene-ontology enrichment analyses to describe metabolic processes involved in adaptation of *Salmonella enterica* subsp. *enterica* to mammalian- (*S.* Dublin), multi- (*S.* Enteritidis), and avian- (*S.* Pullorum and *S.* Gallinarum) hosts.

Results: The 'VARCall' workflow produced a robust phylogenetic inference confirming that the monophyletic clade *S.* Dublin diverged from the polyphyletic clade *S.* Enteritidis which includes the divergent clades *S.* Pullorum and *S.* Gallinarum (i). The scripts 'phyloFixedVar' and 'FixedVar' detected non-synonymous and non-homoplastic fixed variants supporting the phylogenetic reconstruction (ii). The scripts 'GetGOxML' and 'EveryGO' identified representative metabolic pathways related to host adaptation using the first gene-ontology enrichment analysis based on bacterial coregenome variants (iii).

Conclusions: We propose in the present manuscript a new coregenome approach coupling identification of fixed SNPs and InDels with regards to inferred phylogenetic clades, and gene-ontology enrichment analysis in order to describe the adaptation of *Salmonella* serovars Dublin (i.e. mammalian-hosts), Enteritidis (i.e. multi-hosts), Pullorum (i.e. avian-hosts) and Gallinarum (i.e. avian-hosts) at the coregenome scale. All these polyvalent Bioinformatic tools can be applied on other bacterial genus without additional developments.

Keywords: Bacterial genomics, Bacterial fixed variants, Gene-ontology enrichment analysis

* Correspondence: nicolas.radomski@anses.fr
Université PARIS-EST, Anses, Laboratory for food safety, Maisons-Alfort, France

Background

Recent advances in bacterial genomics aim to identify small insertions/deletions (InDels) of the coregenome [1] in addition to single nucleotide polymorphisms (SNPs) [2] and offer a unique opportunity to perform the first gene-ontology enrichment analysis based on sensitive and specific variants previously identified according to inferred bacterial clades [3]. These variants correspond to spontaneous mutations or recombination evens [4] which are usually identified thanks to the detection of SNP hotspots [5]. Rare genetic variants near the final leaves of the phylogeny [6] are described as unstable [7] or transient variants [8], in opposite to fixed variants of the branches which are passed on to descendants in the absence of recombination events [9].

Representing several tens of Bioinformatics tools developed since the beginning of the twenty-first century, the gene-ontology enrichment analysis allows identification of relevant biological processes among large lists of genes [10]. Covering Eukaryotic and Prokaryotic organisms [11], the gene-ontology enrichment analysis can be applied on lists of annotated genes or variants [12]. Based on several annotation databases [13], such as Gene Ontology (GO) [14] where the GO-terms form a directed acyclic graph (DAG) organized as three independent ontologies of GO-terms, namely biological process (BP), molecular function (MF), cellular component (CC), different kinds of gene-ontology enrichment analyses must be distinguished. The enrichment likelihood may indeed be calculated using Chi-square, Fisher's exact test, Binomial probability based on preselected interesting gene lists (i.e. a singular enrichment analysis), entire genes with associated experimental values (i.e. a gene set enrichment analysis) [10], or using hypergeometric distribution based on the approaches term-for-term (i.e. a modular enrichment analysis excluding dependencies between the GO-terms) and parent-child (i.e. a modular enrichment analysis including dependencies between the GO-terms) [15].

With around 80 million cases annually, *Salmonella enterica*, especially *Salmonella enterica* subsp. *enterica* serovars, is the main cause of foodborne gastroenteritis [16]. Host adaptation of *Salmonella* is related to acquired pathogenicity islands (PAIs) called *Salmonella* pathogenicity islands (SPIs), and genes involved in intestinal phase of infection, colonization of deeper tissues, and expansion in host range, successively [17]. Several studies tend to theorize that the gains of large genetic elements by horizontal gene transfer (e.g. plasmids, transposons and phages) would expand the host range of *S. enterica* subsp. *enterica* serovars, especially those causing severe human infections (e.g. *S.* Typhimurium and *S.* Enteritidis), while the reduction of host range (e.g. *S.* Typhi in humans, *S.* Dublin in cattle and humans, *S.* Gallinarum in avian hosts) would mainly be explained by the losses of genes due to large deletions and accumulation of pseudogenes (i.e. start/stop gained codons and/or frameshift mutations) [18]. The few genomic approaches aiming for description of host adaptation between [19–22] or into *Salmonella* serovars [4, 23–25], focused on the accessory genome which refers to gains and losses of large genetic elements [26].

The evolution of serovars, as Agona, Dublin, Hadar, Heidelberg, Javiana, Kentucky, Newport, Saintpaul, Schwarzengrund, Virchow, Weltevreden, Choleraesuis, Enteritidis, Gallinarum, Paratyphi A, Paratyphi B, Paratyphi C, Typhi and Typhimurium, seems to be highly mediated by losses of coding sequences which lead to functional reduction, and horizontal acquisitions of plasmid and phage sequences, hypothetically under regulation of a defense mechanism against exogenous invading sequences, particularly through clustered regularly interspaced short palindromic repeats (CRISPR) [20]. With the exception of structural RNA genes and transposable elements, large stable duplications are rare for *Salmonella* and other *Enterobacteriaceae* (e.g. locus *ccm* of 7.5 kb in S. Typhimurium) [27], and many *S. enterica* serovars, as Abortusequi, Abortusovis, Choleraesuis, Dublin, Enteritidis, Gallinarum, Pullorum, Sendai, Paratyphi C, and Typhimurium, are known to harbor different virulence plasmids [21] heterogeneous by size (50–90 kb) but all sharing five genes (i.e. *spvR*, *spvA*, *spvB*, *spvC*, and *spvD*) in the 7.8-kb *spv* region (*Salmonella* plasmid virulence) which are required for bacterial multiplication in the reticulo-endothelial system [28] and cytopathic effect in the human macrophages [29].

Known as a common serovar involved in domestically acquired foodborne illness in humans, *S.* Enteritis is introduced in the food chain by products from diverse hosts [22]. In contrast, foodborne illness in humans caused by *S.* Dublin is mainly due to contaminated cow's milk and cheese [30], while *S.* Gallinarum [31] and *S.* Pullorum [22] infect avian-hosts. According to a microarray study dividing *S.* Enteritidis into two major clades [19], a recent phylogenetic inference based on SNPs of the coregenome (reference genome: *S.* Enteritidis strain P125109, NC_011294.1) [22] demonstrated that *S.* Dublin and *S.* Enteritidis diverged from a most recent common ancestor (MRCA), and that *S.* Enteritidis was constituted of two clades (e.g. classical and unclassical), one of them being the origin of the avian-adapted serovars *S.* Gallinarum and *S.* Pullorum [22]. Focusing on gene acquisition and functional gene losses explaining host specialization of these bacterial pathogens, Langridge et al. emphasized that evolution of *Salmonella* pathogenicity is strongly associated with the acquisition of SPIs [22], especially SPI-6 and SPI-19 which are mobile genetic elements encoding type VI secretion systems

(T6SSs) [31]. While T6SSs in SPI-19 of *S.* Gallinarum contributes to the colonization of the gastrointestinal tract of chickens, as demonstrated with chickens orally infected by bacterial mutants [31], several genes encoding exported proteins are deleted in SPI-19 of the multi-hosts adapted *S.* Enteritidis, but were intact in the mammalian-adapted serovar *S.* Dublin (i.e. cattle and humans), as well as the avian-adapted serovars *S.* Gallinarum and *S.* Pullorum [22]. Among molecular mechanisms involved in fimbriae (i.e. *std.*Δ, *stiC*, *stfF*, *safC*, *stbC*, *pegC*, *lpfC*, *sefD*, *sefC*, *sthB*, *sthA*, *sthE* in *S.* Gallinarum [25]), the partial losses of *ssf* and *saf* operons are highly specific to avian adaptation of *S.* Gallinarum and *S.* Pullorum [22].

Here, we propose a novel approach coupling identification of clade specific SNPs and InDels [1, 2] and associated gene-ontology enrichment analyses [14, 15] in order to link evolutionary history and biological function. By targeting the coregenome, we propose an additional and complementary method to analyses focusing on the accessory genome only. We tested its performance by studying the host-adaptation processes at the coregenome scale on a dataset constituted by Langridge et al. [22] in order to describe at the accessory genome scale the host adaptations of *Salmonella* serovars: Enteritidis (i.e. multi-hosts), Dublin (i.e. mammalian-hosts), Gallinarum (i.e. avian-hosts) and Pullorum (i.e. avian-hosts) [19, 22, 25].

Results

Robustness of the variant dataset

After trimming and mapping of paired-end reads, the previously published genome dataset (Additional file 1) displayed satisfactory depth (i.e. 175X) and breadth (i.e. 99 ± 1%) coverage (Additional file 2) against the reference genome *S.* Enteritidis (strain P125109) [32]. This depth coverage (i.e. >30X) is sufficient to accurately detect InDels with GATK [32], and this breadth coverage is quantitatively in accordance with results obtained by mapping of *S.* Typhimurium reads against the reference LT2 genome (i.e. 95% mean, 91% min., 98% max.) [24].

The set of sequencing data corresponding to 59 *Salmonella* genomes was analyzed through the variant calling workflow 'VARCall' where high-confidence variants were retained [33]. We identified 14,086 coregenome variants among the four serovars corresponding to 12,929 SNPs and 1157 InDels. Even if InDels are considered more challenging to be called accurately than SNPs [34], the ratio of 0.9 InDels/10 SNPs identified by the 'VARCall' workflow on the current *Salmonella* dataset falls inside the boundaries of 0.5–2 InDels / 10 SNPs which have been previously observed in other prokaryotic species (e.g. *Blochmannia vafer* [35], *Streptococcus* [36], *Salmonella* Typhimurium [24]) and eukaryotic studies (e.g. Rice [37], e.g. Human [38]).

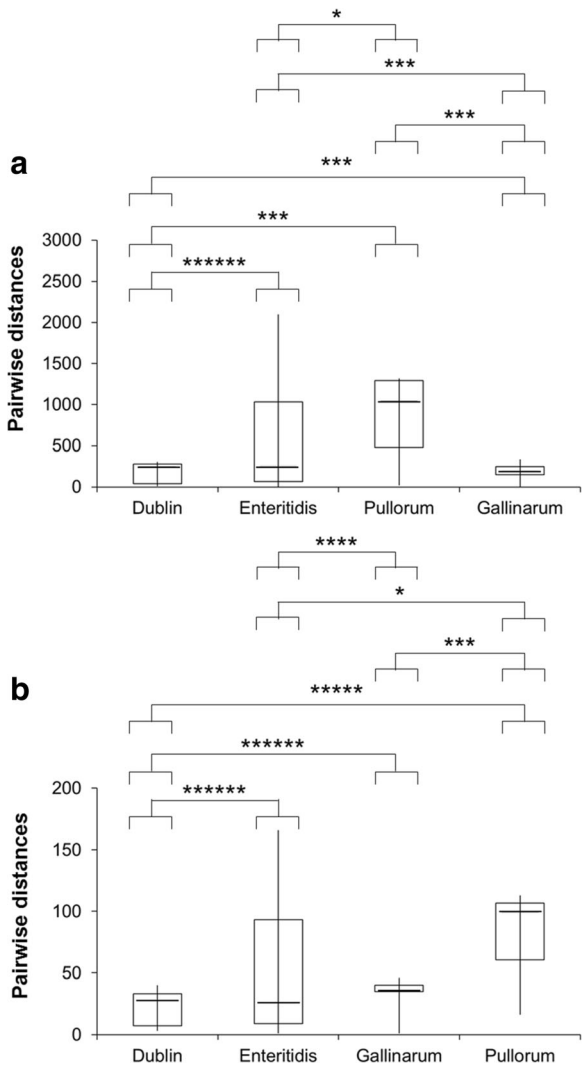

Fig. 1 Boxplots (median, 25th percentile, 75th percentile, minimum and maximum) of pairwise distances expressed as single nucleotide polymorphisms (SNPs) (**a**) or small insertions/deletions (InDels) (**b**) into *Salmonella enterica* subsp. *enterica* serovars Dublin ($n = 60$), Enteritidis ($n = 528$), Pullorum ($n = 10$) and Gallinarum ($n = 28$). Normality of distribution and equality of variances were checked with Shapiro-Wilk and Fisher tests, respectively. Statistical differences (*: $p < 5.0\times10^{-2}$; **: $p < 1.0\times10^{-2}$; ***: $p < 1.0\times10^{-3}$; ****: $p < 1.0\times10^{-4}$; *****: $p < 1.0\times10^{-5}$; ******: $p < 1.0\times10^{-6}$) are calculated with Wilcoxon rank sum (i.e. non-normal distribution with equality of variances) or Kolmogorov-Smirnov (i.e. non-normal distribution without equality of variances) tests

We found that 12,031 variants (11,285 SNPs and 746 InDels) were intragenic, corresponding to 85% of the 14,086 coregenome variants (12,929 SNPs and 1157 InDels). As the coding regions represent about 88% of the *S.* Enteritidis chromosome (strain P125109, accession NC_011294.1) [25], intra and intergenic regions of the chromosome seem to be similarly impacted by variants. As observed by Zhou et al. [4] with 37% of intergenic

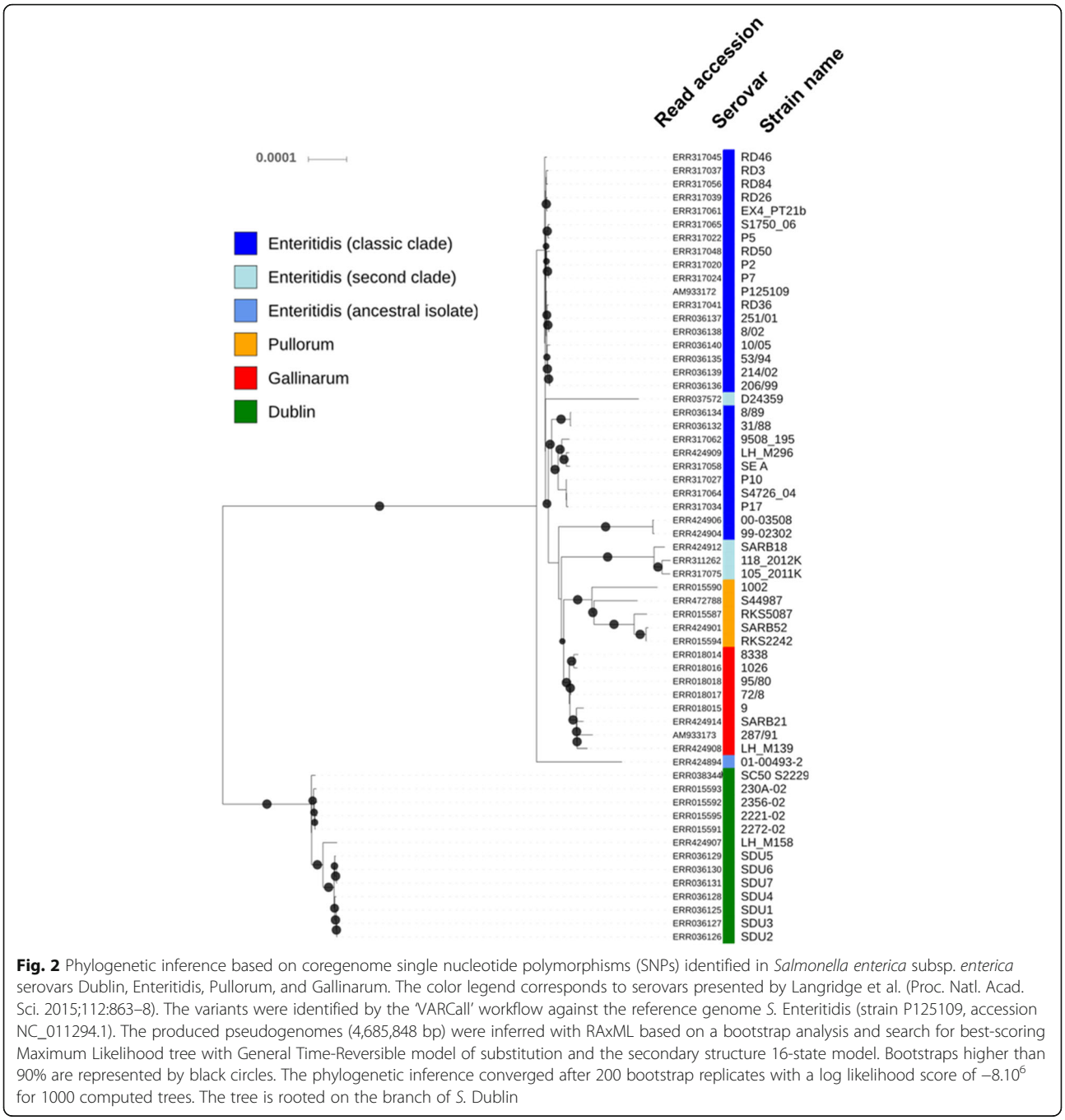

Fig. 2 Phylogenetic inference based on coregenome single nucleotide polymorphisms (SNPs) identified in *Salmonella enterica* subsp. *enterica* serovars Dublin, Enteritidis, Pullorum, and Gallinarum. The color legend corresponds to serovars presented by Langridge et al. (Proc. Natl. Acad. Sci. 2015;112:863–8). The variants were identified by the 'VARCall' workflow against the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The produced pseudogenomes (4,685,848 bp) were inferred with RAxML based on a bootstrap analysis and search for best-scoring Maximum Likelihood tree with General Time-Reversible model of substitution and the secondary structure 16-state model. Bootstraps higher than 90% are represented by black circles. The phylogenetic inference converged after 200 bootstrap replicates with a log likelihood score of -8.10^6 for 1000 computed trees. The tree is rooted on the branch of *S.* Dublin

InDels (i.e. 37/100) across 73 genomes of *S.* Agona serovar, we also detected 35% of intergenic InDels (i.e. 411/1157) across the 59 studied genomes.

Interestingly, the average pairwise distances of SNPs or InDels were significantly different (Fig. 1) between all the combinations of studied serovars ($p < 5.0\times10^{-2}$, Wilcoxon rank sum or Kolmogorov-Smirnov tests). For instance, the average pairwise distances of *S.* Enteritidis clade (531 ± 512 SNPs; 51 ± 48 InDels) were significantly different than those calculated for Pullorum (894 ± 501 SNPs; 84 ± 34 InDels), Gallinarum (195 ± 78 SNPs; 35 ± 9 InDels) and Dublin (163 ± 120 SNPs; 21 ± 13 InDels) serovars ($p < 5.0\times10^{-2}$, Wilcoxon rank sum or Kolmogorov-Smirnov tests).

Robustness of the phylogenetic inference

The phylogenetic inference converged after 200 bootstrap replicates [39] with a log likelihood score of -8.10^6

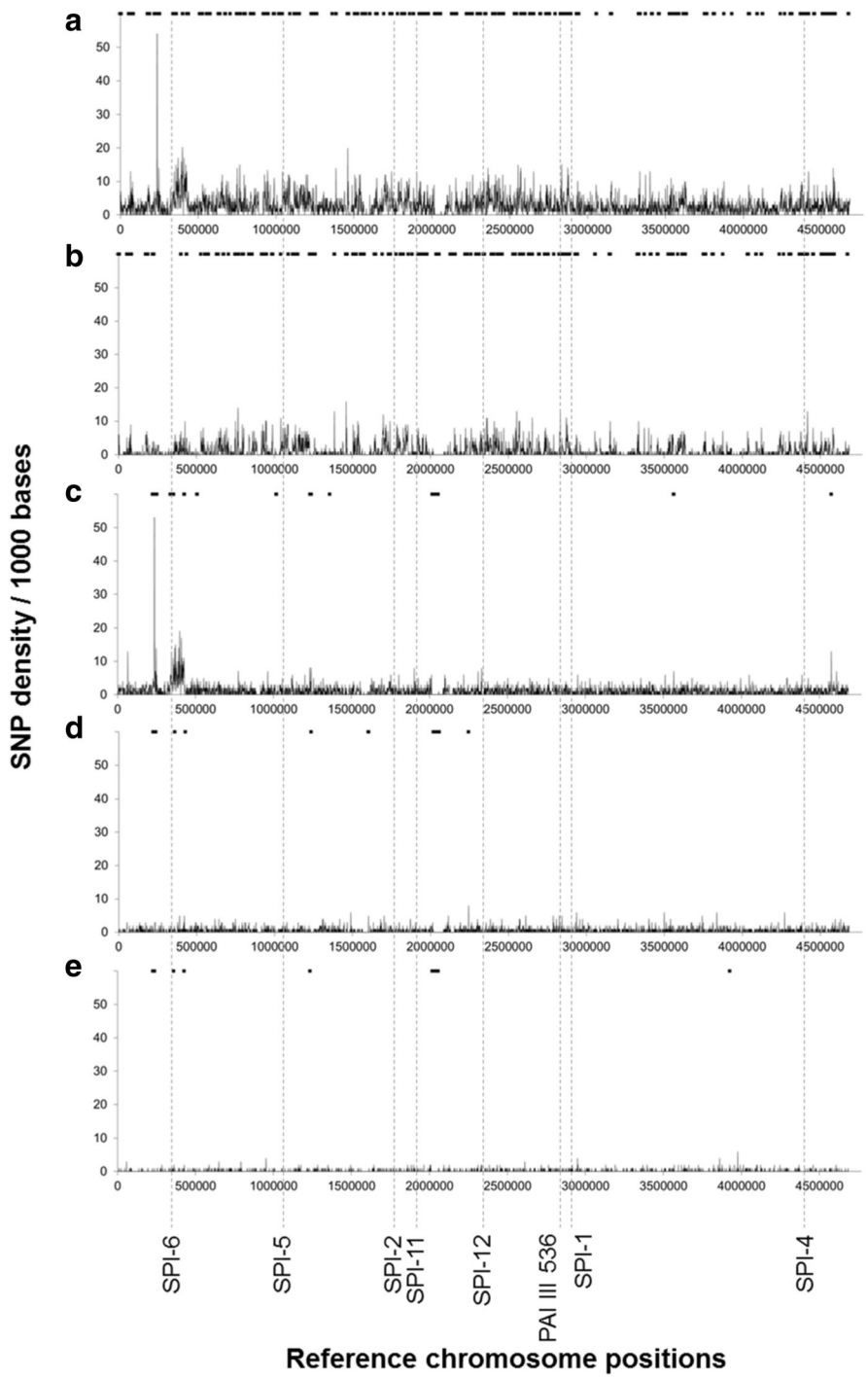

Fig. 3 Densities of single nucleotide polymorphisms (SNPs) per 1000 bp (curves), *Salmonella* pathogenic islands (dotted lines), and recombination events (rectangles) across *Salmonella enterica* subsp. *enterica* serovars (**a**: 59 genomes, 12,929 SNPs), including Dublin (**b**: 13 genomes, 5084 SNPs), Enteritidis (**c**: 33 genomes, 5136 SNPs), Pullorum (**d**: 5 genomes, 2225 SNPs), and Gallinarum (**e**: 8 genomes, 671 SNPs). Pathogenicity island database from KonKuk University (Seoul, South Korea) were used to detect *Salmonella* Pathogenic Islands (SPIs) SPI-1 (2890501–2,934,879), SPI-2 (1727425–1,769,273), SPI-4 (4333507–4,361,514), SPI-5 (1053174–1,074,167), SPI-6 (299796–330,890), SPI-11 (1904313–1,912,607), SPI-12 (2328077–2,347,757) and PAI III 536 (2801306–2,810,695) of the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1)

for 1000 computed trees [40], resulting in best-scoring Maximum Likelihood tree with most of branches presenting bootstraps higher than 90% (Fig. 2) and emphasizing the robustness of the tree reconstructed with SNPs detected by the 'VARCall' workflow.

The phylogenetic tree (Fig. 2) showed the mammalian-adapted serovar *S.* Dublin as a distinct clade with differences of 4945 ± 378 SNPs and 296 ± 17 InDels from the *S.* Enteritidis clade, as confirmed by our genetic structure analysis [41] and according to Langridge et al. [22]. Excepting few differences in terms of genome composition, our phylogenetic inferences excluding or including SNPs from recombination events (Additional file 3) showed two clades in *S.* Enteritidis described as classic and second clades by Langridge et al. [22] (Fig. 2). In addition to previous studies distinguishing these two clades of *S.* Enteritidis by microarray [19] and SNP-based approach [22], we supported also this dichotomy by a recombination event of 414 bp (node K in Additional files 4 and 5) harboring genes (POS 2018948–2,019,361) involved in DNA-binding (SEN_RS23380) and unknown function (SEN_RS09975).

Host adaptation is not associated with a reduced coregenome diversity

S. Pullorum and *S.* Gallinarum are highly specialized toward avian-hosts; they are non-motile members of serogroup D, indistinguishable by serotyping and historically recognized as distinct biotypes considering several criteria: the distinct septocaemic diseases they cause in avian species, biochemical characteristics and multilocus enzyme electrophoresis [42]. As previously shown by Langridge et al. [22], our phylogenetic inference indicates that *S.* Pullorum and *S.* Gallinarum isolates arose from the second clade of *S.* Enteritidis (Fig. 2). Interestingly the pairwise distances of SNPs or InDels inside the two groups were significantly different (Fig. 1): the *S.* Gallinarum clade was homogeneous (195 ± 78 SNPs; 35 ± 9 InDels), while by contrast the *S.* Pullorum clade displayed a larger significant diversity with regards to SNPs (894 ± 501; $p = 4.6\times10^{-13}$; Fisher test) or InDels (84 ± 34; $p = 1.4\times10^{-7}$; Fisher test). These results confirm at the genetic level the biochemical classification performed by Crichton and Old [43].

All together, these observations are in accordance with Langridge et al. [22] who concluded that *S.* Dublin and *S.* Enteritidis have a most recent common ancestor, and that the clade *S.* Pullorum/*S.* Gallinarum diverged from *S.* Enteritidis. According to Langridge et al. [22], we also confirmed that the genome ERR424894 is an ancestral isolate of *S.* Enteritidis (Fig. 2 and Additional file 4). In opposite to studies theorizing that gains of large genetic elements by horizontal gene transfer would expand large host range of *S. enterica* subsp. *enterica* serovars, while the reduction of host range would mainly be explained by losses of genes due to large deletions and accumulation of pseudogenes [18], we observed that the coregenome diversities of the multi-hosts adapted serovar *S.* Enteritidis (Fig. 1), expressed in SNPs or InDels, were not significantly different than those of the avian-adapted serovar *S.* Pullorum (SNPs: $p = 0.985$; InDels: $p = 0.262$; Fisher test), as well as significantly higher than those of the mammalian-adapted serovar *S.* Dublin and avian-adapted serovar *S.* Gallinarum (SNPs: $p < 4.6\times10^{-13}$; InDels: $p < 7.9\times10^{-15}$; Fisher test).

Accumulation of coregenome recombination events in the ancient branches

Following lateral transfer of genetic material, recombination events are a common modality of evolution of bacterial genomes [44]. In *Salmonella*, the PAIs called SPIs have long been recognized to play a determinant role in the virulence properties of the pathogen [45]. These regions are likely acquired by lateral transfer and can be excised from the chromosome by site-specific recombination events [46]. We performed a specific recombination analysis on the whole dataset to identify branch specific recombination events that could explain some characteristics of the corresponding isolates. Detecting high densities of SNPs (Fig. 3), we identified 112 recombination events associated to 12 nodes of the phylogenetic inference (Additional files 4 and 5).

Surprisingly, most of these recombination events ($n = 91$) were detected in *S.* Dublin, the most distant clade of the collection (node A in Additional files 4 and 5). On the other hand, only 21 recombination signatures were detected in the 11 other nodes of the phylogenetic inference, indicating that recombination events are rare among the Enteritidis/Pullorum/Gallinarum serovars. A similar observation was made for the Agona serovar [4]. Certain recombination events were clade-specific, like those detected at nodes E and H, which are specific for *S.* Gallinarum or *S.* Enteritidis/*S.* Pullorum/*S.* Gallinarum, respectively (Additional files 4 and 5). For instance, the recombination event specific to *S.* Gallinarum impacted a segment of 418 bp harboring genes involved in the synthesis of multidrug resistance proteins (SEN_RS19005, SEN_RS19010).

We also detected two recombination events at node H where *S.* Enteritidis/*S.* Pullorum/*S.* Gallinarum split from *S.* Dublin (Additional files 4 and 5). This *S.* Enteritidis/*S.* Pullorum/*S.* Gallinarum (node H in Additional file 4) specific recombination events concerns a 4012 bp DNA sequence corresponding to genes encoding the FhuBCD ATP-dependent iron transport system and a 67,080 bp region carrying genes involved in the production of fimbrial proteins and transport of multidrug (Additional files 4 and 5).

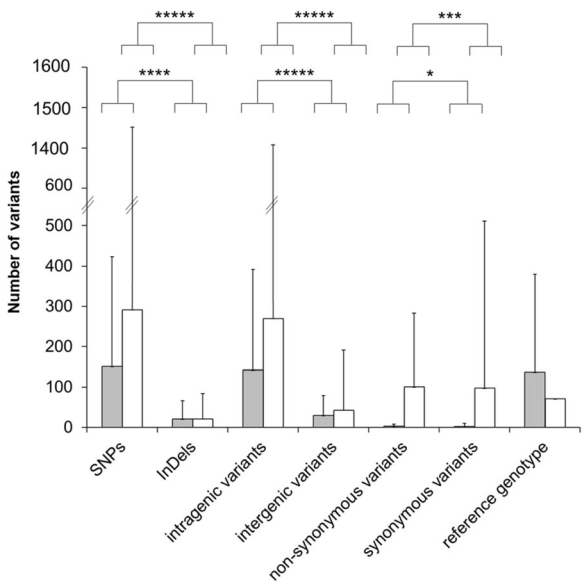

Fig. 4 Homoplastic (grew bars) and non-homoplastic (white bars) variants (SNPs *versus* InDels, intragenic *versus* intergenic, non-synonymous *versus* synonymous) fixed across all branches of the phylogenetic inference including genomes of *Salmonella enterica* subsp. *enterica* serovars Enteritidis ($n = 33$), Pullorum (n = 5), Gallinarum ($n = 8$) and Dublin ($n = 13$). The variant annotation was performed with SnpEff against reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The fixed non-homoplastic variants are defined by common genotypes across the considered group of genomes, as well as different genotypes in all the others compared genomes. The fixed homoplastic variants are defined by common genotypes across the considered group of genomes and genomes of independent phylogenetic clades, as well as different genotypes in genomes of the compared child-branches. The term 'reference genotype' refers to fixed variants presenting genotype of the reference genome. This analysis was performed with the script 'phyloFixedVar' (i.e. dependently of the phylogenetic inference). Statistical differences (*: $p < 1.0\times10^{-6}$; **: $p < 1.0\times10^{-7}$; ***: $p < 1.0\times10^{-8}$; ****: $p < 1.0\times10^{-9}$; *****: $p < 1.0\times10^{-10}$) are calculated with Wilcoxon signed rank tests. The vertical bars represent the standard deviation

Phylogenetically relevant fixed variants

With the general aim to link functional information to the genetic variations identified by the 'VARCall' workflow, we retrieved the SNPs and InDels which are specific and sensitive (i.e. fixed variants) either of branches in the phylogenetic tree (i.e. 'PhyloFixedVar') or between groups defined by the user (i.e. 'FixedVar').

The 'phyloFixedVar' and 'FixedVar' scripts retrieve SNPs and InDels and distinguish intragenic from intergenic variants, as well as homoplastic and non-homoplastic fixed variants at each branch of the phylogenetic tree. Additional information related to impact on protein translation was associated to each fixed variant (e.g. synonymous, non-synonymous, missense, frameshift). According to a study on *S.* Typhimurium estimating non-synonymous SNPs at 44% [24], most of the branches presented significantly more synonymous fixed variants that non-synonymous fixed variants (Fig. 4) with regard to non-homoplastic ($p < 1.0\times10^{-8}$; Wilcoxon signed rank test) variants (Fig. 4).

Focusing on branches having accumulated coregenome variants during evolution of *S.* Dublin, *S.* Enteritidis, *S.* Gallinarum and *S.* Pullorum (Fig. 2 and Additional file 4), most of fixed variants were identified at the node where *S.* Dublin and *S.* Enteritidis diverged (Table 1), as expected and whatever the considered type of variant (SNPs or InDels, intragenic or intragenic, homoplastic or non-homoplastic). For instance, fixed non-homoplastic InDels appeared in *S.* Enteritidis during the divergence with *S.* Dublin in two small intergenic regions between genes encoding a 4-hydroxy-2-oxo-heptane-1,7-dioate aldolase and a 4-hydroxyphenylacetate permease, or genes encoding a tripeptide permease A and a glutathione S-transferase, as well as in a gene encoding the fimbrial outer membrane usher protein SthC (Additional file 6).

Considered in this case as homoplastic variants, the divergence of the clade Pullorum/Gallinarum with *S.* Enteritidis is also supported by these two InDels impacting these two small intergenic regions between pairs of genes encoding a 4-hydroxy-2-oxo-heptane-1,7-dioate aldolase and a 4-hydroxyphenylacetate permease, or a tripeptide permease A and a glutathione S-transferase (Table 1 and Additional file 6). Compared to *S.* Gallinarum, only one homoplastic and non-synonymous SNP has been fixed in *S.* Pullorum (Table 1) impacting a gene coding an hypothetical protein, as well as five InDels impacting genes encoding a type I secretion system permease/ATPase, an effector protein YopJ, a TonB-dependent receptor, a membrane protein and a 5-keto-4-deoxyuronate isomerase (Additional file 6). In comparison with *S.* Pullorum, only one non-homoplastic and non-synonymous SNP was fixed in *S.* Gallinarum impacting a gene coding a ribonuclease Z, as well as three non-homoplastic SNPs impacting two intergenic regions between pairs of genes coding a transcriptional regulator and a NAD^+ synthetase, or an hypothetical protein and an α-glucosidase (Table 1 and Additional file 6).

Because several tens or hundreds of variants are fixed at these branches of interest (Table 1), we developed a tool in order to perform the first gene-ontology enrichment analyses based on sensitive and specific variants which are defined in the present study as intragenic and non-homoplastic fixed variants.

Gene-ontology enrichment analyses

Gene-ontology enrichment analysis was developed to identify relevant biological processes associated to large number of biological objects [10–12]; our aim in the present work was to link fixed variants to GO-terms.

Table 1 Single nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels) fixed at phylogenetic branches where genomes of *Salmonella enterica* subsp. *enterica* serovars Enteritidis ($n = 33$), Pullorum ($n = 5$), Gallinarum ($n = 8$) and Dublin ($n = 13$) diverged

	Serovars	Variants							Total
		Intragenic					Intergenic		
		sSNP	nsSNP	nsInDel	rSNP	rInDel	SNP	InDel	
Homoplastic	Dublin *versus* all [a]	0	0	0	0	0	0	0	0
	Enteritidis *versus* Dublin [b]	0	0	0	3948	93	439	117	4597
	Pullorum + Gallinarum *versus* Enteritidis [b]	0	0	0	0	0	0	2	2
	Pullorum *versus* Gallinarum [a]	1	1	5	6	113	4	47	117
	Gallinarum *versus* Pullorum	0	0	8	236	84	16	38	382
Non-homoplastic	Dublin versus all [a]	3129	819	87	0	0	438	115	4588
	Enteritidis *versus* Dublin [b]	0	0	0	0	1	0	2	3
	Pullorum + Gallinarum *versus* Enteritidis [b]	0	0	31	0	0	0	5	36
	Pullorum *versus* Gallinarum [a]	95	139	81	0	0	15	27	357
	Gallinarum *versus* Pullorum [a]	5	1	108	0	0	3	38	155

The variant calling analysis was performed with the 'VARCall' workflow (i.e. 12,929 SNPs and 1157 small InDels). The fixed non-homoplastic variants are defined by common genotypes across the considered group of genomes, as well as different genotypes in all the others compared genomes. The fixed homoplastic variants are defined by common genotypes across the considered group of genomes and genomes of independent phylogenetic clades, as well as different genotypes in genomes of the compared child-leaves. According to the variant annotation performed with SnpEff against reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1), the fixed variants presenting reference (r) and alternative (synonymous: s; non-synonymous: ns) genotypes are presented

[a]Analysis performed with the script 'phyloFixedVar' (i.e. dependently of the phylogenetic inference)

[b]Analysis performed with the script 'FixedVar' (i.e. independently of the phylogenetic inference)

In order to retrieve biological functions impacted during evolution of the studied serovars, we performed four gene-ontology enrichment analyses: *S.* Dublin compared to all the 3 other serovars (Ontology 1 called 'Dub_All'), divergence of *S.* Pullorum/Gallinarum from *S.* Enteritidis (Ontology 2 called 'Ent_Pull/Gall'), divergences of *S.* Pullorum (Ontology 3 called 'Pull_Gall') and *S.* Gallinarum (Ontology 4 called 'Gall_Pull') with each other (Table 1). Our analysis focused on 4035, 31, 315 and 114 intragenic and non-homoplastic fixed variants specific of the 'Dub_All', 'Ent_Pull/Gall', 'Pull_Gall' and 'Gall_Pull' divergences, respectively; leading to the retrieval of 1841, 195, 1034, 1034 GO-terms that were significantly enriched (Additional file 7). The enriched GO-terms of the 'Dub_All' divergence represented biological process (54%) and cellular component (40%), while those of the other analyses corresponded to biological process (61 ± 3%) and molecular function (32 ± 3%) (Additional file 7). For further analysis we selected the most relevant GO-terms based on accuracy (i.e. GO-level ≥ 5), number of hits (i.e. ≥ 4) and *p*-value (i.e. $< 5.0\times10^{-2}$) (Table 2). Through this selection process we reduced the number of highly relevant GO-terms to 39, 2, 8, and 8 for 'Dub_All', 'Ent_Pull/Gall', 'Pull_Gall' and 'Gall_Pull' ontologies, respectively.

With only 3 non-homoplastic InDels fixed in two intergenic regions and in a gene coding the fimbrial outer membrane usher protein SthC (Table 1 and Additional file 6), the multi-hosts adapted serovar *S.* Enteritidis can be considered as a polyphyletic clade including the avian-adapted serovars *S.* Pullorum and *S.* Gallinarum (Fig. 2 and Additional file 4).

The *S.* Dublin monophyletic clade diverged from the *S.* Enteritidis polyphyletic clade (Fig. 2 and Additional file 4) by accumulating 4035 specific fixed variants (Table 1). These variants were found to be associated with 39 GO-terms mainly involved in central metabolism pathways and especially those of amino acids metabolism (Table 2). Hence, among these 39 GO-terms, 23 were directly related to amino acid metabolism and more specifically the catabolic processes of proline and arginine to glutamate (Table 2).

A similar analysis performed on the avian adapted serovars *S.* Pullorum and *S.* Gallinarum showed that the 31 fixed intragenic and non-homoplastic fixed variants (Table 1) impacted preferentially genes involved in oxidoreductase activity (Table 2). Interestingly, the unique non-homoplastic InDel fixed in all the genomes of *S.* Pullorum and *S.* Gallinarum (Table 1 and Additional file 6) disrupts the gene encoding FAD dependent oxidoreductase (POS 3067326, WP_000271927). This FAD dependent oxidoreductase also called glycolate oxidase (GOX) is an oxidoreductase that oxidizes α-hydroxy acids to α-keto acids with reduction of oxygen to H_2O_2. Its role remains elusive in eubacteria but its inactivation in the avian-restricted strains suggests that its activity is not compatible with an efficient colonization of the avian gut.

Emphasizing a biphyletic divergence, the 315 and 114 intragenic and non-homoplastic fixed variants differentiating

Table 2 Gene-ontology (GO) terms of intragenic and non-homoplastic variants (SNPs and InDels) fixed in *Salmonella enterica* subsp. *enterica* serovars Dublin *versus* all the others genomes (Ontology 1 called 'Dub_All'), Pullorum/Gallinarum *versus* Enteritidis (Ontology 2 called 'Ent_Pull/Gall'), Pullorum *versus* Gallinarum (Ontology 3 called 'Pull_Gall'), and Gallinarum *versus* Pullorum (Ontology 4 called 'Gall_Pull')

Gene-ontology enrichment analysis	GO ID	GO-term	Number of hits	Expected number of hits	GO-level	*p*-value	Corrected *p*-value	Ontology
1	GO:0006105	succinate metabolic process	36	14.692	8	5.8×10^{-13}	8.1×10^{-10}	BP
	GO:0006307	DNA dealkylation involved in DNA repair	4	1.433	9	1.0×10^{-42}	1.0×10^{-39}	BP
	GO:0006468	protein phosphorylation	61	40.850	8	3.9×10^{-05}	5.5×10^{-02}	BP
	GO:0006520	cellular amino acid metabolic process	533	445.766	7	6.4×10^{-08}	9.0×10^{-05}	BP
	GO:0006525	arginine metabolic process	104	53.392	10	8.4×10^{-18}	1.1×10^{-14}	BP
	GO:0006527	arginine catabolic process	62	25.800	11	1.3×10^{-19}	1.9×10^{-16}	BP
	GO:0006545	glycine biosynthetic process	5	1.792	11	1.0×10^{-42}	1.0×10^{-39}	BP
	GO:0006560	proline metabolic process	81	46.583	10	2.4×10^{-10}	3.4×10^{-07}	BP
	GO:0006562	proline catabolic process	27	12.183	11	3.4×10^{-08}	4.9×10^{-05}	BP
	GO:0009064	glutamine family amino acid metabolic process	190	116.817	9	4.6×10^{-17}	6.5×10^{-14}	BP
	GO:0009065	glutamine family amino acid catabolic process	89	37.983	10	2.4×10^{-25}	3.4×10^{-22}	BP
	GO:0009233	menaquinone metabolic process	27	13.258	6	9.0×10^{-07}	1.2×10^{-03}	BP
	GO:0009234	menaquinone biosynthetic process	27	13.258	7	9.0×10^{-07}	1.2×10^{-03}	BP
	GO:0010133	proline catabolic process to glutamate	27	12.183	11	3.4×10^{-08}	4.9×10^{-05}	BP
	GO:0019544	arginine catabolic process to glutamate	10	3.942	12	1.2×10^{-05}	1.7×10^{-02}	BP
	GO:0019545	arginine catabolic process to succinate	36	14.692	9	5.8×10^{-13}	8.1×10^{-10}	BP
	GO:0035510	DNA dealkylation	4	1.433	8	1.0×10^{-42}	1.0×10^{-39}	BP
	GO:1,901,565	organonitrogen compound catabolic process	162	112.517	5	3.8×10^{-09}	5.3×10^{-06}	BP
	GO:1,901,605	alpha-amino acid metabolic process	320	240.441	8	8.0×10^{-11}	1.1×10^{-07}	BP
	GO:1,901,606	alpha-amino acid catabolic process	100	62.350	9	1.9×10^{-09}	2.7×10^{-06}	BP
	GO:0009379	Holliday junction helicase complex	4	1.408	5	1.0×10^{-42}	1.0×10^{-39}	CC
	GO:0003842	1-pyrroline-5-carboxylate dehydrogenase activity	27	12.957	6	1.5×10^{-07}	1.8×10^{-04}	MF
	GO:0003908	methylated-DNA-[protein]-cysteine S-methyltransferase activity	4	1.524	7	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0004020	adenylylsulfate kinase activity	4	1.524	6	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0004072	aspartate kinase activity	6	2.286	6	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0004372	glycine hydroxymethyltransferase activity	5	1.905	6	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0004412	homoserine dehydrogenase activity	6	2.286	6	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0004657	proline dehydrogenase activity	27	12.957	5	1.5×10^{-07}	1.8×10^{-04}	MF
	GO:0004743	pyruvate kinase activity	10	3.811	6	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0004815	aspartate-tRNA ligase activity	5	1.905	7	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0008770	[acyl-carrier-protein] phosphodiesterase activity	4	1.524	7	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0009015	N-succinylarginine dihydrolase activity	12	4.954	6	3.5×10^{-06}	4.1×10^{-03}	MF
	GO:0009017	succinylglutamate desuccinylase activity	10	4.192	6	2.4×10^{-05}	2.8×10^{-02}	MF
	GO:0015166	polyol transmembrane transporter activity	12	4.573	6	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0015169	glycerol-3-phosphate transmembrane transporter activity	12	4.573	8	1.0×10^{-42}	1.0×10^{-39}	MF

Table 2 Gene-ontology (GO) terms of intragenic and non-homoplastic variants (SNPs and InDels) fixed in *Salmonella enterica* subsp. *enterica* serovars Dublin *versus* all the others genomes (Ontology 1 called 'Dub_All'), Pullorum/Gallinarum *versus* Enteritidis (Ontology 2 called 'Ent_Pull/Gall'), Pullorum *versus* Gallinarum (Ontology 3 called 'Pull_Gall'), and Gallinarum *versus* Pullorum (Ontology 4 called 'Gall_Pull') *(Continued)*

Gene-ontology enrichment analysis	GO ID	GO-term	Number of hits	Expected number of hits	GO-level	*p*-value	Corrected *p*-value	Ontology
	GO:0015430	glycerol-3-phosphate-transporting ATPase activity	6	2.286	9	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0015605	organophosphate ester transmembrane transporter activity	12	4.573	5	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0016749	N-succinyltransferase activity	5	1.905	7	1.0×10^{-42}	1.0×10^{-39}	MF
	GO:0018480	5-carboxymethyl-2-hydroxymuconic-semialdehyde dehydrogenase activity	8	3.049	6	1.0×10^{-42}	1.0×10^{-39}	MF
2	GO:0003973	(S)-2-hydroxy-acid oxidase activity	4	0.013	6	1.1×10^{-12}	1.3×10^{-09}	MF
	GO:0016899	oxidoreductase activity, acting on the CH-OH group of donors, oxygen as acceptor	4	0.013	5	1.1×10^{-12}	1.3×10^{-09}	MF
3	GO:0043603	cellular amide metabolic process	36	17.451	5	1.9×10^{-05}	2.7×10^{-02}	BP
	GO:0006428	isoleucyl-tRNA aminoacylation	4	0.445	11	4.7×10^{-05}	6.7×10^{-02}	BP
	GO:0006522	alanine metabolic process	4	0.376	10	1.8×10^{-05}	2.5×10^{-02}	BP
	GO:0009078	pyruvate family amino acid metabolic process	4	0.376	9	1.8×10^{-05}	2.5×10^{-02}	BP
	GO:0004822	isoleucine-tRNA ligase activity	4	0.475	7	6.5×10^{-05}	7.5×10^{-02}	MF
	GO:0015079	potassium ion transmembrane transporter activity	8	1.680	9	3.7×10^{-05}	4.2×10^{-02}	MF
	GO:0008079	translation termination factor activity	5	0.657	7	3.0×10^{-05}	3.4×10^{-02}	MF
	GO:0003747	translation release factor activity	5	0.657	8	3.0×10^{-05}	3.4×10^{-02}	MF
4	GO:0043603	cellular amide metabolic process	36	17.473	5	2.0×10^{-05}	2.8×10^{-02}	BP
	GO:0006428	isoleucyl-tRNA aminoacylation	4	0.445	11	4.8×10^{-05}	6.7×10^{-02}	BP
	GO:0006522	alanine metabolic process	4	0.377	10	1.8×10^{-05}	2.5×10^{-02}	BP
	GO:0009078	pyruvate family amino acid metabolic process	4	0.377	9	1.8×10^{-05}	2.5×10^{-02}	BP
	GO:0004822	isoleucine-tRNA ligase activity	4	0.475	7	6.5×10^{-05}	7.5×10^{-02}	MF
	GO:0015079	potassium ion transmembrane transporter activity	8	1.681	9	3.7×10^{-05}	4.2×10^{-02}	MF
	GO:0008079	translation termination factor activity	5	0.658	7	3.0×10^{-05}	3.4×10^{-02}	MF
	GO:0003747	translation release factor activity	5	0.658	8	3.0×10^{-05}	3.4×10^{-02}	MF

The identification of variants, detection of fixed variants, assignment of GO-terms to variants, and gene-ontology enrichment analysis were performed with the scripts 'VARCall', 'phyloFixedVar', 'GetGOxML', and 'EveryGO', respectively. The level, biological process (BP), molecular function (MF), and cellular component (CC) of GO-terms are represented. The p-values of hypergeometric tests were adjusted by Bonferroni correction. The lowest corrected p-values representing GO-terms highly impacted by fixed variants (i.e. $< 5.0\times10^{-2}$), the highest GO-levels presenting the most accurate GO-terms (i.e. ≥ 5), and the highest number of hits representing relevant GO-terms quantitatively (i.e. ≥ 4) are presented

S. Pullorum from *S.* Gallinarum (Table 1) impacted GO-terms related to various metabolic processes. However, the results indicate clear trends to accumulate modification in regions related to translation (i.e. 8/16 GO-terms), alanine and pyruvate metabolism, as well as potassium transport (Tables 2 and 3).

Amino acid pathways related to divergence between S. Enteritidis *and* S. Dublin

The present analysis revealed that an important number of fixed variants differentiating *S.* Dublin from *S.* Enteritidis *sensu lato* have been accumulated into genes related to amino acid metabolism (Table 2). A more detailed analysis was then undertaken with a list of 22 genes directly involved in amino acid metabolism and in which 123 SNPs specific to *S.* Dublin have been fixed during the evolution process (Fig. 5). These 123 SNPs fall into the synonymous (81%) or missense (19%) categories. It is significant that none of these 123 mutations led to frameshift or interruption of reading frame suggesting that the corresponding gene functions were conserved. However, the biological consequences of these

Table 3 Impacts of translation and function of proteins encoded by genes presenting GO-terms highly impacted by intragenic and non-homoplastic fixed variants in *Salmonella enterica* subsp. *enterica* serovars Pullorum and Gallinarum

GO	Position	Genes	Reference genotype	Genotype in Gallinarum	Genotype in Pullorum	Protein translation impact in Gallinarum	Protein translation impact in Pullorum	Impact on protein function
GO:0006428	54,044	SEN_RS00235	G	G	T	Null	Missense variant	Modification in Pullorum
GO:0006428	54,289	SEN_RS00235	T	T	C	Null	Synonymous variant	Potential modification
GO:0006428	54,658	SEN_RS00235	C	C	T	Null	Synonymous variant	Potential modification
GO:0006428	55,063	SEN_RS00235	G	G	A	Null	Synonymous variant	Potential modification
GO:0006522	1,313,705	SEN_RS06395	C	C	T	Null	Stop gained	Partial lost in Pullorum
GO:0006522	1,313,706	SEN_RS06395	A	A	G	Null	Missense variant	Modification in Pullorum
GO:0004822	54,044	SEN_RS00235	G	G	T	Null	Missense variant	Modification in Pullorum
GO:0004822	54,289	SEN_RS00235	T	T	C	Null	Synonymous variant	Potential modification
GO:0004822	54,658	SEN_RS00235	C	C	T	Null	Synonymous variant	Potential modification
GO:0004822	55,063	SEN_RS00235	G	G	A	Null	Synonymous variant	Potential modification
GO:0015079	99,941	SEN_RS00445	C	CGCTGGG	C	Disruptive inframe insertion	NULL	Partial lost in Gallinarum
GO:0015079	3,489,575	SEN_RS17095	C	CG	C	Frameshift variant	NULL	Modification in Gallinarum
GO:0003747	1,343,057	SEN_RS06530	C	C	A	Null	Synonymous variant	Potential modification
GO:0003747	1,343,237	SEN_RS06530	A	A	G	Null	Synonymous1 variant	Potential modification
GO:0003747	339,779	SEN_RS01530	G	G	T	Null	Missense variant	Modification in Pullorum

The identification of variants, detection of fixed variants, assignment of GO-terms to variants, and gene-ontology enrichment analysis were performed with the scripts 'VARCall', 'phyloFixedVar', 'GetGOxML', and 'EveryGO', respectively. The variant annotation was performed with SnpEff against reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The p-values of hypergeometric tests were adjusted by Bonferroni correction. The lowest corrected p-values representing GO-terms highly impacted by fixed variants (i.e. $< 5.0\times10^{-2}$), the highest GO-levels presenting the most accurate GO-terms (i.e. ≥ 5), and the highest number of hits representing relevant GO-terms quantitatively (i.e. ≥ 4) are presented

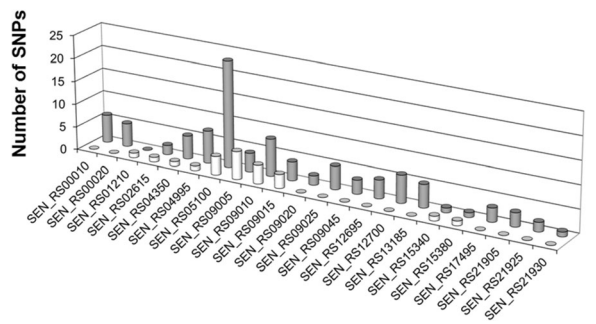

Fig. 5 Genes impacted by single nucleotide polymorphisms (SNPs), involved in the amino acid catabolism, and fixed at the branch representing divergence between *Salmonella* serovars Dublin and Enteritidis/Pullorum/Gallinarum. Round bars represent missense (white) and synonymous SNPs (grew)

fixed SNPs are unclear at this stage. All the concerned genes encode enzymes which are involved in a series of reactions that we extracted using metabolic databases [47]. The Fig. 6 reports one of the most significant reconstructed metabolic network encompassing 12 genes among the 22 and shows that glutamate was a highly connected node.

Salmonella pathogenic islands

Contrary to Langridge et al. who emphasized that evolution of *Salmonella* pathogenicity is strongly associated with the acquisition SPI-6 and SPI-19 [22], the present study identified intragenic and non-homoplastic fixed variants retained for the four presented gene-ontology enrichment analyses in SPI-1, SPI-2, SPI-4, SPI-5, SPI-6 and PAI III 536 (Additional file 8), as well as recombination evens

impacting SPI-1, SPI-2, SPI-5 and SPI-12 (Additional file 9). While intragenic and non-homoplastic fixed variants impacting SPIs define common evolutionary traits of *S.* Dublin, *S.* Pullorum/Gallinarum, *S.* Pullorum and *S.* Gallinarum, we did not observed that SPIs impacted by recombination events are associated with these serovars (Fig. 3).

Discussion

Implementation of a workflow for gene-ontology enrichment analysis based on bacterial coregenome variants

Our objective to retrieve functional information from evolutionary relationships between genomes required first to build a tool to reconstruct robust phylogenetic inferences. The 'VARCall' workflow allows accurate qualitative and quantitative detection of coregenome SNPs and InDels to compute robust downstream phylogenetic inference. Because it is not possible to discriminate between absences of sequencing data and absences of sequences in samples, we did not take into account the variants when reads of at least one genome were missing in the alignments [6–9].

It must be emphasized that the number of coregenome variants is in accordance with the genetic distances between the genomes included in the variant calling analysis. Consequently, we recommend estimating genome pairwise distances before to run the 'VARCall' workflow in order to detect and exclude the potential divergent genomes which may cause a drastic fall in the coregenome variants during variant calling analysis. With this objective and independently of genome sizes, a pangenomic approach, combining *de novo* assemblies (SPAdes [48]) and estimations of Jaccard indexes with a form of locality-sensitive hashing of kmers (MinHash [49]), is currently under development in our team.

Although the removal of variants from recombination events must be theoretically performed when the phylogenetic inference assumes (i.e. Least squares, Minimum Evolution, Neighbor-Joining, UPGMA) or requires (i.e. Maximum Likelihood) a Markov chain model of nucleotide substitution [3], we observed according to Hedge and Wilson [50] that this removal induced a loss of information (Additional file 3), especially in depth branches where fixed variants from recombination events are a majority (Additional file 4).

Because of our objective to detect clade specific traits related to host adaptation, we decided to exclude the homoplastic variants in the analyses of the *S.* Dublin, *S.* Enteritidis, *S.* Gallinarum and *S.* Pullorum genomes. Considering all leaves of the phylogenetic inference, the non-homoplastic fixed variants represented 65% of all fixed variants. It should however be mentioned that it is fully possible to use the scripts 'phyloFixedVar', 'FixedVar', 'GetGOxML' and 'EveryGO' to select homoplastic variants for studies that would focus on coevolution [51].

The gene-ontology enrichment analysis was applied to synonymous and non-synonymous fixed variants because synonymous variants may be involved in regulation of gene expression or level of protein synthesis even if they do not impact the protein sequence [52].

In order to browse easily between tree branches, genes and their annotations, the xml file produced by the scripts 'phyloFixedVar' or 'FixedVar' centralizes the annotations of variants including the homoplastic variants for all the combinations of genomes present in the phylogenetic inference. The polyvalent scripts 'phyloFixedVar', 'FixedVar', 'GetGOxML' and 'EveryGO' were developed to perform the first gene-ontology enrichment analysis based on bacterial coregenome variants and can be applied to all kind of bacterial genome collection.

Coregenome diversity is independent of host specialization in *Salmonella*

Contrary to studies showing that gene losses and accumulation of pseudogenes accompanied host specialization in bacteria [18], we did not observed that mammalian- and avian-adapted *Salmonella* displayed a decreasing in coregenome diversity (Fig. 1). However, and considering the high level of diversity of the *S.* Enteritidis coregenome, we may hypothesize that the MRCA of *S.* Enteritidis and *S.* Dublin was a multi-hosts adapted serovar because several studies concluded that the high genetic diversity of a multi-hosts adapted bacterial lineage represents a source for potential host specializations of less genetically diversified sub-lineages [53]. Because the reduction of the genome size mainly concerns endosymbiotic bacteria [54], we may also hypothesis that the divergence of the *Salmonella* sub-lineages is too recent to make any drastic decreasing of genome diversity related to host specializations observable.

Transition from a free-living state to mammalian intestinal environment

With regards to fixed variants (Table 1), we focused on phylogenetic relevant variants defining the clades *S.* Dublin, *S.* Enteritidis, *S.* Pullorum and *S.* Gallinarum (Additional file 6). The fixed non-homoplastic InDel observed in the gene coding the fimbrial outer membrane usher protein SthC of the multi-hosts adapted serovar *S.* Enteritidis (i.e. fixed genotype ATT) may correspond to frameshift insertions of T or TT in *S.* Enteritidis, or frameshift deletions of T or TT in *S.* Dublin (Additional file 6). Both possibilities may be linked to adaptation to environment and gastrointestinal tracts of mammalian, respectively. The adaptation to free-living state is indeed known to be driven by adhesion on plants or biofilm formation which can be mediated by fimbrial proteins [55]. On the other hand the adaptation to the intestinal tracts of mammalian is known to be mediated

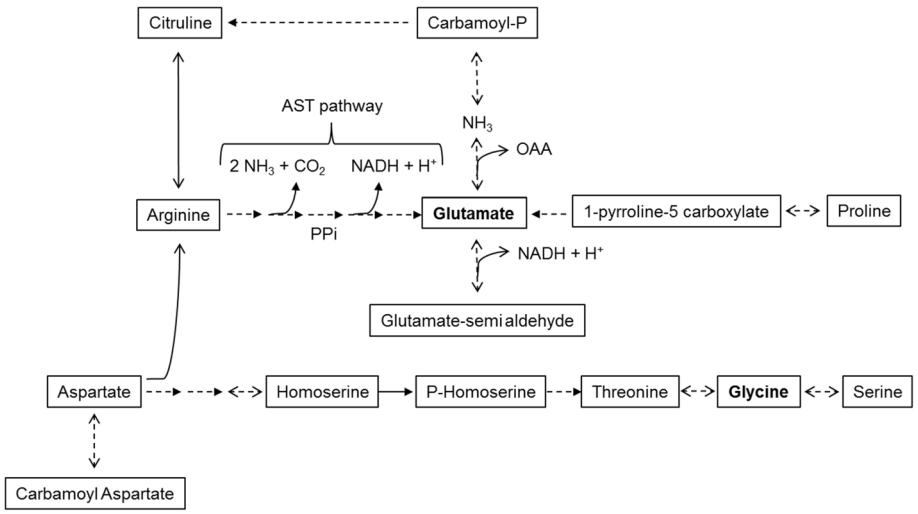

Fig. 6 Amino acid pathways in which intragenic and non-homoplastic fixed single nucleotide polymorphisms (SNPs) differentiating *Salmonella* serovars Dublin *versus* Enteritidis have been detected. The dotted lines represent enzymatic steps for which the corresponding genes encoding enzymes have been specifically mutated. AST, NADH, OAA and PPi stand for ammonia-producing arginine succinyltransferase, nicotinamide adenine dinucleotide, oxaloacetic acid and pyrophosphate, respectively. KEGG database were used as a database for reference pathway (Nucl. Acids Res. 2016;44:D457–62)

by fimbriae operon during the transition from a free-living state to the intestinal environment, especially the *sth*ABCDE fimbrial operon which is frequent in non-typhoidal *Salmonella* serovars [56, 57] and required by *S.* Typhimurium for intestinal persistence in mice [58].

The divergence between Dublin and Enteritidis targets regions involved in metabolic pathways of amino acids

The results of the present study highlighted the accumulation of fixed variants in intragenic regions connected to metabolic pathways of amino acids (Table 2). The most impacted pathway is the arginyl succinyl transferase pathway (AST pathway) involved in arginine catabolism. All 5 genes encoding the enzymes catalyzing the degradation of arginine to glutamate with the concomitant release of NH_3, CO_2 and regeneration of NADH + H^+ accumulated mutations specific to each of the two serovars (Fig. 6). Concerning the GO-terms of interest related to metabolic and catabolism processes of amino acids in *S.* Dublin (Table 2 and Fig. 6), the glycine (GO:0006545, GO:0004372) may form serine or threonine but is not present in sufficient amounts to support growth of *S.* Typhimurium during intra host survival, and the glycine production or conversion of glycine to tetrahydrofuran (oxolane) are essential reactions during mice infection by *S.* Typhimurim [59]. Aerobic replication of *S.* Typhimurium in mice requires also the twin-arginine translocation system (Tat) which exports across the cytoplasmic membrane numerous cofactors containing virulence factors and anaerobic respiratory chain proteins [45] (GO:0006525, GO:0006527, GO:0019545, GO:0016749, GO:0009015). More precisely, with an interconnection between the biosynthesis of arginine and polyamines (i.e. putrescine and spermidine), the carbamoylphosphate is a precursor of arginine and is produced by the carbamoylphosphate synthetase (CPSase) with glutamine as the physiological amino group donor [60] (GO:0009064, GO:0009065, GO:0010133, GO:0019544, GO:0009017). The N-succinylarginine dihydrolase (gene *astB*) (GO:0009015) is indeed the second enzyme of the arginine succinyltransferase pathway involved in arginine catabolism as a sole nitrogen source [61].

Among its role as a protein component and as source of CO_2 and NH_3 through the GS/GOGA cycle, glutamate plays a major role in adaptation of bacteria to hyper osmotic conditions which are notably faced in the lumen of the digestive tract [62]. Our observations suggest a possible impact of the set of fixed SNPs on a differential glutamate accumulation capacity between *S.* Enteritidis and *S.* Dublin cells (Fig. 6). This hypothesis could be tested either by measuring cytoplasmic glutamate accumulation or growth under osmotic stress.

With regard to the GO-terms of interest for *S.* Dublin (Table 2 and Fig. 6), the accumulation of glutamate is correlated with *Salmonella* growth at high external osmolality [63] and aspartate is known to induce significant dysbacteriosis in gut microbiota of animals and humans [64]. Also included in GO-terms of interest for *S.* Dublin, the homoserine dehydrogenase (GO:0004412) is in fact a key enzyme in the biosynthetic pathway from aspartate to homoserine, which is a common precursor for the synthesis of amino acids methionine, threonine and isoleucine.

Regarding the proline metabolic and catabolic processes impacted by non-homoplastic fixed variants in *S.* Dublin (GO:0006560, GO:0006562, GO:0010133, GO:0004657, GO:0003842), hyperosmotic stress and proline limitation in host compartment is indeed known to lead *S.* Typhimurium responds to a decrease in the levels of proline-charged $tRNA^{Pro}$ by promoting expression of the mgtCBR virulence operon [65]. Related to this observation, the 1-pyrroline-5-carboxylate dehydrogenase identified as GO-terms of interest for *S.* Dublin (GO:0003842) is encoded in the *putA* gene of *S.* Typhimurim by a bifunctional membrane-associated dehydrogenase which oxidizes proline to glutamate for use as the sole carbon, nitrogen or energy source [66].

The genes *alr* of *S.* Typhimurim, also called *dal* genes, encode alanine racemases which are biosynthesis sources of D-alanine for cell wall formation and also necessary to the catabolism of L-alanine as a source of carbon, energy and nitrogen [67]. L-arginine being used for growth of laying hens [68], the stop gained and missense variants in alanine racemase leading modifications of function in *S.* Pullorum (SEN_RS06395, POS 1313705 and POS 1313706, GO:0006522) may consequently be due to his adaptation to this avian alimentation.

Compared to similar sized nonflying mammalian, the avian gastrointestinal tracks present a typical shorter retention time and quantitatively much more important passive adsorption of L-glucose. During ontogeny, the architecture and functioning of mammalian and avian gastrointestinal tracts are closely related with their food diet and intake rate. The gestational phase of mammalian is dominated by production of gastrointestinal enzymes required for digestion and absorption of milk (e.g. amino acid transporters and the Na^+-dependent D-glucose transporter SLGT1), whereas enzymes required for solid food and pinocytotic uptake capacity are produced during weaning (e.g. fructose and starch), and activities of sucrase-isomaltase, maltase-glucoamylase, trehalase and fructose transport (i.e. GLUT-5) increase when adult diet switches from lactose to sucrose and starch. In contrast, the pre- and post-natal periods of birds and chicks are associated with a switch of gastrointestinal functions driven by the transition from a lipid-rich yolk diet inside the egg (i.e. sucrase-isomaltase and the D-glucose transporter SLGT1) to a carbohydrate- and protein-based diet after hatch in young chickens and house sparrows (i.e. sucrose, maltase, maltase-glucoamylase and pancreatic amylase activities) [69]. Several specific amino acid transporters and a single proton-oligopeptide transporter (PEPT1) are responsible of the assimilation of protein components by the enterocytes of mammalian small intestines, but the specific amino acid transporters and role in protein nutrition of the ancient PEPT1 are still poorly described in avian [70].

Acid survival in mammalian implies modifications of amino acids pathways

Further with respect to the GO-terms of interest related to metabolic and catabolism processes of amino acids in *S.* Dublin (Table 2 and Fig. 6), *S.* Typhimurium has also an active arginine-dependent arginine mechanism permitting survival at pH 2.5 [71]. *Salmonella* uses the tolerance response of low gastric pH and the arginine decarboxylase (gene *adiA*) acid resistance system to prepare for the stresses of host-cell interactions [72].

Antibiotic resistances in avian implies fitness restorations

Mutations related to resistance acquisition and fitness restoration in the gene *ileS* encoding the essential enzyme isoleucine-tRNA ligase (IleRS) were observed in *S.* Enteritidis under selection pressure induced by the mupirocin, an inhibitor of this enzymatic activity [73]. Even if the mupirocin is not a common antibiotic for poultry farming, the missense variant in isoleucine-tRNA ligase leading modifications of function in *S.* Pullorum (SEN_RS00235, POS 54044, GO:0006428 + GO:0004822) may consequently also be linked to other antibiotics (Table 3).

Limited ion supply in avian tract implies modifications of ion transport

Concerning the loss or modification of potassium ion transmembrane transporter activity in the avian-adapted serovar *S.* Gallinarum (GO:0015079) (Table 3), the modifications of ion biosynthesis is indeed also known to contribute to avian adaptation of *S.* Kentucky (e.g. turkey farms, Hatchery chicks, commercial broiler, layer flocks, commercial broiler environments, broiler processing plants, retail poultry products). For instance, plasmids encoding aerobactin (i.e. *iucABCD* and *iutA*) and Sit iron transport operons (i.e. *sitABCD*), as well as other iron acquisition genes (e.g. *iss*) play indeed a major role in survival abilities of *S.* Kentucky (e.g. IncFIB plasmid) and some *S.* Heidelberg (e.g. pSH163_120 and pSH696_117 plasmids) in the extraintestinal environments of poultry where iron is in limited supply [74].

Conclusions

In conclusion, we proposed the first validated procedure to identify fixed SNPs and InDels according to inferred phylogenetic clades and performed the associated gene-ontology enrichment analysis in order to describe the adaptation of *Salmonella* serovars Dublin (i.e. mammalian-hosts), Enteritidis (i.e. multi-hosts), Pullorum (i.e. avian-hosts) and Gallinarum (i.e. avian-hosts) at the coregenome scale. Among the multiple metabolic pathways impacted by fixed variants during host adaptation of *Salmonella* serovars, our main observation emphasized that glutamate

metabolism could play a major role in adaptation of *S.* Dublin to mammalian-hosts.

Methods

Read dataset

With the objective to validate the method of phylogenetic inference described in detail below, and to perform the first gene-ontology enrichment analysis based on bacterial core-genome variants, a previously published read dataset was used [22]. This collection of reads is made of 59 genomes sequences of *Salmonella* strains and was constituted by Langridge et al. [22] in order to describe host adaptation at the accessory gene scale of *Salmonella* serovars *S.* Enteritidis (i.e. multi-hosts), *S.* Dublin (i.e. mammalian-hosts), *S.* Gallinarum (i.e. avian hosts) and *S.* Pullorum (i.e. avian-hosts) (Additional file 1). While Langridge et al. proposed manual workflows mainly focusing on the accessory genome, we provide in the present manuscript automated workflows aiming to perform the first gene-ontology enrichment analysis based on bacterial coregenome variants.

Variant calling analysis (i.e. the 'VARCall' workflow)

A driver script called 'VARCall' invokes 'BAMmaker', 'VCFmaker_SNP', 'VCFmaker_INDEL', and 'SNP-INDEL_merge', successively (Fig. 7). The script 'BAMmaker' allows trimming of single- and paired-end reads with Trimmomatic (i.e. quality score of 25 and minimal length of 20 bp) [75], read alignment against a reference genome with BWA [76], read sorting with Samtools [77], as well as duplication removal and realignment around InDels with the Genome Analysis Toolkit (GATK) [2], successively. Following an approved framework for variant discovery [78], the scripts 'VCFmaker_SNP' and 'VCFmaker_INDEL' call and filter variants (i.e. SNPs and InDels) according to GATK best practices [2] in order to retain high-confidence variants [33]. After variant combination prioritizing SNPs with GATK [2] and SNPs/InDels flagging with SnpSift [79], the variants presenting allele frequencies equal to one across the dataset of genomes (i.e. variants specific to the reference genome), as well as the variants presenting missing genotypes in at least one genome, are removed from the dataset of variants (i.e. vcf file). The 'VARCall' workflow produces four output files: matrices of pairwise distances ('VCFtoMATRIX), a report about breadth and depth coverages (i.e. script 'reportMaker'), and files of concatenated variants (i.e. script 'VCFtofasta') and pseudogenomes (i.e. 'VCFtoseudogenomes') to build fast or slow phylogenetic inferences, respectively (Fig. 7). The pseudogenomes correspond to the reference genome where the genotypes of detected variants are replaced in each genome of the dataset. Finally, the variants are annotated with SnpEff (version 4.1 g without variants from intron, UTR-5', UTR-3', upstream regions, and downstream regions) [80] based on the reference genome annotation from NCBI (i.e. gbk file). Independently of the 'VARCall' workflow, the density of SNPs were computed with the 'vcf-subset' module and the 'SNPdensity' argument of VCFtools [81].

We develop currently a new version of the 'VARCall' workflow (i.e. iVARCall2) based on the HaplotypeCaller algorism (GATK) in order to call SNPs and InDels at the same time by local *de novo* assembly for each genome which would give us the opportunity to compute a faster phylogenetic inference based on variants independently called for each genome (HaplotypeCaller) rather than to perform time consuming variant calling across reference genome alignments (UnifiedGenotyper) [82].

Phylogenetic inference

Based on the pseudogenomes, the phylogenetic inference was performed with the multi-core architecture of RAxML [40]. Following computation of a parsimony starting tree, a rapid bootstrap analysis and search was computed for best-scoring Maximum Likelihood (ML) tree with General Time-Reversible (GTR) model of substitution [83] and the secondary structure 16-state model (i.e. nwk file). A posterior bootstrap convergence test was performed in order to determine if sufficient bootstrap replicates were computed [39], and the log likelihood score of all trees was also computed with RAxML [40]. Based on the pseudogenomes and the RAxML phylogenetic inference, the recombination events were detected using ClonalFrameML locating regions with high densities of SNPs on each branch [5]. The positions of recombination events detected by ClonalFrameML were also removed from the vcf file with a script 'Clonal_VCFilter' in order to compute phylogenetic inference based on pseudogenomes excluding variants linked to recombination events. Finally, the best-scoring ML trees were graphically represented with iTOL viewer [84]. The comparison of the tree topologies was performed using the cophylo function of 'phytools' R package [85]. In order to support results of phylogenetic inference, genetic structure analyses were also performed with BAPS based on SNPs [41].

Fixed and homoplastic variants

A script called 'phyloFixedVar' uses the annotated vcf file from the variant calling analysis and the nwk file from a binary tree. This script was developed in order to detect sensitive and specific variants at each branch of the phylogenetic inference. The variants are defined with a biological point of view either as fixed (i.e. variants with sensitive and specific genotypes at intermediate branches) or transient (i.e. variants with sensitive and specific genotypes at final leaves). The script

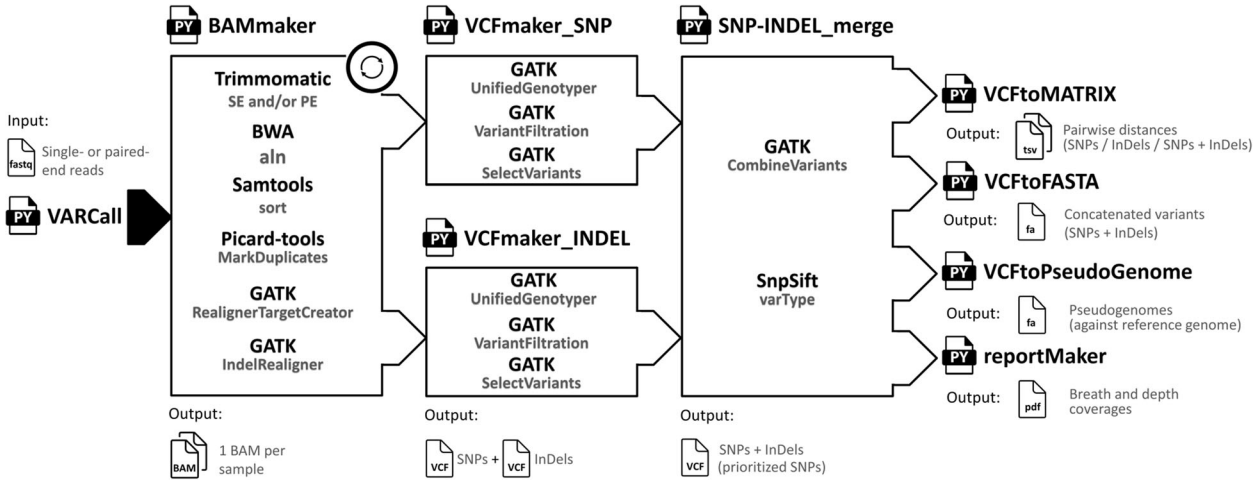

Fig. 7 Programs (i.e. black letters) and commands (i.e. grew letters) implemented in the 'VARCall' workflow aiming to call single nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels). The scripts referring to alignment against reference genome (i.e. 'BAMmaker'), variant calling (i.e. 'VCFmaker_SNP' and 'VCFmaker_INDEL'), variant combination (i.e. 'SNP-INDEL_merge'), pairwise distances (i.e. 'VCFtoMATRIX'), variant concatenation (i.e. 'VCFtoFASTA'), pseudogenome assemblies (i.e. 'VCFtoPseudoGenome'), and report about breadth and depth coverages (i.e. 'reportMaker') were written with Python 2.7 and are invoked by the driven script 'VARCall' (i.e. black arrow). The script 'BAMmaker' is performed for each genome (i.e. circular arrow)

'phyloFixedVar' (a) identifies all comparisons of genome groups referring to all nodes, (b) selects sensitive and specific variants, and (c) defines if these variants are homoplastic (i.e. convergence of genotypes between independent phylogenetic clades), successively. More precisely, right child leaves are listed with corresponding leaves of left child at each node and reversely, then both comparisons (i.e. right versus left children and reversely) are connected with comparison numbers which are associated to identifiers of single nodes, all together grouped into node labels in a new nwk file (a). On the one hand the sensitive variants are detected as presenting common genotypes into leaves of right or left child of previously listed comparisons, and on the other hand the specific variants are identified as presenting different genotypes in corresponding right or left children (b). With regard to these sensitive and specific variants, the genotypes of all the other genomes are screened in order to tag homoplastic variants (i.e. common genotypes between variants of independent phylogenetic clades) (c). The selected variants are finally written in the xml file with their related annotations (i.e. genotype, effect of homoplasy, NCBI identifiers, gene IDs, gene names, type, position, phenotypical impact) for each node and all

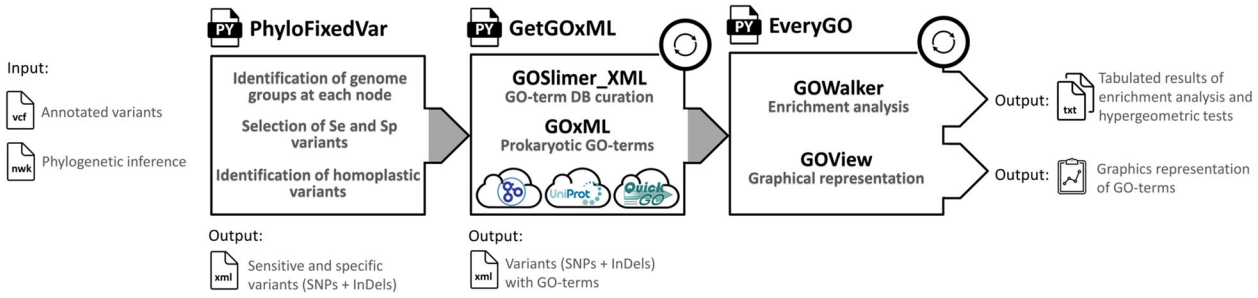

Fig. 8 Programs (i.e. black letters) and corresponding effects (i.e. grew letters) implemented in the scripts 'phyloFixedVar', 'GetGOxML' and 'EveryGO' aiming to identify sensitive (Se) and specific (Sp) variants (SNPs and InDels) at each branches of corresponding phylogenetic inference, associate prokaryotic gene ontology (GO) terms with these homoplastic and/or non-homoplastic variants, and perform gene-ontology enrichment analysis based on the parent-child approach integrating hypergeometric tests and Bonferroni corrections, respectively. Online databases are queried by the scripts 'GOSlimer' and 'GOxML' (i.e. clouds). The GO database of the Gene Ontology Consortium is used by the script 'GOSlimer' to identify prokaryotic GO-terms. The QuickGO browser of the UniProt GO annotation program is queried by the script 'GOxML' to associate the variants with the corresponding GO-terms. These scripts were written with Python 2.7 and implement R libraries 'p.ajust' and 'phyper'. The whole workflow is semi-automated (i.e. grew arrows) and the scripts 'GetGOxML' and 'EveryGO' can be performed for each variant and each branch, respectively (i.e. circular arrow)

comparisons of lists of leaves. Similarly to 'phyloFixedVar', another script called 'FixedVar' requires a vcf file and lists of genomes IDs that have to be compared. This script was developed in order to detect sensitive and specific variants independently of the phylogenetic inference.

Gene-ontology enrichment analysis

With the objective to associate the selected variants to corresponding prokaryotic GO-terms, a driver script called 'GetGOxML' invokes the scripts 'GOSlimer_XML' and 'GOxML', successively (Fig. 8). More precisely, the script 'GOSlimer_XML' aims to generate lists of prokaryote GO-terms based on the GO database of the Gene Ontology Consortium [14] (i.e. go-basic.obo file: http://geneontology.org/page/download-ontology), and the script 'GOxML' associates gene identifiers (i.e. NP or WP) from the xml file with GO-terms available in the QuickGO browser (http://www.ebi.ac.uk/GOA) of the UniProt GO annotation program (https://www.ebi.ac.uk/QuickGO/). In addition, the script 'GOxML' allows comparison of these GO-terms in order to exclude potential eukaryote GO-terms from the dataset and retains prokaryote GO-terms. Finally, the script 'GOxML' integrates the curated GO-terms (i.e. prokaryote GO-terms) and related biological processes to the common xml file in order to centralized the GO-terms and functional annotations of variants (i.e. genotype, effect of homoplasy, NCBI identifiers, gene IDs, gene names, type, position, phenotypical impact). With a view to select intragenic variants (SNPs and InDels), and distinguish between GO-terms from the variants of interest (i.e. tested sample) and all variants (i.e. universe) which are used for the hypergeometric test of the gene-ontology enrichment analysis based on the parent-child approach [15], the driver script 'EveryGO' invokes 'GOWalker.R' and 'GOView.R', successively (Fig. 8). More precisely, the script 'GOWalker.R' counts the GO-terms from the sample (i.e. variants from compared leaves) and universe (i.e. all variants) for each GO-term, as well as the sizes of the sample (i.e. total GO-terms in the sample) and universe (i.e. total GO-terms in the universe). Then, the script 'GOWalker.R" performs the hypergeometric test and Bonferroni correction [86] implemented in the 'stats' and 'phyper' R libraries, respectively [87]. Finally, the script 'GOView.R' aims to compute a graphical representation of the gene-ontology enrichment analysis with the plotting system ggplot2 (i.e. *p*-values of the hypergeometric tests *versus* the branch levels from the GO-terms of the DAG).

Salmonella pathogenic islands

Candidates of PAI-like region overlapping genomic islands (cPAIs) of the Pathogenicity Island Database (http://www.paidb.re.kr) from KonKuk University (Seoul, South Korea) [88] were used to detect SPI-1 (2890501–2,934,879), SPI-2 (1727425–1,769,273), SPI-4 (4333507–4,361,514), SPI-5 (1053174–1,074,167), SPI-6 (299796–330,890), SPI-11 (1904313–1,912,607), SPI-12 (2328077–2,347,757) and PAI III 536 (2801306–2,810,695) according to the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1).

Additional files

Additional file 1: Genome dataset used in the present study. The genomes of *Salmonella enterica* subsp. *enterica* serovars Enteritidis, Pullorum, Gallinarum and Dublin were described by Langridge et al. (Proc. Natl. Acad. Sci. 2015;112:863–8). (XLSX 15 kb)

Additional file 2: Statistical report of the 'VARCall' workflow. The serovars *Salmonella enterica* subsp. *enterica* serovars Dublin, Enteritidis, Pullorum and Gallinarum were described by Langridge et al. (Proc. Natl. Acad. Sci. 2015;112:863–8). (XLSX 16 kb)

Additional file 3: Phylogenetic inferences performed based on coregenome single nucleotide polymorphisms (SNPs) excluding (A) or including (B) variants from recombination events detected in *Salmonella enterica* subsp. *enterica* serovars Dublin, Enteritidis, Pullorum and Gallinarum. The variants were identified by the 'VARCall' workflow against the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The positions of recombination events detected by Maximum Likelihood and default gamma priors of ClonalFrameML are removed with a script 'Clonal_VCFilter' in order to compute phylogenetic inference based on pseudogenomes excluding variants linked to recombination events. The produced pseudogenomes (4,685,848 bp) were inferred with RAxML based on a bootstrap analysis and search for best-scoring Maximum Likelihood tree with General Time-Reversible model of substitution and the secondary structure 16-state model. The color legend corresponds to phylogenetic clustering performed by Langridge et al. (Proc. Natl. Acad. Sci. 2015;112:863–8). The trees are rooted on the branches of *S.* Dublin before comparison. The comparison of the tree topologies were performed using the cophylo function of 'phytools' R package. (PDF 1733 kb)

Additional file 4: Phylogenetic inference based on coregenome single nucleotide polymorphisms (SNPs) and recombination events identified in *Salmonella enterica* subsp. *enterica* serovars Dublin, Enteritidis, Pullorum and Gallinarum. The color legend corresponds to serovars presented by Langridge et al. (Proc. Natl. Acad. Sci. 2015;112:863–8). The variants were identified by the 'VARCall' workflow against the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The produced pseudogenomes (4,685,848 bp) were inferred with RAxML based on a bootstrap analysis and search for best-scoring Maximum Likelihood tree with General Time-Reversible model of substitution and the secondary structure 16-state model. The phylogenetic inference converged after 200 bootstrap replicates with a log likelihood score of -8.10^6 for 1000 computed trees. The tree is rooted on the branch of *S.* Dublin. The pseudogenomes and the RAxML inference were used to perform detection of recombination events based on default gamma priors of ClonalFrameML. The number of recombination events is defined closed to white circles which represent recombination events occurred on a branch of the phylogenetic tree. The recombination events with sizes higher than 400 bp are presented. (PDF 752 kb)

Additional file 5: List of recombination events identified in *Salmonella enterica* subsp. *enterica* serovars Dublin, Enteritidis, Pullorum and Gallinarum. The variants were identified by the 'VARCall' workflow against the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The produced pseudogenomes (4,685,848 bp) were inferred with RAxML based on a bootstrap analysis and search for best-scoring Maximum Likelihood tree with General Time-Reversible model of substitution and the secondary structure 16-state model. The pseudogenomes and the RAxML inference were used to perform detection of recombination events based on default gamma priors of

ClonalFrameML. The recombination events with sizes higher than 400 bp are presented. (XLSX 16 kb)

Additional file 6: Phylogenetically relevant single nucleotide polymorphisms (SNPs) and small insertions/deletions (InDels) fixed at phylogenetic branches where genomes of *Salmonella enterica* subsp. *enterica* serovars Enteritidis ($n = 33$), Pullorum ($n = 5$), Gallinarum ($n = 8$) and Dublin ($n = 13$) diverged. The variant calling analysis was performed with the 'VARCall' workflow (i.e. 12,929 SNPs and 1157 small InDels). The fixed non-homoplastic variants are defined by common genotypes across the considered group of genomes, as well as different genotypes in all the others compared genomes. The fixed homoplastic variants are defined by common genotypes across the considered group of genomes and genomes of independent phylogenetic clades, as well as different genotypes in genomes of the compared child-leaves. The variants were annotated with SnpEff against reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). (XLSX 13 kb)

Additional file 7: Gene-ontology (GO) terms of intragenic and non-homoplastic variants (SNPs and InDels) fixed in *Salmonella enterica* subsp. *enterica* serovars Dublin *versus* all the others genomes (Ontology 1 called 'Dub_All'), Pullorum/Gallinarum *versus* Enteritidis (Ontology 2 called 'Ent_Pull/Gall'), Pullorum *versus* Gallinarum (Ontology 3 called 'Pull_Gall'), and Gallinarum *versus* Pullorum (Ontology 4 called 'Gall_Pull'). The variant annotation was performed with SnpEff against reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The identification of variants, detection of fixed variants, assignment of GO-terms to variants, and gene-ontology enrichment analysis were performed with the scripts 'VARCall', 'phyloFixedVar', 'GetGOxML', and 'EveryGO', respectively. The level, biological process (BP), molecular function (MF), and cellular component (CC) of GO-terms are represented. The *p*-values of hypergeometric tests were adjusted by Bonferroni correction. (XLSX 327 kb)

Additional file 8: Candidates of PAI-like region overlapping genomic islands (cPAIs) impacted by intragenic and non-homoplastic variants (SNPs and InDels) fixed in *Salmonella enterica* subsp. *enterica* serovars Dublin *versus* all the others genomes (Ontology 1 called 'Dub_All'), Pullorum/Gallinarum *versus* Enteritidis (Ontology 2 called 'Ent_Pull/Gall'), Pullorum *versus* Gallinarum (Ontology 3 called 'Pull_Gall'), and Gallinarum *versus* Pullorum (Ontology 4 called 'Gall_Pull'). Pathogenicity Island Database from KonKuk University (Seoul, South Korea) were used to detect *Salmonella* Pathogenic Islands (SPIs) SPI-1 (2890501–2,934,879), SPI-2 (1727425–1,769,273), SPI-4 (4333507–4,361,514), SPI-5 (1053174–1,074,167), SPI-6 (299796–330,890), SPI-11 (1904313–1,912,607), SPI-12 (2328077–2,347,757) and PAI III 536 (2801306–2,810,695) of the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). (XLSX 251 kb)

Additional file 9: Candidates of PAI-like region overlapping genomic islands (cPAIs) impacted recombination events identified in *Salmonella enterica* subsp. *enterica* serovars Dublin, Enteritidis, Pullorum and Gallinarum. The variants were identified by the workflow 'VARCall' against the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). The produced pseudogenomes (4,685,848 bp) were inferred with RAxML based on a bootstrap analysis and search for best-scoring Maximum Likelihood tree with General Time-Reversible model of substitution and the secondary structure 16-state model. The pseudogenomes and the RAxML inference were used to perform detection of recombination events based on default gamma priors of ClonalFrameML. The recombination events with sizes higher than 400 bp are presented. Pathogenicity Island Database from KonKuk University (Seoul, South Korea) were used to detect *Salmonella* Pathogenic Islands (SPIs) SPI-1 (2890501–2,934,879), SPI-2 (1727425–1,769,273), SPI-4 (4333507–4,361,514), SPI-5 (1053174–1,074,167), SPI-6 (299796–330,890), SPI-11 (1904313–1,912,607), SPI-12 (2328077–2,347,757) and PAI III 536 (2801306–2,810,695) of the reference genome *S.* Enteritidis (strain P125109, accession NC_011294.1). (XLSX 19 kb)

Abbreviations

AST: Arginyl succinyl transferase; BP: Biological process; CC: Cellular component; cPAIs: Candidates of PAI-like region overlapping genomic islands; CPSase: Carbamoylphosphate synthetase; CRISPR: Clustered regularly interspaced short palindromic repeats; DAG: Directed acyclic graph; GLUT-5: fructose transport; GO: Gene ontology; GOX: Glycolate oxidase; IleRS: Isoleucine-tRNA ligase; InDels: small insertions/deletions; MF: Molecular function; MRCA: Most recent common ancestor; NADH: Nicotinamide adenine dinucleotide; OAA: Oxaloacetic acid; PAIs: Pathogenicity islands; PEPT1: Proton-oligopeptide transporter; PPi: Pyrophosphate; SLGT1: Na^+-dependent D-glucose transporter; SNPs: Single nucleotide polymorphisms; SPIs: *Salmonella* pathogenicity islands; T6SSs: Type VI secretion systems; Tat: Twin-arginine translocation system

Acknowledgements

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

Funding

This work was supported by a grant from the European Union's Horizon 2020 research and innovation program under grant agreement No. 643476 (COMPARE). The funding body had no role in the design of the study, collection/analysis/interpretation of data, and writing the manuscript.

Availability of data and materials

The scripts developed for computational analyses (i.e. 'VARCall', 'phyloFixedVar', 'FixedVar', 'GetGOxML' and 'EveryGO') can be found in the following repositories https://github.com/afelten-Anses/VARtools/tree/master/VARCall; https://github.com/afelten-Anses/VARtools/tree/master/FixedVarTools and https://github.com/afelten-Anses/VARtools/tree/master/GOtools. Previously published sequencing data [22] is available in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) (Additional file 1). The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Authors' contributions

NR planned the Bioinformatic project and was a major contributor in writing the manuscript. NR set and organized the programs of the 'VARCall' workflow. AF made automatic the 'VARCall' workflow. AF and MVN developed the scripts 'phyloFixedVar' and 'FixedVar'. AF and KD developed the scripts 'GetGOxML' and 'EveryGO'. NR, LG and MYM interpreted the genomic data and drafted the manuscript. 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.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

Received: 17 August 2017 Accepted: 16 November 2017
Published online: 28 November 2017

References

1. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.
2. 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.
3. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 2012;13:303–14.
4. 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. Casadesús J, editor. PLoS Genet 2013;9:e1003471.
5. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. Prlic A, editor. PLoS Comput Biol 2015;11:e1004041.
6. Lees JA, Bentley SD, Bacterial GWAS. Not just gilding the lily. Nat Rev Microbiol. 2016;14:406–6.

7. Andersen JB, Sternberg C, Poulsen LK, Bjorn SP, Givskov M, Molin S. New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. Appl Environ Microbiol. 1998;64:2240–6.
8. Amato SM, Fazen CH, Henry TC, Mok WWK, Orman MA, Sandvik EL, et al. The role of metabolism in bacterial persistence. Front Microbiol. 2014;5:1–9.
9. 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.
10. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37:1–13.
11. Fruzangohar M, Ebrahimie E, Ogunniyi AD, Mahdi LK, Paton JC, Adelson DL. Comparative GO: A Web Application For comparative gene ontology and gene ontology-based gene selection in bacteria. Patterson RL, editor. PLoS One 2013;8:e58759.
12. 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.
13. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol. 2003;4:P3.
14. 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.
15. 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.
16. 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.
17. Bäumler AJ, Tsolis RM, Ficht TA, Adams LG. Evolution of host adaptation in salmonella enterica. Infect Immun. 1998;66:4579–87.
18. Foley SL, Johnson TJ, Ricke SC, Nayak R, Danzeisen J. Salmonella pathogenicity and host adaptation in chicken-associated Serovars. Microbiol Mol Biol Rev. 2013;77:582–607.
19. Porwollik S, Santiviago CA, Cheng P, Florea L, Jackson S, 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.
20. 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.
21. 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.
22. 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.
23. Rabsch W, Andrews HL, Kingsley RA, Prager R, Tschäpe H, Adams LG, et al. Salmonella enterica serotype typhimurium and its host-adapted variants. Infect Immun. 2002;70:2249–55.
24. 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.
25. 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.
26. Soucy SM, Huang J, Gogarten JP. Horizontal gene transfer: building the web of life. Nat Rev Genet. 2015;16:472–82.
27. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, et al. Complete genome sequence of salmonella enterica serovar typhimurium LT2. Nature. 2001;413:852–6.
28. Rotger R, Casadesús J. The virulence plasmids of salmonella. Int Microbiol Off J Span Soc Microbiol. 1999;2:177–84.
29. Libby SJ, Lesnick M, Hasegawa P, Weidenhammer E, Guiney DG. The salmonella virulence plasmid spv genes are required for cytopathology in human monocyte-derived macrophages. Cell Microbiol. 2000;2:49–58.
30. Nielsen LR. Review of pathogenesis and diagnostic methods of immediate relevance for epidemiology and control of salmonella Dublin in cattle. Vet Microbiol. 2013;162:1–9.
31. Blondel CJ, Yang H-J, Castro B, Chiang S, Toro CS, Zaldívar M, et al. Contribution of the type VI secretion system encoded in SPI-19 to chicken colonization by salmonella enterica serotypes Gallinarum and Enteritidis. Otto M, editor. PLoS One 2010;5:e11724.
32. Neuman JA, Isakov O, Shomron N. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. Brief Bioinform. 2013;14:46–55.
33. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline: The Genome Analysis Toolkit Best Practices Pipeline. In: Bateman A, Pearson WR, Stein LD, Stormo GD, Yates JR, editors. Curr. Protoc. Bioinforma. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013. p. 1–33.
34. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011;21:961–73.
35. Fridjonsson O, Olafsson K, Tompsett S, Bjornsdottir S, Consuegra S, Knox D, et al. Detection and mapping of mtDNA SNPs in Atlantic salmon using high throughput DNA sequencing. BMC Genomics. 2011;12:1–10.
36. Beres SB, Carroll RK, Shea PR, Sitkiewicz I, Martinez-Gutierrez JC, Low DE, et al. Molecular complexity of successive bacterial epidemics deconvoluted by comparative pathogenomics. Proc Natl Acad Sci. 2010;107:4371–6.
37. Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, et al. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res. 2017;45:D1075–81.
38. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. Hum Mol Genet. 2010;19:R131–6.
39. Pattengale ND, Alipour M, Bininda-Emonds ORP, Moret BME, Stamatakis A. How many bootstrap replicates are necessary? J Comput Biol. 2010;17:337–54.
40. Aberer AJ, Pattengale ND, Stamatakis A. Parallelized phylogenetic post-analysis on multi-core architectures. J Comput Sci. 2010;1:107–14.
41. Corander J, Marttinen P, Sirén J, Tang J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinformatics. 2008;9:539.
42. Barrow PA, Neto OCF. Pullorum disease and fowl typhoid—new thoughts on old diseases: a review. Avian Pathol. 2011;40:1–13.
43. Crichton PB, Old DC. Salmonellae of serotypes Gallinarum and Pullorum grouped by biotyping and Fimbrial-gene probing. J Med Microbiol. 1990;32:145–52.
44. Joyce EA, Chan K, Salama NR, Falkow S. Redefining bacterial populations: a post-genomic reformation. Nat Rev Genet. 2002;3:462–73.
45. Craig M, Sadik AY, Golubeva YA, Tidhar A, Slauch JM. Twin-arginine translocation system (*tat*) mutants of *Salmonella* are attenuated due to envelope defects, not respiratory defects: role of tat in *Salmonella* virulence. Mol Microbiol. 2013;89:887–902.
46. Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev. 2009;33:376–93.
47. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe MKEGG. As a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
48. 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.
49. 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.
50. Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. MBio. 2014;5:e02158–14.
51. Makendi C, Page AJ, Wren BW, Le Thi Phuong T, Clare S, Hale C, et al. A phylogenetic and phenotypic analysis of salmonella enterica Serovar Weltevreden, an emerging agent of diarrheal disease in tropical regions. Ryan ET. PLoS Negl Trop Dis. 2016;10:e0004446.
52. 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.
53. Radomski N, Thibault VC, Karoui C, de Cruz K, Cochard T, Gutiérrez C, et al. Genotypic diversity of Mycobacterium Avium subspecies from human and animal origins, studied by MIRU-VNTR and IS1311 RFLP typing methods. J Clin Microbiol. 2010;48:1026–34.
54. Wernegreen JJ. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet. 2002;3:850–61.
55. Wiedemann A, Virlogeux-Payant I, Chaussé A-M, Schikora A, Velge P. Interactions of salmonella with animals and plants. Front Microbiol. 2015;5:1–18.

56. Dhanani AS, Block G, Dewar K, Forgetta V, Topp E, Beiko RG, et al. Genomic comparison of non-Typhoidal salmonella enterica Serovars typhimurium, Enteritidis, Heidelberg, Hadar and Kentucky isolates from broiler chickens. PLoS One. 2015;10:e0128773.
57. Dhanani AS, Block G, Dewar K, Forgetta V, Topp E, Beiko RG, et al. Correction: genomic comparison of non-Typhoidal salmonella enterica Serovars typhimurium, Enteritidis, Heidelberg, Hadar and Kentucky isolates from broiler chickens. PLoS One. 2015;10:e0137697.
58. Weening EH, Barker JD, Laarakker MC, Humphries AD, Tsolis RM, Bäumler AJ. The salmonella enterica serotype typhimurium lpf, bcf, stb, stc, std, and sth fimbrial operons are required for intestinal persistence in mice. Infect Immun. 2005;73:3358–66.
59. Jelsbak L, Hartman H, Schroll C, Rosenkrantz JT, Lemire S, Wallrodt I, et al. Identification of metabolic pathways essential for fitness of salmonella typhimurium in vivo. Cloeckaert A, editor. PLoS One 2014;9:e101869.
60. Charlier D, Glansdorff N. Biosynthesis of arginine and polyamines. EcoSal Plus. 2004;1:1–54.
61. Tocilj A, Schrag JD, Li Y, Schneider BL, Reitzer L, Matte A, et al. Crystal structure of N-Succinylarginine Dihydrolase AstB, bound to substrate and product, an enzyme from the arginine catabolic pathway of Escherichia Coli. J Biol Chem. 2005;280:15800–8.
62. Botsford JL, Alvarez M, Hernandez R, Nichols R. Accumulation of glutamate by salmonella typhimurium in response to osmotic stress. Appl Environ Microbiol. 1994;60:2568–74.
63. Csonka LN, Ikeda TP, Fletcher SA, Kustu S. The accumulation of glutamate is necessary for optimal growth of salmonella typhimurium in media of high osmolality but not induction of the proU operon. J Bacteriol. 1994;176:6324–33.
64. Claus SP, Guillou H, Ellero-Simatos S. The gut microbiota: a major player in the toxicity of environmental pollutants? Npj Biofilms Microbiomes. 2016;2:16003.
65. Lee E-J, Choi J, Groisman EA. Control of a salmonella virulence operon by proline-charged tRNAPro. Proc Natl Acad Sci. 2014;111:3140–5.
66. Allen SW, Senti-Willis A, Maloy SRDNA. Sequence of the putA gene from salmonella typhimurium: a bifunctional membrane-associated dehydrogenase that binds DNA. Nucleic Acids Res. 1993;11:1676.
67. Esaki N, Walsh CT. Biosynthetic alanine racemase of salmonella typhimurium: purification and characterization of the enzyme encoded by the alr gene. Biochemistry (Mosc). 1986;25:3261–7.
68. Yuan C, Li JM, Ding Y, He Q, Yan HX, JJ L, et al. Estimation of L-arginine requirement for Xinyang black laying hens from 33 to 45 weeks of age. J Appl Poult Res. 2015;24:463–9.
69. H. Karasov W, Douglas AE. Comparative Digestive Physiology. In: Terjung R, editor. Compr. Physiol . Hoboken, NJ, USA: John Wiley & Sons, Inc; 2013.
70. Gilbert ER, Li H, Emmerson DA, Webb KE, Wong EA. Developmental regulation of nutrient transporter and enzyme mRNA abundance in the small intestine of broilers. Poult Sci. 2007;86:1739–53.
71. Kieboom J, Abee T. Arginine-dependent acid resistance in salmonella enterica Serovar typhimurium. J Bacteriol. 2006;188:5650–3.
72. Brenneman KE, Willingham C, Kong W, Curtiss R, Roland KL. Low-pH Rescue of Acid-Sensitive Salmonella enterica Serovar Typhi strains by a Rhamnose-regulated arginine decarboxylase system. J Bacteriol. 2013;195:3062–72.
73. Paulander W, Andersson DI. Amplification of the gene for Isoleucyl-tRNA Synthetase facilitates adaptation to the fitness cost of mupirocin resistance in salmonella enterica. Genetics. 2010;185:305–12.
74. Han J, Lynne AM, David DE, Tang H, Xu J, Nayak R, et al. DNA sequence analysis of plasmids from multidrug resistant salmonella enterica serotype Heidelberg isolates. Webber MA. PLoS One. 2012;7:e51160.
75. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30:2114–20.
76. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics. 2009;25:1754–60.
77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
78. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43:491–8.
79. 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;3:35.
80. 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.
81. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
82. Lescai F, Marasco E, Bacchelli C, Stanier P, Mantovani V, Beales P. Identification and validation of loss of function variants in clinical contexts. Mol. Genet. Genomic Med. 2014;2:58–63.
83. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
84. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.
85. Revell LJ. Phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. Methods Ecol Evol. 2012;3:217–23.
86. Armstrong RA. When to use the Bonferroni correction. Ophthalmic Physiol Opt. 2014;34:502–8.
87. Development Core Team R. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. Available from: http://www.R-project.org/
88. Yoon SH, Park Y-K, Lee S, Choi D, TK O, Hur C-G, et al. Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic Acids Res. 2007;35:D395–400.