Persistent, Private and Mobile genes: a model for gene dynamics in evolving pangenomes

Jasmine Gamblin[*,1], Amaury Lambert[†,1,2], and François Blanquart[†,1]

[1]Center for Interdisciplinary Research in Biology (CIRB), College de France, CNRS, INSERM, Université PSL, Paris, France

[2]Institut de Biologie de l'ENS (IBENS), École Normale Supérieure (ENS), CNRS, INSERM, Université PSL, Paris, France.

July 15, 2024

Abstract

The pangenome of a species is the set of all genes carried by at least one member of the species. In bacteria, pangenomes can be much larger than the set of genes carried by a single organism. Many questions remain unanswered regarding the evolutionary forces shaping the patterns of presence/absence of genes in pangenomes of a given species. We introduce a new model for bacterial pangenome evolution along a species phylogeny that explicitly describes the timing of appearance of each gene in the species and accounts for three generic types of gene evolutionary dynamics: persistent genes that are present in the ancestral genome, private genes that are specific to a given clade, and mobile genes that are imported once into the gene pool and then undergo frequent horizontal gene transfers. We call this model the Persistent-Private-Mobile (PPM) model. We develop an algorithm fitting the PPM model and apply it to a dataset of 902 *Salmonella enterica* genomes. We show that the best fitting model is able to reproduce the global pattern of some multivariate statistics like the gene frequency spectrum and the parsimony vs. frequency plot. Moreover, the gene classification induced by the PPM model allows us to study the position of accessory genes on the chromosome depending on their category, as well as the gene functions that are most present in each category. This work paves the way for a mechanistic understanding of pangenome evolution, and the PPM model developed here could be used for dynamics-aware gene classification.

[*]Corresponding author: `jasmine.gamblin@polytechnique.org`

[†]Co-last authors

The authors declare no competing interests.

Introduction

Many bacterial species present an impressive diversity in terms of gene content. Due to pervasive intra- and inter-specific horizontal gene transfer (HGT), the number of genes present in a given species is often much higher than the number of genes contained in a typical genome of this species, and can potentially be very large. As multiple genomes from the same species began to be available around the year 2005, the observation that each new sequenced genome contained new genes lead Tettelin et al. (2005) to coin the term 'pangenome' for the set of all genes present in a species (Tettelin & Medini, 2020). Genes from a pangenome are usually classified as 'core' (if they are present in every genome) or 'accessory' (if they are not) (Page et al., 2015). The diversity and organization of bacterial genes could be used to understand the mechanisms governing their evolutionary dynamics. Understanding the dynamics of accessory genes is also of interest for public health as most genes conferring higher virulence or antimicrobial resistance belong to this category.

Patterns and processes of pangenome evolution

In the genomic era, very rich information is available to study the patterns created by pangenome evolution: tens of thousands of sequenced genomes in certain bacterial species, including hundreds of completely assembled genomes. These patterns can be found in various summaries of the data. To name a few, the mechanisms of pangenome evolution probably influences the distribution of gene frequencies, the presence/absence patterns of genes, the diversity carried by gene families and the order of genes along a chromosome. For example, the gene frequency spectrum (GFS) - which is the histogram of gene frequencies in a set of genomes - often presents a 'U' shape that seems ubiquitous in bacterial pangenomes, whereby more genes are rare and very frequent than present at intermediate frequencies (Figure 5a). While these patterns can be precisely described and compared between different populations (Cummins et al., 2022; Botelho et al., 2023), an open problem is to identify the main processes generating them. To this end, we seek to develop a minimal model describing gene dynamics in a pangenome, that is able to reproduce the observed patterns. We started with three biologically relevant types of evolutionary gene dynamics, partially based on previous models of pangenome evolution, and tested several versions of our model to retain the one that was best able to reproduce these patterns.

Existing models of bacterial pangenome evolution

Most models of bacterial gene evolution study gene gains and losses along a reference species tree, usually obtained using the alignment of core genome sequences. A simple model is the classical Markovian model of binary character evolution along a phylogenetic tree, in which a gene can be gained ($0 \to 1$) and lost ($1 \to 0$) at constant rates along the tree lineages (Pagel, 1997). Cohen and Pupko (2010) used this model to detect horizontally transferred gene families, by allowing the rates to vary across gene families and using stochastic mapping to map gain and loss events on branches. Using

a similar model that allowed rate variation across time and lineages, and considering small genome chunks rather than genes as the evolutionary units, Didelot et al. (2008) inferred variations of genomic flux across lineages in several bacterial species and correlated it to their lifestyle. This model class was later called the Finitely Many Genes (FMG) model by Zamani-Dahaj et al. (2016). Indeed, this model assumes that a finite pool of genes is available to the population, which is fixed from the beginning of the species' history, and remains constant throughout evolution.

Around 2012, alternative models emerged to account for gene immigration from other species. The Infinitely Many Genes (IMG) model (Baumdicker et al., 2012) assumes that new genes are imported at a constant rate along the phylogeny, then are lost at a constant rate and cannot be regained once lost. In this model, imported genes are picked from an infinite pool. Haegeman and Weitz (2012) presented a similar model with a fixed genome size, thus imposing that a gene importation must coincide with a gene loss. These models were originally designed to explain the U-shaped gene frequency spectrum (GFS) of pangenomes, and were fitted using exclusively the GFS.

Zamani-Dahaj et al. (2016) compared these two approaches and estimated the proportion of genes that were better explained by one gain (IMG model) versus multiple gains (FMG model), finding that both types of gene dynamics are present in 40 genomes of cyanobacteria. The two models are now implemented in the software Panaroo to allow for gain and loss rates estimation (Tonkin-Hill et al., 2020).

There are several limitations to the aforementioned approaches. First, some assumptions of existing models are limiting: the IMG model does not account for intra-species gene transfers, while the FMG model does not account for inter-species transfers. When comparing both types of gene dynamics, Zamani-Dahaj et al. (2016) find that under the FMG model, 15% of genes have patterns that are better explained by multiple gains, suggesting that they have been transferred. These genes are typically not well fitted by the IMG model. Meanwhile, the FMG model assumes that all genes from the pangenome are already present in the gene pool of the species from the root of the phylogeny. In reality, many genes are imported into the focal species over its evolutionary history, sometimes very recently compared to the whole evolutionary history.

We also identify methodological limitations. On the one hand, studies aiming at fitting the GFS use a very restricted fraction of the available data (the gene frequencies). As a result, they are not able to reproduce correctly the level of parsimony observed in gene presence/absence patterns. On the other hand, while Zamani-Dahaj et al. (2016) reproduce the GFS of 40 cyanobacteria species thanks to several types of gene dynamics and allowing variability in parameters, the best model has five categories of genes. With much larger genomic datasets available, it is necessary to define parsimonious models that lend themselves more easily to interpretation. In this direction, we suggest a middle ground approach. We fix a priori the number of categories and parameters based on biological observations of gene dynamics, and check how well we are able to reproduce some of the global patterns observed in the data to decide whether more complex models are needed.

Persistent, Private and Mobile gene dynamics

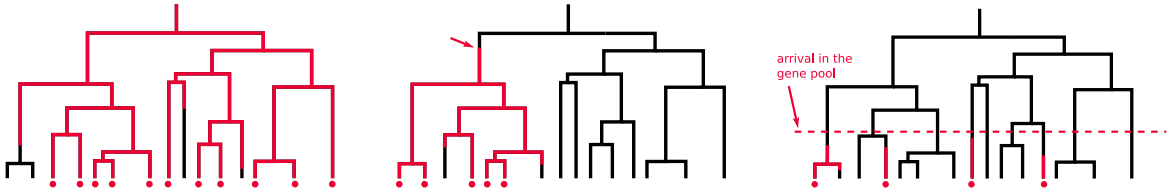
In the following, we list three qualitatively distinct gene behaviors that are relevant for bacterial populations. Our goal is to design a model of pangenome evolution accounting for these three types of gene dynamics.

First, we model the dynamics of genes that are in most circumstances essential to the bacteria, and thus are present in the ancestor and in the majority of sequenced genomes (Figure 1a). We use the adjective 'persistent' for these genes, as we expect them to largely overlap the set of genes usually designated as persistent (i.e., genes being present in a high percentage of genomes, typically between 95 and 100%: Perrin and Rocha, 2021). We hypothesise that the inferred loss rates of these genes will be very small. A loss can happen for example when a taxon colonizes a new niche where this gene is no longer essential, or by gene redundancy following the acquisition of a new gene performing the same function.

Second, we model genes that are specific to a given clade (Figure 1b). This specificity can have several explanations: a single event of transfer into the focal species, adaptation to a particular environment (e.g. bovine-adapted *Escherichia coli* lineages: Arimizu et al., 2019), defense against a phage specific to this clade (e.g. in *Vibrio crassostreae*: Piel et al., 2021), or epistatic interactions with genes already present in this clade (Whelan et al., 2021; McInerney, 2022; Beavan et al., 2024). We call such genes 'private'. Some of these genes are already present in the ancestor of the population, while others appear along the species tree either by *de novo* gene birth or by inter-species transfer.

Finally, we model the dynamics of 'mobile' genes, i.e. genes that undergo frequent intra-species HGT such as transposons, prophages or genes located on plasmids. These genes have highly non-monophyletic presence/absence patterns at the leaves of the species tree (Figure 1c). A mobile gene is first introduced into the gene pool of the population by inter-species transfer, which is a rare event, and then is able to spread in the population through frequent intra-species HGT (Tenaillon et al., 2010). For example, the $bla_{\text{CTX-M}}$ gene family encodes CTX-M β-lactamases allowing their host to tolerate β-lactam antibiotics. Originally found as a chromosomal gene in genus *Kluyvera*, it has introgressed several times in other Enterobacteriaceae species and has rapidly spread within these species on various mobile genetic elements (D'Andrea et al., 2013).

In this paper we introduce the PPM model for pangenome evolution. It includes three generic types of evolutionary gene dynamics along a species phylogeny, designed to model the aforementioned behaviors: (i) Persistent genes are present in the ancestral genome and can only be lost, (ii) Private genes are gained once by some ancestral lineage, are lost sometimes and undergo no further transfers, and (iii) Mobile genes undergo intra-species transfers following their arrival in the gene pool. We inferred the parameters of our model on a dataset of 902 *S. enterica* genomes by fitting the presence/absence patterns of genes. Using these parameter values, we simulated genomes evolving according to our model and compared the obtained patterns to the ones from observed data. By assigning each gene to its most likely category, we were able to study the distribution of the three categories along the chromosome and on plasmids, as well as the gene functions present in each category.



(a) A Persistent gene (b) A Private gene (c) A Mobile gene

Figure 1: The three behaviors captured by our model: (a) Persistent genes are present at the root of the phylogeny and rarely lost (b) Private genes arrive once in the phylogeny, transmit vertically and are occasionally lost, while (c) Mobile genes arrive once in the gene pool of the population, then undergo intra-species HGT (thus multiple gains and losses along the phylogeny).

The model

In the Persistent-Private-Mobile (PPM) model, genes evolve on the species phylogeny along which they can be gained and lost. The following events are accounted for: gene gain and loss, intra- and inter-specific horizontal gene transfer, sequencing and bioinformatic errors. For the sake of concision, we index the three gene categories of the model as follows: 0 for the Persistent category, 1 for the Private category and 2 for the Mobile category. Parameters referring to each of these categories are indexed accordingly.

- N_0 Persistent genes are present in the ancestral genome. They are lost at constant rate l_0 along the phylogeny.
- N_1 Private genes are present in the ancestral genome. At constant rate i_1, new Private genes arrive on branches of the phylogeny. Once they are present, they are lost at constant rate l_1 along the phylogeny. In the following, we fixed N_1 to be equal to the expected number of Private genes at stationarity, which is i_1/l_1 (Huson & Steel, 2004). This category follows exactly the IMG model (Baumdicker et al., 2012).
- Mobile genes enter the gene pool of the population at constant rate i_2 across time (i.e. along the tree height). Once they are in the gene pool, Mobile genes can be gained and lost at constant rates g_2 and l_2 by all lineages of the population phylogeny below their arrival date. Gains are interpreted as transfers from the pool to a tree lineage. The fact that we only model this type of transfers (rather than transfers between lineages) is consistent with the predominance of so-called 'transfers from the dead': assuming that the number of sequenced genomes is small compared to the total number of individuals in the species (which is usually the case), a transferred gene has much more chance to come from a non-sampled ('dead') lineage than from another lineage of the reconstructed tree (Szöllosi et al., 2013).

These events are represented in Figure 2.

Lastly, our model accounts for errors in the presence/absence patterns of genes at the leaves of the

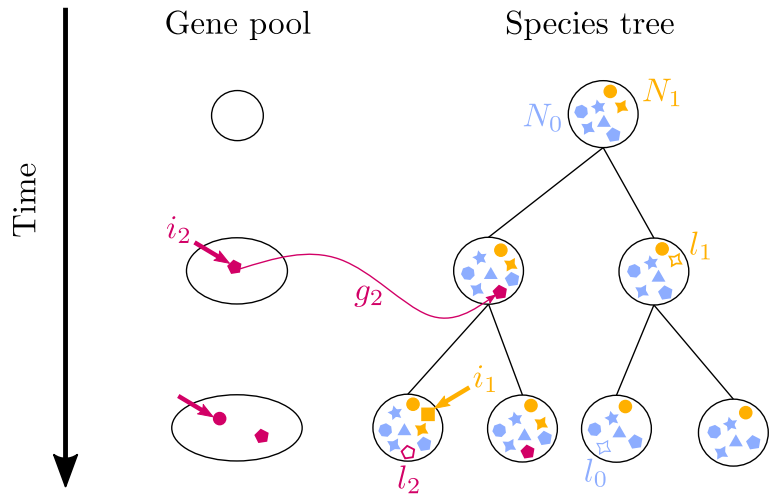


Figure 2: Schematic description of genomes evolving according to the PPM model for pangenome evolution. Persistent genes are represented in blue, Private genes in orange and Mobile genes in Burgundy. Different shapes represent different gene functions. The following events are represented, for Persistent genes: loss (rate l_0); for Private genes: immigration on the tree (rate i_1), loss (rate l_1); and for Mobile genes: immigration into the gene pool (rate i_2), gain (rate g_2) and loss (rate l_2).

phylogeny. Pangenome datasets can be subject to several sources of errors (Tonkin-Hill et al., 2020). These sources include sequencing errors (Salzberg, 2019) and bioinformatic errors (Denton et al., 2014). Moreover, the genuine biological process of rapid gain and loss events of genes on the leaves of the phylogeny could be accounted for by our 'error' term. For example if a bacterium having very recently lost an essential gene is sequenced, it would probably not have left descent over several generations of replication *in vivo* (which is the case of bacteria at internal nodes of the phylogeny). This illustrates that gain and loss events at the leaves are happening at an evolutionary scale which is different from that of the rest of the phylogeny. Our model contains three error parameters ϵ_0, ϵ_1 and ϵ_2. ϵ_0 and ϵ_1 are the false negative rates for Persistent and Private genes, that is the probability that a gene was lost just before sampling or was not detected because of incorrect sequencing or annotation. ϵ_2 is both the false positive and false negative rate for Mobile genes. A false positive would be a gene gained just before sampling or spuriously detected. We assume that the false positive rate is zero for Persistent and Private genes, as these genes can be gained only once.

Results

Simulation study

In order to assess the accuracy of our inference procedure (see Methods section), we simulated sets of presence/absence patterns and checked if we were able to infer the parameters that generated them. We randomly picked a 200-leaf tree, and 100 sets of values for parameters $N_0, l_0, i_1, l_1, i_2, g_2, l_2, \epsilon_0, \epsilon_1$

and ϵ_2. For each set of values we simulated genes evolving along the tree, and used the resulting presence/absence patterns to compute the maximum likelihood parameters. More details on this simulation study are given in the Supplementary Information. Figure 3 shows the relations between the true parameters (i.e. those used for simulations) and the inferred ones.

Most parameters show good correlations, with relative errors below 10% (or log-scale error below 0.1). In particular this is true for parameters governing the number of genes in each category: N_0, i_1 and i_2. The error rates of the Private and Mobile categories (ϵ_1 and ϵ_2) show higher relative errors (128% and 76%). The gain and loss rates for the Mobile category are not very well estimated with log-scale errors of 0.52 and 0.21, meaning they can be estimated up to a factor of respectively 3 and 1.5. This is due to very poor inference when l_2 is more than 100 times bigger than g_2 (Figure 3h). We discuss this phenomenon in the Supplementary Information, and show on Figure S4 that this ratio should be well estimated in the case of the *S. enterica* dataset. The average proportion of genes correctly assigned to their category is 97% across the 100 simulated sets (Figure 3l).

Maximum likelihood estimates of parameters

We inferred the Maximum Likelihood parameters of our model on a dataset of 902 *S. enterica* genomes, and found the following estimates:

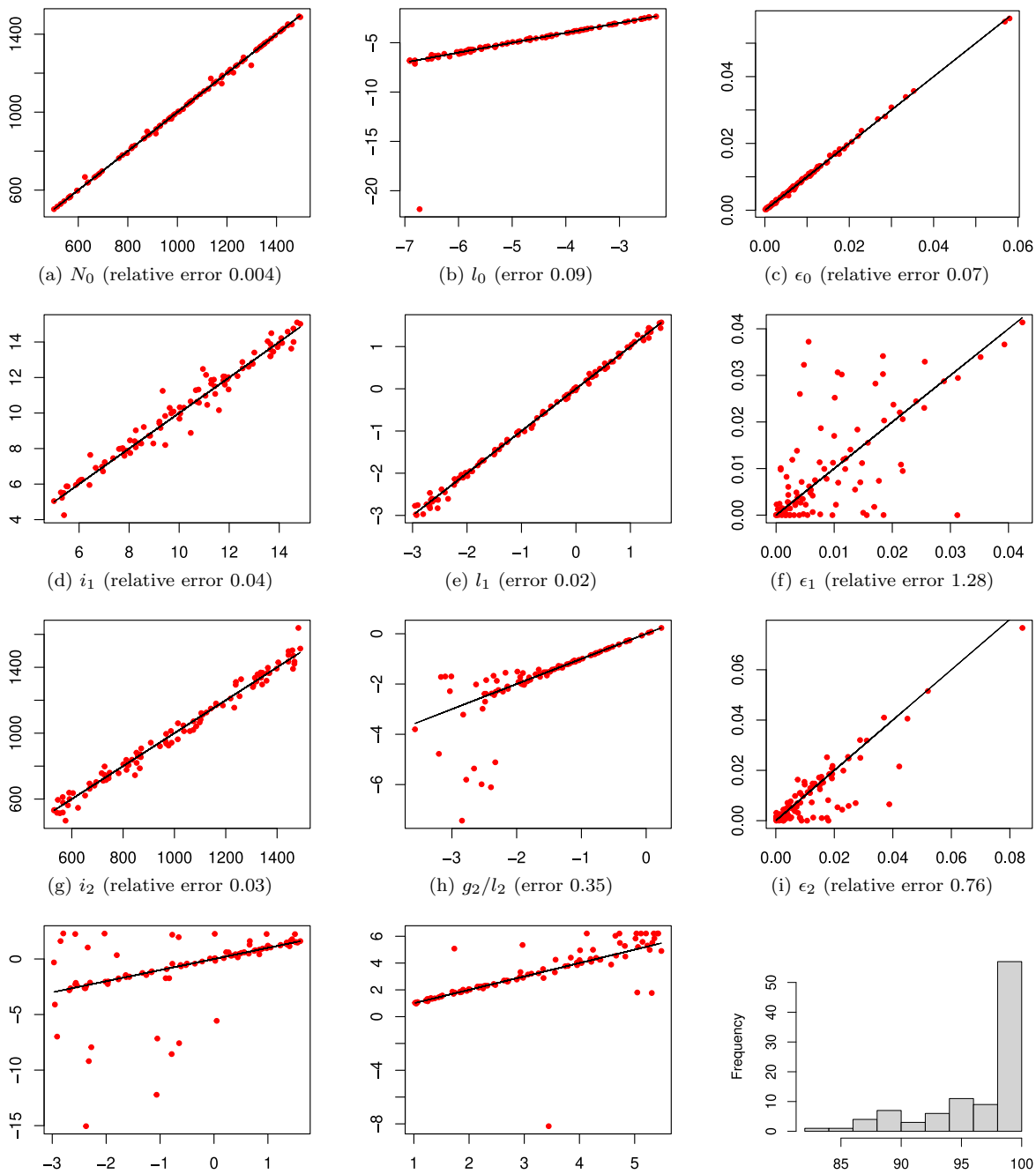
$$\left.\begin{array}{lll} \hat{N}_0 & = 4,014 & \text{CI: } [3894, 4134] \\ \hat{l}_0 & = 2.25 & \text{CI: } [2.18, 2.32] \\ \hat{\epsilon}_0 & = 0.00213 & \text{CI: } [0.00208, 0.00218] \end{array}\right\} \begin{array}{l} 4,007 \text{ observed} \\ \text{Persistent genes} \end{array} \tag{1}$$

$$\left.\begin{array}{lll} \hat{i}_1 & = 45,548 & \text{CI: } [45150, 45946] \\ \hat{l}_1 & = 152.4 & \text{CI: } [150.8, 154.0] \\ \hat{\epsilon}_1 & = 0.0188 & \text{CI: } [0.0182, 0.0194] \end{array}\right\} \begin{array}{l} 26,797 \text{ observed} \\ \text{Private genes} \end{array} \tag{2}$$

$$\left.\begin{array}{lll} \hat{i}_2 & = 462,793 & \text{CI: } [448832, 475463] \\ \hat{g}_2 & = 222.8 & \text{CI: } [220.0, 225.6] \\ \hat{l}_2 & = 12,858 & \text{CI: } [12717, 12999] \\ \hat{\epsilon}_2 & = 0.00225 & \text{CI: } [0.00220, 0.00230] \end{array}\right\} \begin{array}{l} 14,951 \text{ observed} \\ \text{Mobile genes} \end{array} \tag{3}$$

From these estimates, we computed several quantities of interest. The predicted number of genes in the ancestral genome is $\hat{N}_0 + \hat{i}_1/\hat{l}_1 = 4,313$ (CI: $[4187, 4439]$), close to the mean number of genes per genome in the dataset (4,556).

The expected number of genes of each category in the observed pangenome is computed using the forthcoming set of equations (6), and we find an expected number of 4,007 Persistent genes, 26,797 Private genes and 14,951 Mobile genes. This means that the observed pangenome is composed of 8% of Persistent genes, 60% of Private genes and 32% of Mobile genes, as shown on Figure 4. We also assigned each gene to its most likely category, and computed the proportion of each category present in an average chromosome and in an average plasmid (Figure 4). As Persistent genes are typically present in many genomes, they represent a small proportion of the observed pangenome (8%) but a high proportion of the genes carried by a chromosome (83%). The opposite holds for Private and



(a) N_0 (relative error 0.004)
(b) l_0 (error 0.09)
(c) ϵ_0 (relative error 0.07)
(d) i_1 (relative error 0.04)
(e) l_1 (error 0.02)
(f) ϵ_1 (relative error 1.28)
(g) i_2 (relative error 0.03)
(h) g_2/l_2 (error 0.35)
(i) ϵ_2 (relative error 0.76)
(j) g_2 (error 0.52)
(k) l_2 (error 0.21)
(l) Proportion of genes assigned to the correct category

Figure 3: Results of the cross-validation study on a 200-leaf tree. (a-k) show correlations between true values (x-axis) and point estimates (y-axis) of parameters used to simulate sets of presence/absence patterns. Correlations for gain and loss parameters ($l_0, l_1, g_2/l_2, g_2$ and l_2) are plotted on a log scale. The legends of subplots display the mean relative error for each parameter (or mean error for log-scale plots). (l) is a histogram showing the proportion of patterns that were assigned to the correct category by our method across the 100 simulated sets.

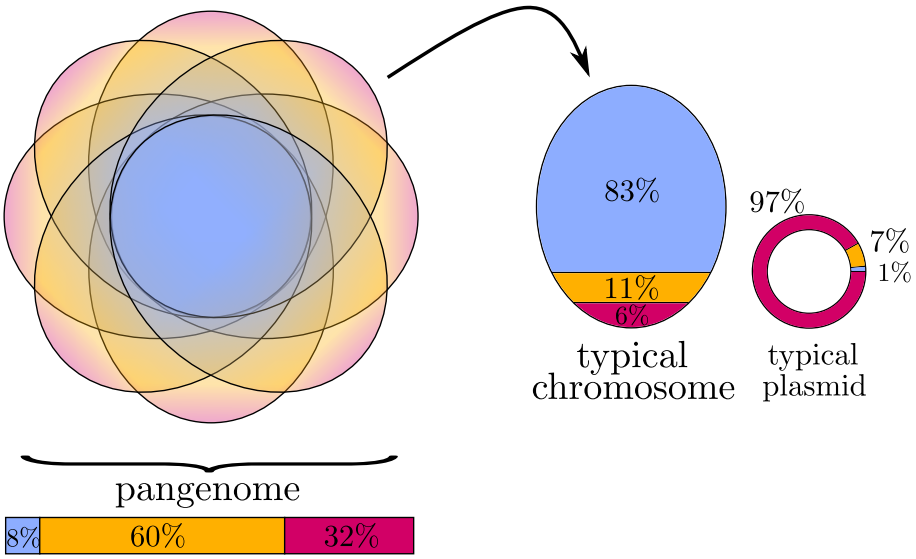


Figure 4: Proportion of the Persistent, Private and Mobile categories in our *S. enterica* dataset as inferred by our model. Colors indicate gene categories: blue for Persistent, orange for Private and Burgundy for Mobile. The observed pangenome is composed of 8% of Persistent genes, 60% of Private genes and 32% of Mobile genes. A typical chromosome carries 83% of Persistent genes, 11% of Private genes and 6% of Mobile genes, while a typical plasmid carries 1% of Persistent genes, 7% of Private genes and 97% of Mobile genes.

Mobile genes: they compose a small proportion of the chromosome (11% and 6%, respectively) but the majority of the pangenome. A typical genome from our dataset carries 4,556 genes, of which 3,701 are Persistent, 504 are Private and 352 are Mobile.

The loss rates are ordered as we expected, increasing from Persistent to Private to Mobile gene categories (although from the simulation study we know that the absolute value of l_2 may not be very well inferred). The values can be compared with the average branch length of the tree, which is 5×10^{-4} substitution per site. This length corresponds to a few thousand years if we divide by the estimated mutation rate of 10^{-7} substitution per site per year for *S. enterica* (Duchêne et al., 2016). During this time interval, Persistent, Private and Mobile genes have a probability 0.001, 0.07 and 0.98 to be lost respectively. Additionally, a Mobile gene has a probability 0.02 to be gained if it was absent. These probabilities were computed using the forthcoming transition matrix (15).

During these thousand years, a given genome is expected to gain 21 Private genes that may result from *de novo* gene birth and direct transfers from other species, and 722 genes from the pool (if we assume that we are in the middle of the tree and the pool has around $i_2 \times H/2$ genes, where H is the tree height). The large probability of loss and large number of new gains of Mobile genes over thousand years implies that Mobile genes completely turnover over that period of time. By construction, a Private gene is introduced only once in the phylogeny. On the other hand, Mobile genes can be introduced several times from the gene pool to lineages of the phylogeny. We computed the expected number of distinct introductions for a Mobile gene using forthcoming equation (23), and found that

they are introduced 25 times on average.

The error rate is quite small for the Persistent and Mobile categories (around 0.2%), and more important for the Private category (around 2%). This may indicate that the presence/absence patterns at the leaves are primarily driven by the underlying processes and not by random error.

Goodness of fit on two summary statistics: the GFS and parsimony plot

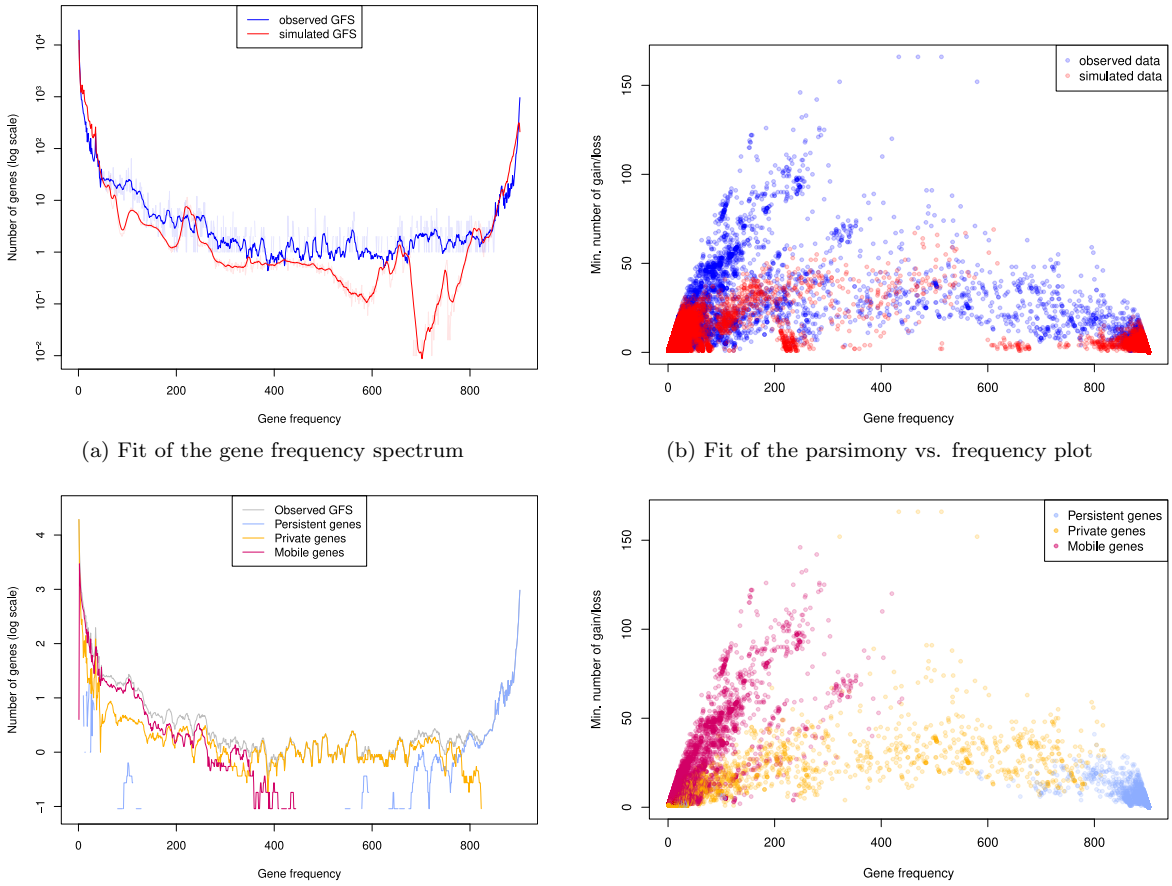
While we fitted our model using the presence/absence matrix, we verified that it is also able to reproduce two important patterns of summary statistics characterising the observed data. To that end, we simulated a pangenome using our model with the maximum likelihood parameter values inferred on the *S. enterica* dataset, and compared patterns from observed data versus simulated data.

The first pattern is the gene frequency spectrum (GFS), i.e. the distribution of gene frequencies in the observed pangenome. We chose the GFS as a summary statistic because it was used to fit pangenome evolution models in several previous studies. We plotted on Figure 5a the observed GFS (in blue) as well as the simulated GFS (in red). While the two curves do not perfectly correspond, in particular with a deficit of medium-high frequency genes in our model, the model qualitatively reproduces the U-shape found in the data.

The second pattern is the parsimony vs. frequency plot, which shows the parsimony of gene patterns as a function of their frequency. The y-axis is the minimal number of gains and losses needed along the species tree to explain a gene's presence/absence pattern, thus small values correspond to very parsimonious patterns. We represented on Figure 5b the observed parsimony plot in blue and the simulated one in red. Our model reproduces the qualitative patterns of this plot, but the presence/absence patterns it generates are more parsimonious than in the observed data. In particular, the maximum score reached by simulated genes is around 50 while some observed genes go as high as 150. For a more detailed inspection of this parsimony vs. frequency plot, we plotted separate plots for each category on Supplementary Figure S7.

Inferred categories

Using the ML parameter estimates, we computed for each gene the category with highest posterior probability. We assigned each gene to a category–Persistent, Private or Mobile. We then compared different characteristics of the three categories, such as frequency, parsimony, position on the genome and enrichment in certain gene functions. The ternary plot on Supplementary Figure S3 show that the assigned categories are clear-cut, with genes clustered around the corners of the triangle.
Figures 5c and 5d show the GFS and parsimony plot colored according to assigned categories. Genes assigned to the Persistent category have typically very high frequency, i.e. they are positioned on the right of these plots. Surprisingly, we also spot a few genes at low frequency that are assigned to the Persistent category. Taking a closer look at their presence/absence patterns revealed that these genes



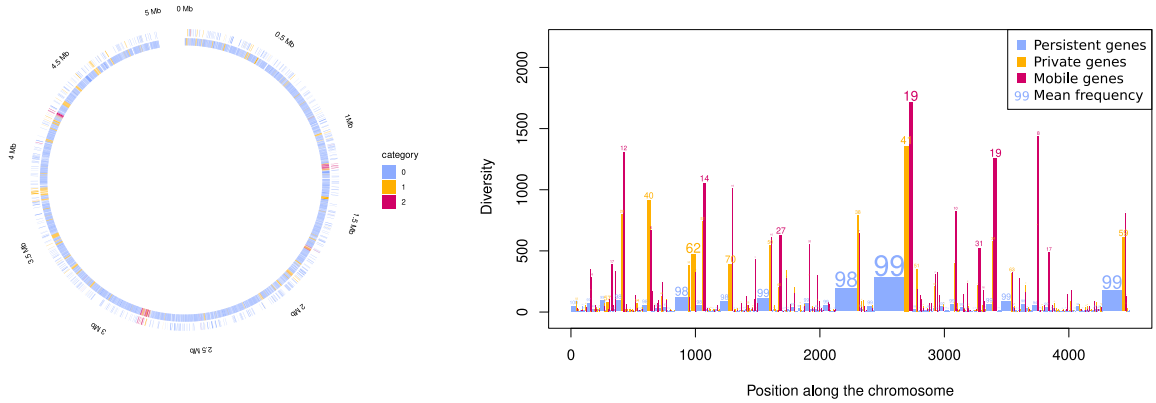
(a) Fit of the gene frequency spectrum

(b) Fit of the parsimony vs. frequency plot

(c) Gene frequency spectra of the three categories

(d) Parsimony vs. frequency plot, colored according to assigned categories

Figure 5: Two multivariate summary statistics of pangenomic data that we use in order to assess the goodness-of-fit of our model, here plotted for a *S. enterica* dataset of 902 genomes. (a) and (c) show the histogram of gene frequencies, called the Gene Frequency Spectrum (GFS). Curves are smoothed for better visualization; raw curves are shown in transparency on (a). (b) and (d) show the parsimony vs. frequency plot: the y-axis is the minimum number of gain and loss events needed along the species tree to explain the presence/absence pattern of a gene with frequency given on x-axis. (a) and (b) show the fit of these two statistics with our model. Observed data are plotted in blue, while simulated data are in red. Simulated data were obtained by simulating a set of genes evolving according to our model with Maximum Likelihood parameter estimates. The simulated GFS is an average over 100 simulations. (c) and (d) show the gene frequency spectrum and parsimony vs. frequency plot colored according to assigned categories.



(a) Chromosome of genome n°1

(b) Diversity of gene categories along the chromosome

Figure 6: Distribution of gene categories along the chromosome. (a) shows gene categories along the chromosome of the first genome (chosen arbitrarily) of the *S. enterica* dataset. (b) shows the diversity of gene categories along the *S. enterica* chromosome, computed over 641 genomes sharing the same ordering of core genes. The position along the chromosome refers to the order of genes. Each bar represents genes from a given category present in a given interval between core genes. The width of the bar is the mean number of genes of this category present in this interval, and the height is the number of unique genes of this category present in this interval (i.e. the diversity). The number above the bar is the mean frequency of these genes.

are present at high frequency in two or more small clades. As a consequence, they are not well-fitted by the Private or the Mobile category. Genes assigned to the Private category present a wide range of frequencies, from very low (the majority of singletons is 'Private') to intermediate and high frequencies. While we could have expected singletons to be assigned preferentially to the Mobile category or to be encompassed by the error term, it is not the case as Mobile genes have a tendency to be introduced several times (and thus, to be present in several genomes) and the error term is constrained by other patterns. Genes assigned to the Mobile category have low to intermediate frequencies and represent the majority of genes having low parsimony (i.e., in the upper part of the parsimony plot).

Gene position

We studied the spatial distribution of the three gene categories on the bacterial chromosome and on plasmids. On Figure 4 is represented the relative proportion of the three categories on a typical chromosome versus on a typical plasmid from the *S. enterica* dataset. As expected, we observe that chromosomes carry a majority of Persistent genes (83%), with some Private and Mobile genes (11% and 6%, respectively). On the other hand, plasmids carry almost exclusively Mobile genes (97%), which is consistent with their easy transfer.

We looked more closely at the spatial distribution of genes along a chromosome, starting arbitrarily with the first genome of the dataset, represented on Figure 6a. Gene are colored according to their

assigned category (Persistent in blue, Private in orange, Mobile in Burgundy). In accordance with the relative category proportions described above, we observe a majority of Persistent genes on the chromosome. Moreover, Private and Mobile genes are not uniformly distributed but appear clustered.

To investigate whether the structure in gene category along the chromosome and the hotspots of Private and Mobile genes are conserved across strains, we used the 770 core genes as reference positions. We identified 641 genomes from the *S. enterica* dataset that had the exact same ordering of core genes. For each genome and each interval between two core genes, we computed the number of genes of each category that were present in this interval. We illustrate the diversity and distribution of gene categories along the chromosome of these 641 genomes with conserved synteny (Figure 6b). Hotspots of Private and Mobile genes that are conserved across the genomes appear all along the chromosome. These hotspots include from one gene to a few dozen genes with low frequency, but across all genomes carry a substantial gene diversity. Notably, the six most diverse hotspots of Mobile genes each carry more than 1000 unique genes across the 641 genomes. Mean diversity of Private genes in an interval is strongly correlated with mean diversity of Mobile genes in the same interval (Pearson correlation coefficient 0.91). The correlation is weaker for the mean number of genes of each of these categories in a given interval (0.58). The concentration of accessory genes in a few hotspots is in agreement with a study from Oliveira et al. (2017), where the authors find that genes imported by HGT are concentrated in hotspots representing around 1% of the genome in a dataset comprising 80 bacterial species.

Gene functions

In order to study the gene functions present in each category, we selected genes that were annotated with a Cluster of Orthologous Genes (COG) database function (Galperin et al., 2021). A total of 2,743 Persistent genes, 5,613 Private genes and 2,638 Mobile genes were annotated with a function, representing respectively 70%, 21% and 18% of these categories. From this we built an association table between assigned gene category and COG function, illustrated on Figure 7. Overall, there is a highly significant association between gene category and gene function ($p << 10^{-6}$).

In particular, there is a positive association between the Persistent category and key cellular functions such as translation (J), cell cycle control (D) and the transport and metabolism of coenzymes (H), amino acids (E), lipids (I), inorganic ions (P) and nucleotides (F). On the contrary, the Private and Mobile categories have nearly opposite associations to the Persistent genes. They are both positively associated with the mobilome (X). The Private category is also associated with carbohydrate transport and metabolism (G) and extracellular structures (W). In addition, the Mobile category is positively associated with replication, recombination and repair (L), defense mechanisms (V), signal transduction mechanisms (T) and cell motility (N). The fact that many genes involved in 'replication, recombination and repair' are inferred to be Mobile genes is probably due to transposases (enzymes helping moving transposons) annotated as recombinases.

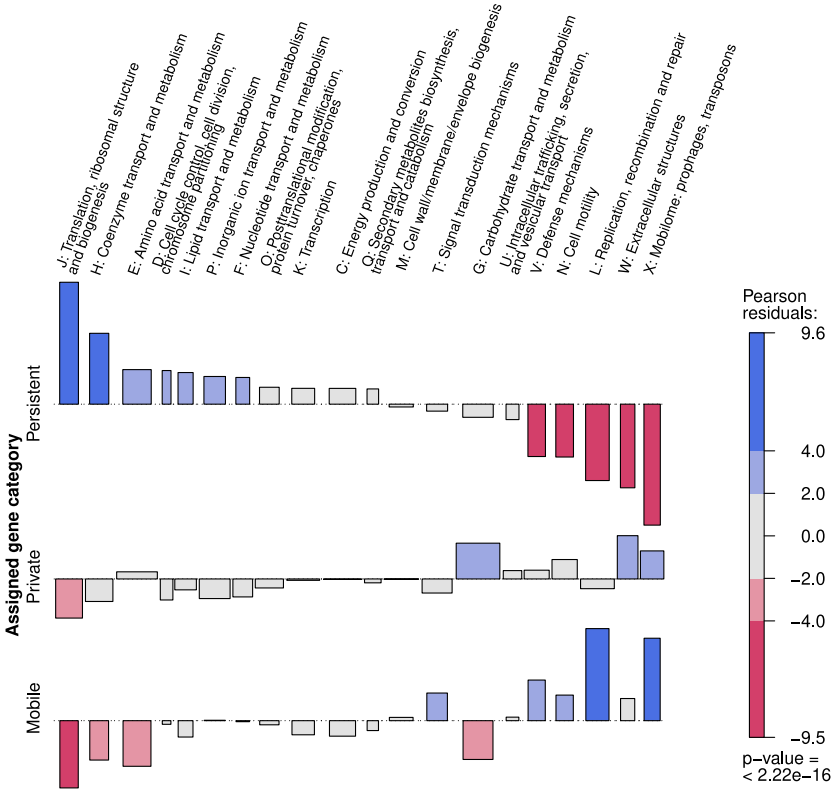


Figure 7: Association between inferred gene categories (Persistent, Private, Mobile) and COG functions. Rectangle heights represent the Pearson residuals, i.e. the observed number of genes corresponding to a given pair of category and function, minus the expected number of such genes if category and function were independent, divided by the square root of the expected value. Rectangle widths are proportional to the square root of expected values, so that rectangles areas are proportional to the difference between observed and expected value. A positive residual indicates a positive association between a gene category and a gene function, while a negative one indicates the opposite.

Discussion

In this paper, we introduce the "Persistent, Private, Mobile" model for bacterial pangenome dynamics. This model proposes to classify bacterial genes into three qualitatively different dynamics on a phylogenetic tree, accounts for both intra- and inter- species gene transfer, and is scalable to large datasets. The model fitted to a dataset of 902 *S. enterica* genomes reproduces the U-shaped gene frequency spectrum and the parsimony plot, a new representation of gene configuration on a phylogeny. Moreover, the category assignment is clear-cut and the behaviors that we intended to capture in the three gene classes are consistent with the inferred gain and loss parameters, as well as with the function and position of genes in the genome. This model could therefore be used for dynamics-aware gene classification.

A model for dynamics-aware category assignment

The PPM model was intended to model three gene categories with different behavior and characteristics. The Persistent category was designed for housekeeping genes present at high frequency. Indeed, we find that genes assigned to this category have a low loss rate and a high mean frequency (854). They are inferred to form the majority of genes present on the chromosome of *S. enterica*. Moreover, this category is enriched in essential functions such as translation, cell cycle control and metabolism. The Private category was designed for accessory genes that are specific to a given clade. Genes assigned to this category have an intermediate loss rate and a small mean frequency (17). Interestingly, they are inferred to form the majority of the observed pangenome, mainly because singletons are classified as 'Private' and therefore represent 70% of this category. The 'Private' category is enriched in genes linked to carbohydrate transport and metabolism, and to extracellular structures. These two associations could be explained by a lasting clade-specific adaptation to different nutrients and environments. Studies performed on *S. enterica* have shown that different serovars display specific carbohydrate metabolism pathways (Seif et al., 2018) and adhesive mechanisms (e.g. pili and adhesins, see Wagner and Hensel, 2011). This category is also enriched in genes from the mobilome (transposons, prophages), which may be driven by singletons. The Mobile category was designed for accessory genes undergoing frequent intra-species HGT. Genes assigned to this category have a high loss rate and a low mean frequency (22). They represent the overwhelming majority of genes found on plasmids. This category is positively associated with gene functions that are known to be highly transferred: prophages, transposons, transposases, defense mechanisms. It is also associated with signal transduction mechanisms and cell motility.

Overall, our classification identifies persistent genes, genes adapted to certain clades and highly mobile genes. This classification cannot be obtained solely based on the gene frequency, even when additionally considering parsimony (Figures 5c and 5d). For example, Private and Mobile genes can have similar frequency and parsimony (Figure 5d), but are actually organized differently on the species phylogeny. The strength of our classification of bacterial genes is that it is informed by patterns and relies on processes. The fact that it uses the gene's inferred dynamics to classify it allows for a straightforward interpretation. Comparatively, the classification done by the tool PPanGGOLiN also uses the

presence/absence matrix, but relies only implicitly on processes by assuming that neighboring genes have more chance to share the same category (Gautreau et al., 2020).
Classifying genes on the basis of their dynamics may provide clues as to how genes of interest spread. This has also enabled us to study the distribution of different types of genes along the chromosome and on plasmids. As a result, we have identified hotspots of accessory gene insertion (Private and Mobile) along the chromosome of *S. enterica* (Figures 6a and 6b). In addition, dynamics-informed classification is potentially of great interest in comparative pangenomic analysis, providing an easy-to-interpret overview of the types of dynamics at work in different populations or species.

Limitations

Our model has several limitations. First, it does not account for gene duplications. Although the method we use to generate gene families allows them to have multiple members per genome, we ignore duplicated genes by taking a binarized version of the presence/absence matrix. Thus, duplicated genes are treated as one gene even if they are sometimes present in multiple copies. That being said, families containing at least one duplicate represent 8% of all families, and those containing at least two duplicates are only 3% of all families.

Second, we do not model the simultaneous transfer of multiples genes. Kloub et al. (2021) designed a method called HoMer to detect potential horizontal multi-gene transfers (HMGTs). In a dataset of 103 *Aeromonas* genomes, they detected that 8.5% of potential intra-species gene transfers were contained in an HMGT event (this percentage was much higher for inter-species transfers, of which 20% were estimated to be part of a HMGT). We expect this phenomenon to be particular relevant for Mobile genes, which are the majority of genes carried by plasmids. HMGT would be evident in our data as sets of mobile genes that are in strong linkage disequilibrium and therefore share similar frequency and similar parsimony.

Third, we do not explicitly include selection acting on gene presence/absence, although the interpretation of the dynamics of gene categories can involve selection: Persistent genes are under purifying selection, Private genes can interact epistatically with the genetic background of the clade, Mobile genes can be favorable in some environmental conditions but not in others.

Last, the importance of gene transfers in bacteria rises the question of how reliable are phylogenetic reconstructions, and what they represent. Studies have shown that a bacterial species tree can be reliably inferred even in the presence of gene transfers and conversions (in *Escherichia coli*: Touchon et al., 2009), and that phylogenetic distance is well correlated with gene content (in *Acinetobacter*: Touchon et al., 2014, in Archea: Wolf et al., 2016). In our *S. enterica* dataset, we evidence a clear phylogenetic signal in the gene patterns by comparing the parsimony of observed patterns to that of patterns simulated without tree (Supplementary Figure S2). Phylogenetic trees are usually supposed to represent relatedness relationships between individuals, or clonal relationship in the case of bacteria. A different view has recently emerged, suggesting that reconstructed bacterial trees reflect the structure of recombination fluxes inside the population rather than actual clonal relationships (Sakoparnig et

al., 2021). While the present study does not bring new arguments supporting either of these views, we believe that the PPM model remains relevant even in the latter one. In this view, the fact that Private genes are restricted to a given clade would not be a consequence of vertical inheritance (i.e., no transfers) but of transfers restricted to a certain subpopulation, represented by a clade in the core tree. Methods exist to describe how the molecular diversity of genes is shaped by gene duplication, transfer and loss on a species tree, but these computationally intensive methods would currently not be adapted for comparative pangenomics on large datasets (Szöllosi et al., 2012; Szöllosi et al., 2015; Jacox et al., 2016; Bansal et al., 2018).

A new pangenomic statistic to visualize goodness-of-fit

As the field of pangenome evolution is largely in its infancy, it is important to devise new statistics to visualize pangenome data and assess goodness-of-fit of models (Baumdicker & Kupczok, 2023). While the first statistics used to describe bacterial pangenomes were the core genome rarefaction and pangenome enrichment curves, the Gene Frequency Spectrum (GFS) was subsequently used to calibrate models of pangenome evolution (Baumdicker et al., 2012; Haegeman & Weitz, 2012), compare different models (Lobkovsky et al., 2013), and compare processes at work in clonal versus recombinant bacterial species (Bolotin & Hershberg, 2015).
We introduce a new statistic for representing the distribution of genes in a set of genomes: the parsimony vs. frequency plot. Each gene is represented by a point, the abscissa of which corresponds to the gene's frequency, and the ordinate to the minimum number of gains and losses required along the species tree to generate the gene's presence/absence pattern. A gene's parsimony score can be seen as indicating whether the gene is preferentially present in phylogenetically close genomes or not. The parsimony vs. frequency plot is a two-dimensional statistic, and therefore contains more information than the GFS (the latter can be found by projecting the scatterplot onto the x-axis). It can be used to visually compare pangenomes from different populations, or to estimate the goodness-of-fit of a model, as we do.

These two statistics (GFS and parsimony vs. frequency plot), along with the visualization of some patterns with poor likelihood, help us suggest ways to improve our model. We chose to formulate a parsimonious model with a small number of gene categories to interpret the classification resulting from the model in the light of genome organisation, and gene function. Some developments of the PPM model could help increase the goodness-of-fit. In particular, in our data many genes have low to moderate frequency and non-parsimonious patterns (with more than 50 gain/loss events) and many genes have medium-large frequency (carried by around 600-800 genomes), features that our model can hardly reproduce (Figures 5a and 5b). One solution would be to split the Private and Mobile categories into subcategories with different rates. However, the genes with patterns that are poorly fitted often exhibit behaviours intermediate between Private and Mobile: either present at high frequency in several clades, or present at high frequency in one clade and low frequency in the rest of the tree. These could be better fitted if Private genes were allowed to be gained several times, or if genes were allowed to change categories over time: Private genes could become Persistent at some point (i.e., reduce their loss rate) or Mobile could become Private in certain clades (i.e., be lost forever when they are lost, but

at a lower rate). Allowing genes to change category would also be a way to ensure stationarity in the gene number of each category; in the current version of the model Persistent genes can only decrease in number through time and Mobile genes can only increase in number.

Gene pool model

One of the new modelling ideas introduced in this paper is the gene pool model used for Mobile genes. This model accommodates intra- and inter-species exchanges, and is both less cumbersome and more relevant than a model based on explicit exchanges between lineages (because, however large the dataset, sequenced genomes will always represent a tiny fraction of a bacterial species). Even more interestingly, our simulation study suggests that it should be possible to infer a date of introduction for each Mobile gene, if applied to a set of genomes from a recently emerged clade (see Supplementary Information and Figures S5 and S6 for details on arrival time inference). This would allow us to infer the temporal construction of the pangenome of the clade by progressive addition of genes. In future works, this model could be used to build a phylo-ecological model taking into account both gene dynamics and ecological niche, to study the formation of multiple "niche gene pools" as lineages move between different environments.

In conclusion, the PPM model classifies gene dynamics into three qualitatively different behaviours, infers relevant rates, and identifies hotspots of gene exchanges along the genome. Applied to the large number of bacterial whole-genome sequences, it could significantly improve our understanding of bacterial pangenome dynamics and evolution.

Methods

Salmonella enterica dataset

We applied our model to a dataset of 902 *S. enterica* genomes and one *Salmonella bongori* as outgroup, downloaded from the Refseq database. We chose to work with complete genome assemblies, as draft assemblies can cause an artificial increase of the gene number due to fragmented genes (Denton et al., 2014). We thus selected all complete *S. enterica* genomes that were available in this database, which yielded 1374 genomes (genomes downloaded the 07/09/2023, Refseq accession numbers listed in Supplementary Material). We performed the pangenome analysis using the PanACoTA pipeline (v.1.4.0, Perrin and Rocha, 2021). After quality filtering 903 genomes were kept (with a maximal distance of 0.1 between 2 genomes). 46,146 gene families were detected. The persistent genome, that we defined as gene families having exactly one member in at least 99% of genomes, is composed of 2,887 families. The species tree was reconstructed using sequences from the persistent genome with IQ-TREE 2 (v.2.1.2, Minh et al., 2020). We used an ultrametric version of this tree obtained with LSD2 (v.2.4.4, To et al., 2016), which is shown on Supplementary Figure S1. This tree has a height of

$H = 0.03$ and a total branch length of $L = 0.84$ substitution per site.

The data that we use in order to fit our model is the reconstructed species tree, as well as the binary presence/absence matrix describing which gene is present in which genome. As the gene clustering step of the pipeline allows gene families to have more than one member per genome, the presence/absence matrix is in fact non-binary. However, our model does not account for gene duplication, thus we use the binarized version of this matrix.

Likelihood computation

In the following, we describe how to compute the likelihood of the data according to our model. Recall that the Persistent, Private and Mobile categories are indexed by 0, 1 and 2, respectively. Let us first introduce some variables. The number of genes in the pangenome of the bacterial population described by the model is

$$N = N_0 + N_1 + I_1 + I_2 \tag{4}$$

where $I_1 \sim Pois(i_1 L)$ and $I_2 \sim Pois(i_2 H)$. L is the total length of the species tree $\mathcal{T}$, H is its height. Remember that we chose to fix $N_1 = i_1/l_1$, which is the mean number of Private genes in a genome at equilibrium (Huson & Steel, 2004). The number of genes observed at the leaves is:

$$N^{obs} = N_0^{obs} + N_1^{obs} + I_1^{obs} + I_2^{obs} \tag{5}$$

where

$$\begin{aligned}
\mathbb{E}[N_0^{obs}] &= p_0^{obs}(root) N_0 \\
\mathbb{E}[N_1^{obs}] &= p_1^{obs}(root) \frac{i_1}{l_1} \\
\mathbb{E}[I_1^{obs}] &= \int_{\mathcal{T}} i_1 p_1^{obs}(x) \, \mathrm{d}x \\
\mathbb{E}[I_2^{obs}] &= \int_{t=0}^{H} i_2 p_2^{obs}(t) \, \mathrm{d}t
\end{aligned} \tag{6}$$

with $p_0^{obs}(root)$ the probability for a Persistent gene to be observed a the leaves, $p_1^{obs}(x)$ the probability for a Private gene that arrived at position x on the tree to be observed at the leaves, and $p_2^{obs}(t)$ the probability for a Mobile gene that arrived in the population at time t to be observed at the leaves. It can be noted that the integrals used to compute $\mathbb{E}[I_1^{obs}]$ and $\mathbb{E}[I_2^{obs}]$ are not of the same nature. In the first one, we integrate over every possible position in the tree, while in the second one we integrate over time. In the following, we will approximate $N_0^{obs}, N_1^{obs}, I_1^{obs}$ and I_2^{obs} by their expectation.

Although immigrant genes from the Private category (resp. from the Mobile category) share the same realization of the random variable I_1 (resp. I_2), we make the simplifying assumption that all genes are independent and compute a pseudo-likelihood of observed data:

$$\mathcal{L}(\text{data}; \Theta) = \prod_{r_g \text{ observed}} \mathbb{P}_\Theta(r_g \,|\, g \text{ observed})^{\#r_g \text{ observations}} \tag{7}$$

where $\Theta = (N_0, i_1, l_1, i_2, l_2, g_2, \epsilon_0, \epsilon_1, \epsilon_2)$ and r_g is the pattern of presence/absence of gene g. Then for each observed pattern we have 4 components to compute. The four components correspond to the

three categories, with the Private category having two components (for genes that were present in the ancestral genome and those that were not).

$$\begin{aligned}
&\mathbb{P}_\Theta(r_g \,|\, g \text{ observed}) = \\
&\qquad \mathbb{P}_\Theta(\text{cat}_g = 0 \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 0, g \text{ observed}) \\
&\quad + \mathbb{P}_\Theta(\text{cat}_g = 1, g \in root \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 1, g \in root, g \text{ observed}) \\
&\quad + \mathbb{P}_\Theta(\text{cat}_g = 1, g \notin root \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 1, g \notin root, g \text{ observed}) \\
&\quad + \mathbb{P}_\Theta(\text{cat}_g = 2 \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 2, g \text{ observed})
\end{aligned} \tag{8}$$

Likelihood for Persistent genes

For a Persistent gene g, we denote by $\mathbb{P}^{(0)}_{\text{root}}(r_g)$ the probability to observe presence/absence pattern r_g given that gene g was present at the root (which is always the case for Persistent genes). The (0) exponent is used to indicate that this probability is computed for a Persistent gene, i.e. a gene from category 0. This quantity only depends on the loss rate l_0 and the false negative rate ϵ_0, and can be computed using Felsenstein's pruning algorithm (Felsenstein, 1981). Formally, we have:

$$\mathbb{P}^{(0)}_{\text{root}}(r_g) := \mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 0) \tag{9}$$

Thus:

$$\begin{aligned}
&\mathbb{P}_\Theta(\text{cat}_g = 0 \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 0, g \text{ observed}) \\
&\quad = \frac{N_0^{obs}}{N^{obs}} \times \mathbb{P}^{(0)}_{root}(r_g \,|\ \text{obs}) \\
&\quad = \frac{N_0 p_0^{obs}(root)}{N^{obs}} \times \frac{\mathbb{P}^{(0)}_{root}(r_g)}{p_0^{obs}(root)} \\
&\quad = \frac{N_0}{N^{obs}} \times \mathbb{P}^{(0)}_{\text{root}}(r_g)
\end{aligned} \tag{10}$$

The first equality is obtained by writing the definition of each of the two factors. The second equality is obtained using equation (6) and the fact that we approximate N_0^{obs} by its expectation.

Likelihood for Private genes

For Private genes, the term corresponding to ancestral genes is similar to the one for Persistent genes:

$$\begin{aligned}
&\mathbb{P}_\Theta(\text{cat}_g = 1, g \in root \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 1, g \in root, g \text{ observed}) \\
&\qquad\qquad = \frac{i_1/l_1}{N^{obs}}\mathbb{P}^{(1)}_{\text{root}}(r_g)
\end{aligned} \tag{11}$$

For an imported Private gene g, we have to integrate over all possible tree positions where this gene could have appeared. These possible positions are located on the path between the most recent common ancestor of the leaves carrying this gene (denoted by $mrca(r_g)$) and the root of the tree. This is due to the fact that Private genes can be gained only once.

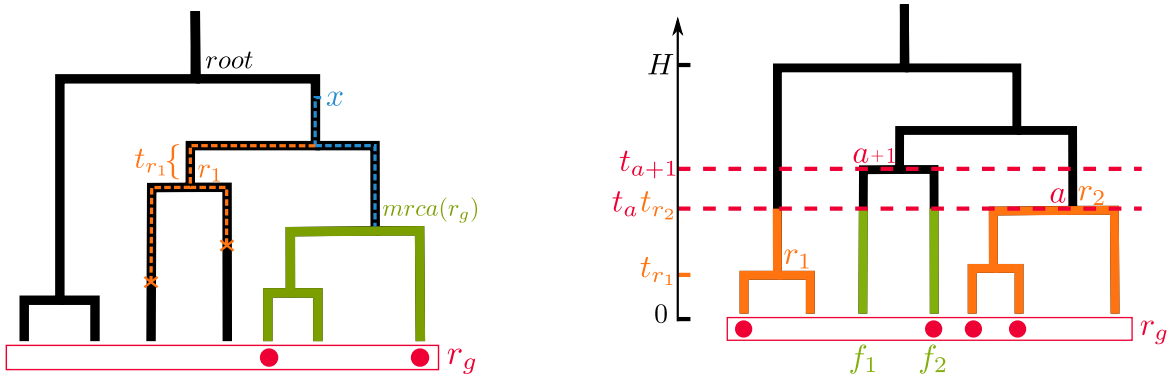
$$
\begin{aligned}
&\mathbb{P}_\Theta(r_g \,|\, \mathrm{cat}_g = 1, g \notin root) \\
&\quad = \frac{1}{i_1 L} \int_{mrca(r_g)}^{root} \mathbb{P}_x^{(1)}(r_g)\, i_1 \mathrm{d}x \\
&\quad = \frac{1}{L} \int_{mrca(r_g)}^{root} p_{1\to1}^{(1)}(x, mrca(r_g)) q_1(x, mrca(r_g)) \mathbb{P}_{mrca(r_g)}^{(1)}(r_g)\, \mathrm{d}x \\
&\quad = \frac{1}{L} \mathbb{P}_{mrca(r_g)}^{(1)}(r_g) \int_{mrca(r_g)}^{root} e^{-l_1(x - mrca(r_g))} \prod_{\substack{T_i \text{ subtree between} \\ x \text{ and } mrca(r_g) \\ \text{with root } r_i}} \left((1 - e^{-l_1 t_{r_i}}) + e^{-l_1 t_{r_i}} \mathbb{P}_{r_i}^{(1)} \left(0_{|T_i} \right) \right) \mathrm{d}x
\end{aligned}
\tag{12}
$$

Here, the first equality integrates over all possible positions for the appearance of the focal gene on the tree, i.e. on the path between $mrca(r_g)$ and $root$. L is the total tree length, and $\mathbb{P}_x^{(1)}(r_g)$ is the probability to observe pattern r_g knowing that gene g was present at position x in the tree.
The second equality decomposes $\mathbb{P}_x^{(1)}(r_g)$ into the probability that g survive between its arrival at position x and $mrca(r_g)$ (denoted by $p_{1\to1}^{(1)}(x, mrca(r_g))$), times the probability that g was lost in all subtrees branching on the path between x and $mrca(r_g)$ (denoted by $q_1(x, mrca(r_g))$), times the probability to observe pattern r_g given that g was present at $mrca(r_g)$. See Figure 8a for a graphical explanation.
The third equality is obtained by factoring out $\mathbb{P}_{mrca(r_g)}^{(1)}(r_g)$ and writing the expressions for $p_{1\to1}$ and q_1. $q_1(x, mrca(r_g))$ is the product for each subtree T_i branching between x and $mrca(r_g)$ of the probability that g is lost before the root r_i of T_i, plus the probability that g survive until r_i and that the null pattern (denoted by 0) is observed at the leaves of T_i. $\mathbb{P}_{mrca(r_g)}^{(1)}(r_g)$ and $\mathbb{P}_{r_i}^{(1)}(0_{|T_i})$ can be computed using Felsenstein's pruning algorithm.

It follows that:

$$
\begin{aligned}
&\mathbb{P}_\Theta(\mathrm{cat}_g = 1, g \notin root \,|\, g \text{ observed}) \mathbb{P}_\Theta(r_g \,|\, \mathrm{cat}_g = 1, g \notin root, g \text{ observed}) \\
&\quad = \frac{I_1^{obs}}{N^{obs}} \times \mathbb{P}_\Theta(r_g \,|\, \mathrm{cat}_g = 1, g \notin root, g \text{ observed}) \\
&\quad = \frac{\int_{\mathcal{T}} i_1 p_1^{obs}(x)\, \mathrm{d}x}{N^{obs}} \times \frac{\mathbb{P}_\Theta(r_g \,|\, \mathrm{cat}_g = 1, g \notin root)}{\frac{1}{L} \int_{\mathcal{T}} p_1^{obs}(x)\, \mathrm{d}x} \\
&\quad = \frac{i_1}{N^{obs}} \mathbb{P}_{mrca(r_g)}^{(1)}(r_g) \int_{mrca(r_g)}^{root} e^{-l_1(x - mrca(r_g))} \prod_{\substack{T_i \text{ subtree between} \\ x \text{ and } mrca(r_g) \\ \text{with root } r_i}} \left((1 - e^{-l_1 t_{r_i}}) + e^{-l_1 t_{r_i}} \mathbb{P}_{r_i}^{(1)} \left(0_{|T_i} \right) \right) \mathrm{d}x
\end{aligned}
\tag{13}
$$

To obtain the second equality, we used the approximation of I_1^{obs} by its expectation as computed in equation (6), and the fact that $\mathbb{P}_\Theta(g \text{ observed} \,|\, \mathrm{cat}_g = 1, g \notin root) = \frac{1}{L} \int_{\mathcal{T}} p_1^{obs}(x)\, \mathrm{d}x$.



(a) Likelihood for an imported Private gene

(b) Likelihood of a Mobile gene

Figure 8: Schematic explanation of how to compute the likelihood of imported genes. (a) For imported Private genes we have to compute the probability to observe pattern r_g given that g was present at $mrca(r_g)$ (in green), times the integral over x of the probability to survive between x and $mrca(r_g)$ (in blue) times the probability to be lost in all subtrees branching between x and $mrca(r_g)$ (in orange). See equation (12). (b) For Mobile genes we have to sum over all time intervals of the tree (i.e., intervals between 2 nodes) the product of the probabilities to observe pattern r_g restricted to subtrees created by cutting the tree at this interval (in orange), times the product of probabilities to observe pattern r_g restricted to single leaves (in green). See equation (16).

Likelihood for Mobile genes

For a Mobile gene, we have to integrate over the tree height for possible appearance times in the gene pool of the population. Once the gene is present in the gene pool, it evolves following a 2-state Markov model (0: absent and 1: present) with transition rates g_2 and l_2. The rate matrix is:

$$Q = \begin{pmatrix} -g_2 & g_2 \\ l_2 & -l_2 \end{pmatrix} \tag{14}$$

Thus the transition matrix for a time duration of t is:

$$P(t) = \begin{pmatrix} p^{(2)}_{0\to 0}(t) & p^{(2)}_{0\to 1}(t) \\ p^{(2)}_{1\to 0}(t) & p^{(2)}_{1\to 1}(t) \end{pmatrix} = e^{Qt} = \begin{pmatrix} \frac{l_2}{g_2+l_2} + \frac{g_2}{g_2+l_2}e^{-(g_2+l_2)t} & \frac{g_2}{g_2+l_2} - \frac{g_2}{g_2+l_2}e^{-(g_2+l_2)t} \\ \frac{l_2}{g_2+l_2} - \frac{l_2}{g_2+l_2}e^{-(g_2+l_2)t} & \frac{g_2}{g_2+l_2} + \frac{l_2}{g_2+l_2}e^{-(g_2+l_2)t} \end{pmatrix} \tag{15}$$

In the following time flows backward, i.e. greater time is further away in the past.

$$
\begin{aligned}
&\mathbb{P}_\Theta(r_g \,|\, \mathrm{cat}_g = 2) \\
&\quad = \int_{t=0}^{H} \frac{1}{H} \mathbb{P}_t^{(2)}(r_g)\, \mathrm{d}t \\
&\quad = \frac{1}{H} \sum_{a=1}^{2G-2} \int_{t=t_a}^{t_{a+1}} \mathbb{P}_t^{(2)}\left(r_g\right)\, \mathrm{d}t \\
&\quad = \frac{1}{H} \sum_{a=G}^{2G-2} \int_{t=t_a}^{t_{a+1}} \prod_{T_i} \mathbb{P}_t^{(2)}\left(r_g|_{T_i}\right)\, \mathrm{d}t \\
&\quad = \frac{1}{H} \sum_{a=G}^{2G-2} \int_{t=t_a}^{t_{a+1}} \prod_{\substack{T_i \text{ subtree} \\ \text{with root } r_i}} \left(p_{0\to1}^{(2)}(t-t_{r_i})\mathbb{P}_{r_i}^{(2)}\left(r_g|_{T_i}\right) + p_{0\to0}^{(2)}(t-t_{r_i})\mathbb{P}_{\bar{r_i}}^{(2)}\left(r_g|_{T_i}\right)\right) \\
&\qquad\qquad \times \prod_{f_i \text{ leaf}} p_{0\to r_g|_{f_i}}^{(2)}(t-t_{f_i})\, \mathrm{d}t
\end{aligned}
\tag{16}
$$

The first equality integrates over all possible heights where the gene was imported in the gene pool of the population. $\mathbb{P}_t^{(2)}(r_g)$ is the probability to obtain pattern r_g given that g arrived at height t of the tree.

The second equality decomposes the integral, summing the probability that gene g arrived between node a and node $a+1$ over all nodes of the tree. We assume here that nodes are ordered by increasing height, i.e. $t_1 = 0$ and $t_{2G-1} = H$. As $\mathcal{T}$ is ultrametric, we have in fact $t_1 = t_2 = ... = t_G = 0$.

In the third equality, $\mathbb{P}_t^{(2)}(r_g)$ is decomposed into a product over all subtrees obtained by cutting the phylogenetic tree at time t (the time of appearance of the gene).

In the fourth equality, we distinguish between subtrees that are a single leaf (second product) and other subtrees (first product), see Figure 8b. In the first product, the sum represents the fact that the gene can either be gained by the focal lineage from time t to t_{r_i} (with probability $p_{0\to1}^{(2)}(t)$), or not be gained (with probability $p_{0\to0}^{(2)}(t)$). $\mathbb{P}_{r_i}^{(2)}(r_g|_{T_i})$ is the probability to obtain pattern r_g restricted to the leaves of T_i given that g was present at root r_i of subtree T_i, while $\mathbb{P}_{\bar{r_i}}^{(2)}(r_g|_{T_i})$ is the probability to obtain this pattern given that is was absent at root r_i. These two quantities can be computed using Felsenstein's pruning algorithm. $p_{i\to j}^{(2)}(t)$ is the probability for a Mobile gene to switch from state i to state j during a time interval of length t, with $i, j \in \{0, 1\}$. From (15) we have:

$$
\begin{aligned}
p_{0\to1}^{(2)}(t) &= \frac{g_2}{g_2+l_2} - \frac{g_2}{g_2+l_2} e^{-(g_2+l_2)t} \\
p_{0\to0}^{(2)}(t) &= \frac{l_2}{g_2+l_2} + \frac{g_2}{g_2+l_2} e^{-(g_2+l_2)t} \\
p_{0\to r_g|_{f_i}}^{(2)}(t) &= \begin{cases} p_{0\to1}^{(2)}(t)(1-\epsilon_2) + p_{0\to0}^{(2)}(t)\epsilon_2, & \text{if } r_g|_{f_i} = 1 \\ p_{0\to1}^{(2)}(t)\epsilon_2 + p_{0\to0}^{(2)}(t)(1-\epsilon_2), & \text{if } r_g|_{f_i} = 0 \end{cases}
\end{aligned}
\tag{17}
$$

with ϵ_2 the rate of false positives and false negatives for Mobile genes. It follows that:

$$\begin{aligned}
&\mathbb{P}_\Theta(\text{cat}_g = 2 \,|\, g \text{ observed})\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 2, g \text{ observed}) \\
&= \frac{I_2^{obs}}{N^{obs}} \times \mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 2, g \text{ observed}) \\
&= \frac{\int_0^t i_2 p_1^{obs}(t)\,\mathrm{d}t}{N^{obs}} \times \frac{\mathbb{P}_\Theta(r_g \,|\, \text{cat}_g = 2)}{\frac{1}{H}\int_0^t p_2^{obs}(t)\,\mathrm{d}t} \\
&= \frac{i_2}{N^{obs}} \sum_{a=G}^{2G-2} \int_{t=t_a}^{t_{a+1}} \prod_{T_i \text{ subtree}} \left(p^{(2)}_{0\to 1}(t - t_{r_i}) \mathbb{P}^{(2)}_{r_i}\left(r_g|_{T_i} \right) + p^{(2)}_{0\to 0}(t - t_{r_i}) \mathbb{P}^{(2)}_{\bar{r}_i}\left(r_g|_{T_i} \right) \right) \\
&\qquad \times \prod_{f_i \text{ leaf}} p^{(2)}_{0\to r_g|_{f_i}}(t - t_{f_i})\,\mathrm{d}t
\end{aligned} \tag{18}$$

To obtain the second equality, we used the approximation of I_2^{obs} by its expectation as computed in equation (6), and the fact that $\mathbb{P}_\Theta(g \text{ observed} \,|\, \text{cat}_g = 2) = \frac{1}{H}\int_0^t p_2^{obs}(t)\,\mathrm{d}t$.

Inference method

Optimization method

In order to fit our model on the *S. enterica* dataset, we computed the Maximum Likelihood parameters of the model:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{L}(\text{data}; \Theta) \tag{19}$$

Substituting the expectations of equation (6) in equation (5), we observe that the 10 parameters ($N_0, l_0, i_1, l_1, i_2, g_2, l_2, \epsilon_0, \epsilon_1$ and ϵ_2) are linked by the following relation:

$$N^{obs} = p_0^{obs}(root)N_0 + \frac{i_1}{l_1} p_1^{obs}(root) + \int_{\mathcal{T}} i_1 p_1^{obs}(x)\,\mathrm{d}x + \int_{t=0}^{H} i_2 p_2^{obs}(t)\,\mathrm{d}t \tag{20}$$

(given that $p_0^{obs}(root)$ depends on l_0 and ϵ_0, $p_1^{obs}(root)$ and $p_1^{obs}(x)$ depend on l_1 and ϵ_1, and $p_2^{obs}(t)$ depends on g_2, l_2 and ϵ_2). N^{obs} being the total number of observed genes in the dataset, we know its value. This relation is a constraint that we have to put on parameters to ensure that the expected number of observed genes predicted by the model coincides with the number of actually observed genes. Hence, we have only 9 degrees of freedom in the model. For example, we can express i_2 as a function of other parameters:

$$i_2 = \frac{N^{obs} - p_0^{obs}(root)N_0 - \frac{i_1}{l_1} p_1^{obs}(root) - \int_{\mathcal{T}} i_1 p_1^{obs}(x)\,\mathrm{d}x}{\int_{t=0}^{H} p_2^{obs}(t)\,\mathrm{d}t} \tag{21}$$

We performed likelihood maximization over the 9 free parameters using the Nelder-Mead algorithm, which is a derivative-free optimization method. As this optimization procedure offers no guarantee to find the global maximum, we tried three different approaches to have some confidence in the result. We performed 5 optimizations with different, randomly chosen initial conditions and took the maximum over these 5 replicates. Then, we designed two methods to find a good starting point for the optimization. The first one consists in making 10 optimizations on subsamples consisting in 1/10 of the gene families, and then use the median of inferred parameters as starting point for the whole dataset. The second one consists in assigning genes a priori to a gene category (based only on the frequency and parsimony of the presence/absence pattern), performing the optimization separately for each category (thus, with fewer parameters) and using the separately inferred parameters as starting point for the complete optimization. Of course, this last method was only possible because we had gained some knowledge on the model and the data while studying previous versions of the model. All the above approaches yielded the same result, which make us confident to have found the global maximum. Optimizations that converged to the global maximum took from 6 to 15 hours to run on 50 cores, depending on the starting point. The time complexity is linear in the number of gene families and quadratic in the number of genomes. The inference tool is available at https://github.com/JasmineGamblin/PPMmodelPangenome. More details on the optimization method (including the boundaries imposed on parameters and the way to draw random starting points) are given in the Supplementary Information.

Category assignment

Once the MLE of parameters was computed, we assigned each gene to the category with highest posterior probability, i.e. gene g with pattern r_g was assigned to category c such that:

$$c = \underset{c=0,1,2}{\operatorname{argmax}}\, \mathbb{P}_{\hat{\Theta}}(\mathrm{cat}_g = c \,|\, r_g) \tag{22}$$

Expected number of gains for a Mobile gene

Let G_g be the number of introductions for a Mobile gene g. By 'introduction', we mean a transfer between the gene pool and a lineage of the tree where this gene has never been present. We can compute G_g as follows, integrating over all possible arrival times in the gene pool and summing over

all possible branches where an introduction could happen:

$$
\begin{aligned}
&\mathbb{E}(G_g \,|\, \mathrm{cat}_g = 2) \\
&\qquad = \frac{1}{H} \int_0^H \mathbb{E}(G_g \,|\, T_g = t)\, \mathrm{d}t \\
&\qquad = \frac{1}{H} \int_{t=0}^H \sum_{\substack{T_i \text{ subtree} \\ \text{below } t}} \sum_{\substack{(b,c) \\ \text{branch of } T_i}} \mathbb{E}(\mathbb{1}_{g \text{ introduced in } (b,c)})\, \mathrm{d}t \\
&\qquad = \frac{1}{H} \int_0^H \sum_{\substack{T_i \text{ subtree} \\ \text{below } t}} \sum_{\substack{(b,c) \\ \text{branch of } T_i}} \int_{t_c}^{t_b} g_2 e^{-g_2(t-u)}\, \mathrm{d}u\, \mathrm{d}t
\end{aligned}
\tag{23}
$$

Conditionally on g having arrived at time t in the pool, the probability that it is introduced on a branch (b, c) below this time t is $\int_{t_c}^{t_b} g_2 e^{-g_2(t-u)}\, \mathrm{d}u$. Indeed, $g_2\, \mathrm{d}u$ is the probability that g is gained during time interval $\mathrm{d}u$ and $e^{-g_2(t-u)}$ is the probability that no gain happened before time u.

Acknowledgements

All authors are grateful to the Center for Interdisciplinary Research in Biology (CIRB, Collège de France) for funding. JG's PhD scholarship was funded by the French Ministry of Research (MESRI) and by the Fondation pour la Recherche Médicale (FRM, scholarship n°FDT202304016561). The authors thank the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing help and computing resources. They also thank Marie Touchon and Bastien Boussau for insightful comments about this work.

Data availability

The inference tool for the PPM model and the code used to run simulations is available at https://github.com/JasmineGamblin/PPMmodelPangenome. Genomic data and softwares used to analyse it are referenced in the Methods section.

References

Arimizu, Y., Kirino, Y., Sato, M. P., Uno, K., Sato, T., Gotoh, Y., Auvray, F., Brugere, H., Oswald, E., ... Ogura, Y. (2019). Large-scale genome analysis of bovine commensal Escherichia coli reveals that bovine-adapted E. coli lineages are serving as evolutionary sources of the emergence of human intestinal pathogenic strains. *Genome Research*, *29*(9), 1495–1505.

Bansal, M. S., Kellis, M., Kordi, M., & Kundu, S. (2018). RANGER-DTL 2.0: Rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, *34*(18), 3214–3216.

Baumdicker, F., Hess, W. R., & Pfaffelhuber, P. (2012). The Infinitely Many Genes Model for the Distributed Genome of Bacteria. *Genome Biology and Evolution*, *4*(4), 443–456.

Baumdicker, F., & Kupczok, A. (2023). Tackling the Pangenome Dilemma Requires the Concerted Analysis of Multiple Population Genetic Processes. *Genome Biology and Evolution*, *15*(5), evad067.

Beavan, A., Domingo-Sananes, M. R., & McInerney, J. O. (2024). Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. *Proceedings of the National Academy of Sciences*, *121*(1), e2304934120.

Bolotin, E., & Hershberg, R. (2015). Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biology and Evolution*, *7*(8), 2173–2187.

Botelho, J., Tüffers, L., Fuss, J., Buchholz, F., Utpatel, C., Klockgether, J., Niemann, S., Tümmler, B., & Schulenburg, H. (2023). Phylogroup-specific variation shapes the clustering of antimicrobial resistance genes and defence systems across regions of genome plasticity in *Pseudomonas aeruginosa*. *eBioMedicine*, *90*, 104532.

Cohen, O., & Pupko, T. (2010). Inference and Characterization of Horizontally Transferred Gene Families Using Stochastic Mapping. *Molecular Biology and Evolution*, *27*(3), 703–713.

Cummins, E. A., Hall, R. A., Chris, C., McInerney, J. O., & McNally, A. (2022). Distinct evolutionary trajectories in the Escherichia coli pangenome occur within sequence types. *Microbial genomics*, *8*(11).

D'Andrea, M. M., Arena, F., Pallecchi, L., & Rossolini, G. M. (2013). CTX-M-type β-lactamases: A successful story of antibiotic resistance. *International Journal of Medical Microbiology*, *303*(6), 305–317.

Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., & Hahn, M. W. (2014). Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. *PLOS Computational Biology*, *10*(12), e1003998.

Didelot, X., Darling, A. E., & Falush, D. (2008). Inferring genomic flux in bacteria. *Genome Research*, *19*(2), 306–317.

Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., Fourment, M., & Holmes, E. C. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, *2*(11), e000094.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376.

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281.

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., … Vallenet, D. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLOS Computational Biology*, *16*(3), e1007732.

Haegeman, B., & Weitz, J. S. (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics*, *13*(1), 196–196.

Huson, D. H., & Steel, M. (2004). Phylogenetic trees based on gene content. *Bioinformatics*, *20*(13), 2044–2049.

Jacox, E., Chauve, C., Szöllosi, G. J., Ponty, Y., & Scornavacca, C. (2016). ecceTERA: Comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, *32*(13), 2056–2058.

Kloub, L., Gosselin, S., Fullmer, M. S., Graf, J., Gogarten, J. P., Gogarten, J. P., & Bansal, M. S. (2021). Systematic detection of large scale multi gene horizontal transfer in prokaryotes. *Molecular Biology and Evolution*, *38*(6), 2639–2659.

Lobkovsky, A. E., Wolf, Y. I., & Koonin, E. V. (2013). Gene Frequency Distributions Reject a Neutral Model of Genome Evolution. *Genome Biology and Evolution*, *5*(1), 233–242.

McInerney, J. O. (2022). Prokaryotic Pangenomes Act as Evolving Ecosystems. *Molecular Biology and Evolution*.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534.

Oliveira, P. H., Touchon, M., Cury, J., & Rocha, E. P. C. (2017). The chromosomal organization of horizontal gene transfer in bacteria. *Nature Communications*, *8*(1), 841.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693.

Pagel, M. (1997). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *255*(1342), 37–45.

Perrin, A., & Rocha, E. P. C. (2021). PanACoTA: A modular tool for massive microbial comparative genomics. *NAR Genomics and Bioinformatics*, *3*(1), lqaa106.

Piel, D., Bruto, M., Labreuche, Y., Blanquart, F., Chenivesse, S., Lepanse, S., James, A., Barcia-Cruz, R., Dubert, J., … Le Roux, F. (2021). Genetic determinism of phage-bacteria coevolution in natural populations. *bioRxiv*.

Sakoparnig, T., Field, C., & van Nimwegen, E. (2021). Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species (A. Nourmohammad & A. M. Walczak, Eds.). *eLife*, *10*, e65366.

Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. *Genome Biology*, *20*(1), 92.

Seif, Y., Kavvas, E., Lachance, J.-C., Yurkovich, J. T., Nuccio, S.-P., Fang, X., Catoiu, E., Raffatellu, M., Palsson, B. O., & Monk, J. M. (2018). Genome-scale metabolic reconstructions of multiple Salmonella strains reveal serovar-specific metabolic traits. *Nature Communications*, *9*(1), 3771.

Szöllosi, G. J., Boussau, B., Abby, S. S., Tannier, E., & Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(43), 17513–17518.

Szöllosi, G. J., Tannier, E., Daubin, V., & Boussau, B. (2015). The Inference of Gene Trees with Species Trees. *Systematic Biology*, *64*(1).

Szöllosi, G. J., Tannier, E., Lartillot, N., & Daubin, V. (2013). Lateral Gene Transfer from the Dead. *Systematic Biology*, *62*(3), 386–397.

Tenaillon, O., Skurnik, D., Picard, B., & Denamur, E. (2010). The population genetics of commensal Escherichia coli. *Nature Reviews Microbiology*, *8*(3), 207–217.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., ... Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955.

Tettelin, H., & Medini, D. (Eds.). (2020). *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing.

To, T.-H., Jung, M., Lycett, S., & Gascuel, O. (2016). Fast Dating Using Least-Squares Criteria and Algorithms. *Systematic Biology*, *65*(1), 82–97.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., ... Parkhill, J. (2020). Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology*, *21*(1), 180.

Touchon, M., Cury, J., Yoon, E.-J., Krizova, L., Cerqueira, G. C., Murphy, C., Feldgarden, M., Wortman, J., Clermont, D., ... Rocha, E. P. (2014). The Genomic Diversification of the Whole Acinetobacter Genus: Origins, Mechanisms, and Consequences. *Genome Biology and Evolution*, *6*(10), 2866–2882.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., ... Denamur, E. (2009). Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics*, *5*(1), e1000344.

Wagner, C., & Hensel, M. (2011). Adhesive Mechanisms of Salmonella enterica. In D. Linke & A. Goldman (Eds.), *Bacterial Adhesion: Chemistry, Biology and Physics* (pp. 17–34). Springer Netherlands.

Whelan, F. J., Hall, R. J., & McInerney, J. O. (2021). Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Molecular Biology and Evolution*, *38*(9), 3697–3708.

Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E., & Koonin, E. V. (2016). Two fundamentally different classes of microbial genes. *Nature Microbiology*, *2*(3), 1–6.

Zamani-Dahaj, S. A., Okasha, M., Kosakowski, J., & Higgs, P. (2016). Estimating the Frequency of Horizontal Gene Transfer Using Phylogenetic Models of Gene Gain and Loss. *Molecular Biology and Evolution*, *33*(7), 1843–1857.