Bacterial Microevolution and the Pangenome

Florent Lassalle and Xavier Didelot

Abstract The comparison of multiple genome sequences sampled from a bacterial population reveals considerable diversity in both the core and the accessory parts of the pangenome. This diversity can be analysed in terms of microevolutionary events that took place since the genomes shared a common ancestor, especially deletion, duplication, and recombination. We review the basic modelling ingredients used implicitly or explicitly when performing such a pangenome analysis. In particular, we describe a basic neutral phylogenetic framework of bacterial pangenome micro-evolution, which is not incompatible with evaluating the role of natural selection. We survey the different ways in which pangenome data is summarised in order to be included in microevolutionary models, as well as the main methodological approaches that have been proposed to reconstruct pangenome microevolutionary history.

Keywords Pangenome · Bacterial microevolution · Evolutionary model · Recombination · Duplication · Deletion · Gene content · Ancestral state reconstruction · Reconciliation

1 Atomic Events in Bacterial Microevolution

Bacterial microevolution is the study of the evolutionary forces that shape the genetic diversity of a natural population of bacteria. This evolutionary process takes place as a result of the genetic changes happening within each of the genomes of the bacterial cells constituting the population. Over time, these changes are amplified or weakened by the effects of both genetic drift and natural selection.

F. Lassalle
Department of Infectious Disease Epidemiology, Imperial College London, London, UK

X. Didelot (✉)
School of Life Sciences and Department of Statistics, University of Warwick, Coventry, UK
e-mail: xavier.didelot@warwick.ac.uk

129

Genetic drift represents the evolution caused by the death and birth of cells in the bacterial population, and it acts at random on all genetic variants (Charlesworth 2009). The effects of genetic drift are higher when the population size is small, and so it could be thought given the large number of cells in bacterial populations that genetic drift would be weak. However, bacterial populations sometimes go through punctual bottlenecks during which genetic drift has a large effect, for example during transmission of pathogens from one host to another (Didelot et al. 2016). It is also believed that the strong structure of bacterial habitat, sometimes at the microscale can lead to much smaller effective population sizes than intuition suggests (Vos et al. 2013). Natural selection on the other hand acts in a nonrandom fashion, amplifying some variations and suppressing others, and is a very potent evolutionary force in shaping the diversity of bacterial species (Petersen et al. 2007; Buckee et al. 2008; Pepperell et al. 2013).

The genetic changes occurring on a single bacterial cell can be classified into mutation and recombination events, and the events of interest differ whether the focus is on the core genome (the regions shared by all genomes in the population) or the accessory genome (the regions that are found in some but not all of the genomes). As far as the core genome is concerned, the main type of mutation is the point mutation, whereby a single nucleotide is replaced, and the main type of recombination is called homologous recombination, in which a relatively short fragment of the genome is replaced with a homologous fragment coming from another bacterial cell (Didelot and Maiden 2010). There are three biological mechanisms that can lead to homologous recombination, namely conjugation (where two bacterial cells come in contact so that DNA can be transmitted from donor to recipient), transduction (where a phage acts as vector from donor to recipient) and transformation (where naked DNA is picked up by the recipient from the environment, possibly following the death of the donor cell) (Thomas and Nielsen 2005). But since their outcomes are hard to distinguish this diversity of mechanisms is usually ignored in evolutionary models of homologous recombination.

Point mutation and homologous recombination events clearly act on the evolution of the accessory genome in the same way as they do for the core genome. However, they do not change the genetic content of core genomes. There are two types of endogenous mutations that can alter the genetic content of a genome, duplication and deletion, and they can be thought of as opposite forces, with the former increasing the number of copies of a gene by one and the latter decreasing it by one. Finally, the accessory genome is also subject to non-homologous recombination, where a bacterial cell imports a DNA fragment from another cell and inserts it in its genome, without overwriting a previously existing homologous fragment (Ochman et al. 2000). Non-homologous recombination is often called lateral gene transfer or horizontal gene transfer, and in this chapter we will be using these three terms interchangeably. It should be noted, however, that this terminology is not always consistently used in the literature, with some authors using the term horizontal gene transfer to refer to both homologous and non-homologous recombination.

The three biological mechanisms mentioned above for homologous recombination (conjugation, transduction, and transformation) can lead to non-homologous recombination and once again, it is helpful when studying the bacterial
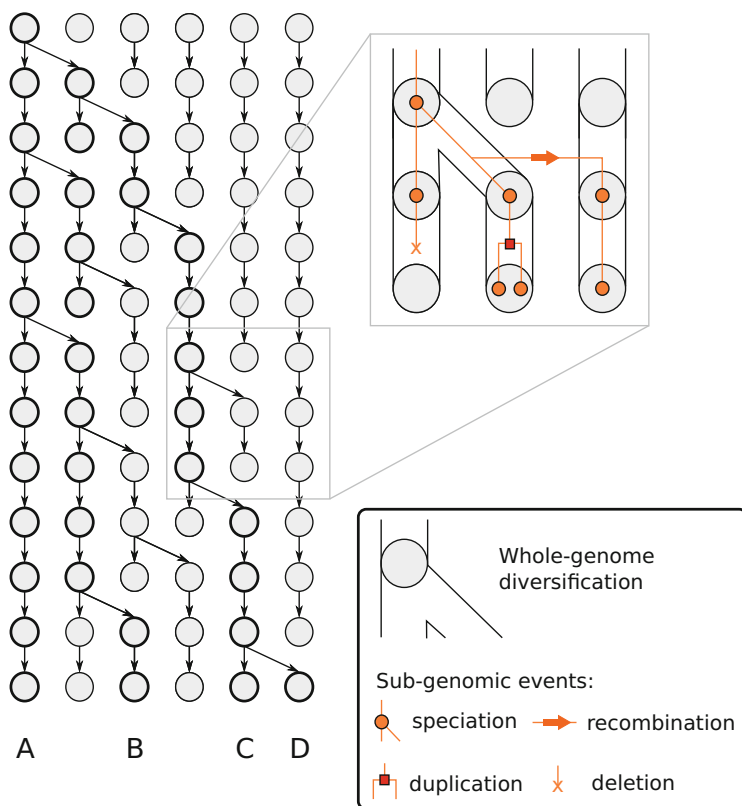
microevolution of the accessory genome to set aside the mechanism at play. Likewise, genetic duplication and deletion can have multiple causes that we will not explore. It should in fact be noted that even though it is useful to present and study them as separate, the atomic evolutionary events briefly described above are not biologically independent (Lawrence 1999; Everitt et al. 2014; Oliveira et al. 2017). For example, a single event of recombination could involve the replacement of some genes (homologous recombination), the insertion of new genes (non-homologous recombination) and the loss of some other genes (genetic deletion).

Furthermore, non-homologous recombination can sometimes be duplicative if the newly imported material is homologous to a sequence found somewhere else in the genome. In this case, the number of copies of the genes concerned is increased by one, as in a duplication event. The evolutionary distance between donor and recipient of such a non-homologous recombination event is then a crucial factor: if this distance is small the effect is similar to a duplication event, which can be seen as a transfer event where recipient and donor are the same organism. If distance between organisms is high, then the difference between the newly imported copy and the copy already present will likely be high too, providing a clear sign that duplication was not involved. This situation is analogous to the detection of homologous recombination in the core genome, where events from a closely related source do not leave a trace, or perhaps involve just a single substitution in which case they are undistinguishable from point mutation (Didelot et al. 2010).

2 Neutral Phylogenetic Framework of Bacterial Pangenome Microevolution

2.1 Challenges with a Comprehensive Model

The microevolutionary events that act on the bacterial pangenome, as briefly described above, can be combined into an evolutionary model of how the pangenome evolves over time. Let us consider a comprehensive model, which would account for the whole population of bacterial cells, including the fact that cells die and reproduce over time (so that genetic drift is included) and that various selective pressures are exerted. In this model, the genome of each cell is affected by various mutation and recombination events, all of which happens at a certain rates over time for each cell. All the rates involved in this model (birth and death of individuals, selection for specific variants and various evolutionary events) would not be assumed to be constant, but would be allowed to vary over time. This model falls in the class of forward-in-time models, due to the fact that it considers evolution as it unfolds over time, and famous examples of such models in the general population genetics literature are the Wright–Fisher model (Fisher 1931; Wright 1931) and the Moran model (Moran 1958). Figure 1 illustrates such a forward-in-time model of pangenome evolution.

Fig. 1 Illustration of the forward-in-time evolution of a bacterial population and its pangenome. At each time step, an individual is removed and another gives birth as in a standard Moran model. Furthermore, at each time step the accessory genome of each individual may evolve via deletion (orange cross), duplication (red square) and recombination (orange arrow)

The idea of this comprehensive model is to replicate exactly the processes that we know occur in nature, so that it is of the highest possible realism. However, a comprehensive model would also integrate the diversification process of the whole community of microorganism found at a given spot, with the impact of their biotic interactions and genetic exchange, but most importantly, of the competitive process leading to natural selection of the fittest. Such level of description of natural processes would render the model impossible to use, and that is why it is not found in the literature. It is educative though to ask ourselves why this model is unusable, as this will guide us towards more practical models that feature some of these ideal properties.

The first problem with this comprehensive model is computational: it would require very large amounts of computer memory to store the state of a population at a single time point, even much more so to track its evolution over time, and an equally impossible amount of computer power to consider the evolutionary events

happening to all members of the population. But even more importantly, there is statistical problem with the comprehensive model, in the sense that there are too many unknown quantities involved, for which we would not be able to take even a very rough guess at what their value might be. Therefore, even if the computational problems could be overcome, and analysis conducted under the comprehensive model, the results would be worthless since the quantities to be estimated would be unidentifiable. Simplifications will therefore have to be made to reach a model that has practical use, with the best model being not the most comprehensive one, but the one that achieves the best trade-off between biological realism on one hand, and computational and statistical considerations on the other hand.

Beyond the degree of complexity of a model and the search for a trade-off between computational efficiency and model realism, models may rely on different conceptual formalisation of bacterial genomes and their evolutionary process. These different concepts will generate different approaches and methods that are in general complementary. We will thus present different elements of phylogenetic models of pangenome evolution, which flavours may be combined to provide a variety of practical models.

2.2 Analysing Selection Based on Neutral Models

Perhaps the greatest challenge posed by the comprehensive model above would be its attempt at encompassing the role of natural selection. As previously mentioned, it is clear that natural selection plays a crucial role in shaping the microevolution of the pangenomes of bacterial populations, but the effect of this force is different for all genes or nucleotidic site and their allelic variants, may vary significantly over time, and be different for different segments of the populations, for example if some lineages are adapted to a certain environment. Such adaptation of a lineage will involve many traits distributed in the pangenome of that population, and new mutation arising in this background might interact with it; this leads to complex epistatic (i.e. non-additive fitness) interactions between genomic traits, affecting the probability of selecting new genetic variants in one or another genomic background—a process that could add infinite degrees of complexity to the exhaustive model. Model design can, therefore, be greatly simplified by considering no effect of selection, or in other words neutrality of evolution.

Even if the role of selection is not explicitly included in a model, it does not mean that analyses based on this model are completely uninformative about selection. Neutral evolutionary models provide a framework to search for evidence of natural selection. This can be achieved formally by contrasting observed patterns in compared genomic data to expectations under neutral models. Another approach to detect selection is to fit a neutral model to genomic data, having heterogeneous parameters to describe the evolution process of each species lineage and/or gene family; outlier species or genes with 'abnormally' high or low parameters can provide a clue to non-neutral processes taking place. Similarly, the identification

of historical changes of processes (e.g. acceleration or slowing down of diversification rates) in the scenario of pangenome evolution can provide strong clues of selection affecting the species lineage or gene under focus (Boussau et al. 2004; David and Alm 2011; Lassalle et al. 2017).

This approach is similar to the way that the role of selection is being investigated in the core genome. In this more frequently explored setting, a typical pipeline (Hedge and Wilson 2016) involves reconstructing a phylogenetic tree, classifying substitutions in terms of whether they are synonymous or not and estimating evolutionary rates so that selection tests can be applied based on variations in the rates of synonymous and non-synonymous substitutions along the genome (Wilson and McVean 2006; Castillo-Ramírez et al. 2011) or between populations (McDonald and Kreitman 1991; Vos 2011). In this popular approach, the evolutionary models used to build the phylogenetic tree and reconstruct substitution events are purely neutral, but still lead to invaluable insights into the natural selection process.

2.3 Phylogenetic Approach

A neutral version of the comprehensive model described above would still be impossible to use in practice. A major difficulty is that it considers the evolution of the population forward-in-time, so that every single cell in the population has to be included. However, any dataset we may have available for analysis will only include a small fraction of the population, sampled typically at a single time point (or a few recent time points in the best-case scenario). However, considering the evolution of the whole population over time can appear wasteful, since most cells in the past would not have had any descendants surviving in the present-time sample. A much more tractable approach is therefore to only consider the genealogical process of the sampled genomes, which is a backward-in-time process. Under relatively mild assumptions, and without introducing too much approximation, this genealogical process can be described without reference to the whole forward-in-time process. In particular, the coalescent model (Kingman 1982) describes the genealogical process of a population following either the Wright–Fisher or the Moran model of forward-in-time evolution. Extensions of the basic coalescent model have been derived to deal with fluctuating population size (Griffiths and Tavare 1994), homologous recombination (Griffiths and Marjoram 1997), which for bacteria is analogous to gene conversion (Wiuf and Hein 2000), and many other forms of relaxation of the assumptions ruling the evolutionary process (Donnelly and Tavare 1995; Nordborg 2001; Rosenberg and Nordborg 2002).

Considering this genealogical process, and the ability to reconstruct it with relatively high accuracy from genome sequences, is pivotal to lead to a usable model of pangenome evolution. A simple approach is to focus on core genome elements and apply a standard phylogenetic method typically based on maximum likelihood or Bayesian inference under an evolutionary model of neutral point mutations (Yang and Rannala 2012). Bacteria reproduce clonally and most species

recombine relatively rarely (Vos and Didelot 2009; Yang et al. 2018), so that this simple approach can often be sufficient for our purposes. Phylogenetic methods have also been developed that can account for the effect of homologous recombination while still reconstructing a single tree (Didelot and Wilson 2015; Croucher et al. 2015). Methods that attempt to reconstruct a graph of ancestry rather than a single tree are superior in principle, but rarely used in practice due to their high computational cost (Didelot et al. 2010; Vaughan et al. 2017).

In the context of a phylogenetic tree reconstructed from the core genome, we can consider the events of duplication, deletion and non-homologous recombination that shape the accessory genome. These events happen on the branches of the phylogeny at certain rates that may vary over time and lineages. Notwithstanding such remaining complexities, a phylogenetic model of bacterial pangenome microevolution represents a practical approach relative to the comprehensive forward-in-time model. The events that affect the accessory genome are relatively rare, which results in a strong phylogenetic inertia of genome gene contents, i.e. a large correlation between gene contents and core genome-based diversity (Konstantinidis et al. 2006; Kislyuk et al. 2011). Ignoring this effect would lead to strong misinterpretation of gene distribution patterns, especially in case of a diversity bias in genome sampling, e.g. when surveying a pathogen epidemics where clusters of closely related strains occur. Modelling the pangenome evolution within a phylogenetic framework where evolution of the gene content takes place along the genealogical tree avoids such pitfalls, in the same way as a phylogenetic framework avoids false conclusions to be reached when performing bacterial genome-wide association mapping (Collins and Didelot 2018).

3 Description of Pangenome Data for Inclusion in Microevolutionary Models

3.1 Units of Pangenome Evolution

In order to describe further the existing models of pangenome microevolution, it is necessary to consider the unit in which the pangenome is being described. Figure 2 illustrates the different approaches that have been used for that purpose. The ideal starting point would be a complete sequence of each genome of interest, but this is rarely available due to repeat regions in the genomes that obscure the exact ordering of sequences along the genomes, at least based on short read sequencing. For that reason, the most frequently used data is a de novo assembly of each genome, which can be performed, for example using Velvet (Zerbino and Birney 2008) or SPAdes (Bankevich et al. 2012). This results a set of genomic regions called contigs, which occur in an unknown order either on chromosomes or on plasmids.

A first approach considers a genome alignment, where every part of a genome is assigned to a syntenic block—segments of genome sequences that are all
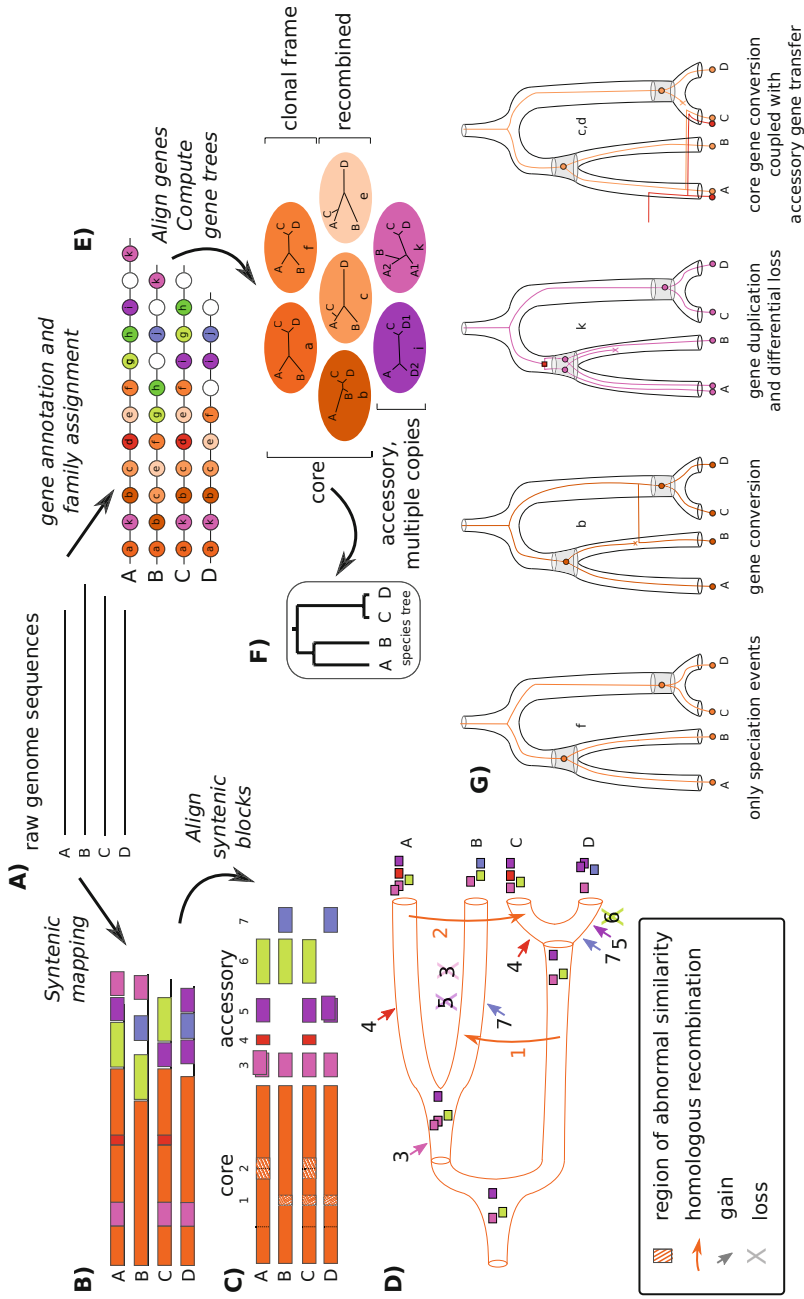
Fig. 2 (**a**) Datasets of raw genome sequence assemblies can be dealt with different approaches. (**b**) Whole-genome alignment of the compared genomes. Extensive homologous regions (syntenic blocks) are mapped in different colours. Orange indicate the core genome, the other colours indicate accessory syntenic

blocks; the absence of a colour block indicate genome-specific (orphan) sequence. (**c**) From there, core and accessory genomes can be treated separately. An ungapped core genome alignment is produced (left hand, orange blocks; dotted vertical lines indicate insertion point of accessory material). Regions of abnormal sequence similarity (with respect to genome-wide divergence) are identified (hatched boxes) as clues of homologous recombination events. Presence of accessory genomic blocks is counted in each genome, an information that can be easily stored in a numeric matrix. (**d**) The clonal genealogy (orange tubes) is computed from the core genome alignment purged of recombined regions. Homologous recombination events and gain/loss of accessory genome blocks can be inferred by confronting this tree to the information gathered in (**c**). (**e**) Another approach consists of annotating the genes present in the genomes (e.g. protein-coding sequences) and to classify them onto homologous gene families based on sequence similarity. (**f**) The occurrence pattern of each family allow to classify them as core (when a single copy of the gene is present in every genome) or accessory. Aligning gene family sequences allow to infer gene-specific phylogenies, or gene trees. Core gene trees can be used to generate a consensus tree representative of the core genome history (species tree). Differences of topologies between the species tree and the individual core gene trees reveal genes subject to homologous recombination. (**g**) Ancestral states and evolution scenarios can be reconstructed for each gene family, through reconciliation of the gene family tree with the species tree under a model with events of gene duplication, transfer and loss. Sample scenarios for families *f, b, k, c* and *d* are presented, with the reconciled gene tree (colour line) embedded in the species tree (tube); shaded areas indicate events of whole-genome diversification, where gene speciation events occur. Note that simpler gene gain/loss scenarios and ancestral gene content reconstruction can be reconstructed based on gene family occurrence data using the method depicted in (**d**)

homologous and can be aligned (Fig. 2b, c). These sequence segments can have boundaries falling anywhere in the genome, notably between or within protein-coding sequences, and the often span many genes. While this is a flexible view of pangenome evolution that is probably the most realistic—evolving genomes ignore human annotation of functional elements—it may be cumbersome to implement with a growing number of compared genome. Indeed, every genome added to the dataset may result in the breakage of a syntenic block into several parts due to insertion, deletions, or rearrangements in one of the homologous genome segments. Homologous genome segments can themselves be difficult to align at the nucleotide level when they include fast-evolving genome regions. For these reasons, even the best software for this task such as progressiveMauve (Darling et al. 2010) or MUGSY (Angiuoli and Salzberg 2011) can only deal with between 10 and 100 genomes, depending on how diverse they are. This first alignment approach works best on the well-conserved parts of pangenomes, i.e. the core genomes and possibly large conserved accessory regions of the genomes. This partial sampling of genome sequences is practical because it allows to represent the homology between genomes as a concatenated alignment of all these syntenic blocks, which amounts to a representative map of the genome. Alternatively, a representative whole-genome alignment can be obtained by mapping all homologous sequences in compared genomes to the genome of a reference isolate, using, for example MUMmer (Kurtz et al. 2004). This can result in reference-biased representation, which may be avoided by restricting the alignment to the core genome.

A second approach focuses on genes, or more specifically on families of homologous genes. These are usually defined based on sequence similarity and restricted to protein-coding sequences, even though it can be applied to conserved intergenic sequences as well (Fig. 2e). In this representation, rather than a reference whole-genome map, we consider independent gene families, which members need not be localised in a genome. The diversity of the gene family can conveniently be represented with a phylogenetic tree based on all nucleotide positions in the aligned genes, which allows the estimation of statistical support (Fig. 2f). This information can, in turn, be used to inform the ancestral reconstruction of genome gene content (as discussed below). There are several ways in which this gene family content identification can be performed. If a representative from each family is known in advance, similarity search tools like BLAST (Altschul et al. 1997) can be used to search them in each genome, and, for example BIGSdb automates this process (Jolley and Maiden 2010). Alternatively, each de novo assembled genome can be annotated separately, using, for example Prokka (Seemann 2014) or RAST (Aziz et al. 2008). Homologs can then be identified by using a combination of similarity search between genes from different genomes (e.g. with BLAST) and similarity network analysis, as implemented, for example in the software OrthoMCL (Li et al. 2003), with integrated pipelines implemented in software like Roary (Page et al. 2015) and MMseqs (Steinegger and Söding 2017).

A consequence of this distinction is the way genetic exchange between genomes is considered. In the nucleotide-centred vision, genetic exchange will result either in the replacement of a region (homologous recombination) or the insertion of a

sequence at a defined position in the genome map (non-homologous recombination) (Fig. 2c). Similarly, a genetic duplication event will consist in recopying a segment of genome sequence and inserting it next to its template (tandem duplication) or away from it. Homologous recombination events can be evidenced based on a scan of the genome map, looking for increased or decreased sequence similarity (or phylogenetic relatedness) between compared genomes along the genome map (Didelot et al. 2007). Non-homologous recombination and duplication events consist of insertion events and are simply evidenced by some region being only represented in some genomes in the alignment—the others featuring a large 'gap', or long string of missing characters. Distinguishing non-homologous recombination from duplication events can be tricky: even comparing the inserted segment to the rest of the genome and finding a similar region is not conclusive that it would be the duplication template (or copy) of the studied region. Such a pattern could also result from a recombination with a related organism leading to the insertion of genetic material that had homologous counterpart already residing in the recipient genome. Not finding a similar region is not conclusive of the insertion resulting from a non-homologous recombination event either, as an ancient duplication followed by a loss (or deletion) in the compared lineage may result in the same pattern. The answer to this conundrum is modelling of the possible sequences of events, or scenarios, and determine the most likely based on patterns of sequence divergence.

In the second approach centred on genes, the exchange of genetic material is made most evident in the phylogeny of genes, or gene tree, because the gene from the recipient will be more closely related to genes from the donor than to genes from closely related species. In this context, the event is rather called horizontal gene transfer, in opposition to vertical evolution, which would have resulted in the 'normal' clustering of genes from closely related species. Again, this representation ignores the locus where genes sit, and it is therefore not straightforward to know from the gene tree whether the horizontal gene transfer event resulted in the replacement of a resident sequence or in the addition of a new one.

There are also other evolving units that can be considered as the basis of pangenome microevolution modelling, including conserved protein domains or short sequences of a constant length, which are also known as words, features, or k-mers (Sims and Kim 2011; Sheppard et al. 2013b). Some units may seem more natural than others from a theoretical point of view, but in practice all units have pros and cons, and the choice of unit is guided by the evolutionary resolution required by each pangenome investigation.

3.2   Granularity of Homologous Groups

When modelling the pangenome diversity with homologous gene families, a further distinction can be done on which homologous link to consider clustering genes into families. A popular approach is to consider orthology relationships. In theory, genes are orthologous when they are related only by events of speciation (i.e. diversification

of the whole genome), not by duplication of horizontal gene transfer. Because the true course of gene diversification events is unknown, we must rely on practical definitions of orthology. This theoretical definition implies that two orthologues cannot occur in the same genome. A usual criteria is thus to look for the bidirectional best hit (BBH) in a similarity search of all the genes in a genome against all of the genes in another genome pairwise genome (Altenhoff and Dessimoz 2009).

This pairwise relationship can be used to build a network of genes covering the whole pangenome dataset, where cliques (groups where the found relationship is transitive among members) are recognised as clusters of orthologous genes or COGs (Tatusov et al. 1997). Using this practical definition, it is straightforward to classify any gene into a cluster, many of which will however be clusters of genes on their own: orphan genes with no homologues, but also those resulting from a recent transfer or duplication. By construction, these COGs can only be absent or present in a single copy in a genome, which proves very convenient for representing the distribution of genes in the pangenome by a genome-to-COG binary matrix filled with zeros and ones. This representation can be handled by many simple methods that model the transition between these binary states over the tree of the genomes, i.e. events of gene gain and loss (Fig. 3a) (Mirkin et al. 2003). This approach has been widely used, but suffers from its stringent definition that leaves many homologous genes out of COGs under scrutiny, which may strongly flaw the inferred ancestral genome gene contents and the derived conclusions on ancestral functional repertoires.

Instead, it is possible to consider a whole family of homologues, which distribution of the family in genomes can again be represented in a matrix of counts, where this time values range from zero to any integer number. Models of pangenome evolution can account for this multiplicity of gene copy number by invoking extra gene gain events (Fig. 3b) (Csurös 2008; Csurös and Miklós 2009). The nature of these gain events—duplication or horizontal gene transfer—is, however, not inferred as it fundamentally requires to know the phylogenetic relationships between genes within a homologous family.

3.3   Linkage of Genes and Syntenic Blocks

Notwithstanding the type of evolving unit considered (aligned genome segment or gene family), all units are usually considered to evolve independently on the phylogeny. This is, however, not always realistic given the high linkage disequilibrium found in bacterial genome—the non-independence of physically linked characters in evolution, a consequence of their clonal mode of reproduction.

Linkage can be introduced in a pangenome evolution model by specifying the location of genes on a genomic map. The evolution of genes on the map is then modelled through events of insertion, deletion and rearrangement. This map can relate the absolute position of genes in contemporary genomes (i.e. with nucleotide site coordinates) by chopping all genomes in the dataset into syntenic blocks, where
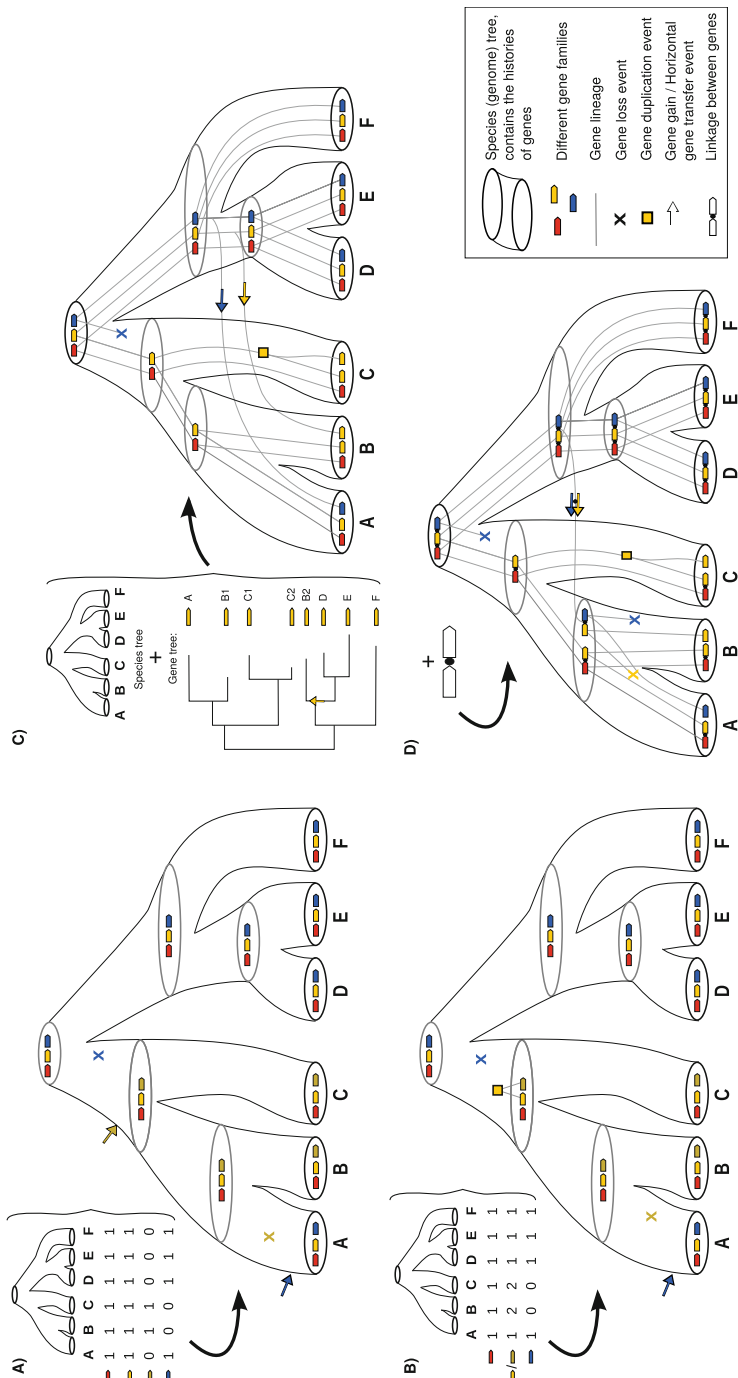
Fig. 3 Schematic representation of scenarios of gene family evolution, inferred with varied levels of information input. (**a**, **b**) Phylogenetic profile: only the information of presence/absence of gene families is used to infer the scenarios of evolution. (**a**) Genes are classified into orthologous groups; (**b**) genes are classified into homologous groups, retaining the information that yellow and brown gene families are related, the emergence of the latter being explained by a duplication of the former. (**c**) Gene tree/species tree reconciliation: the information of topology of the gene tree is used, revealing a more complex scenario, with an additional transfer. (**d**) Synteny-aware reconciliation: the information of linkage of genes (synteny) is used, suggesting that apparently independent transfer events may have happen as one joint event

homology and overall gene order between genomes is conserved (Vallenet et al. 2006; Darling et al. 2010). However, as mentioned above, the larger the genome sample, the more syntenic blocks will split and shrink. Based on such genome maps, the history of each syntenic blocks can be estimated, describing the ancestral events of pangenome evolution. Even though in theory the map evolves over time due to genome rearrangements (Darling et al. 2008), in practice the maps are assumed to be constant in order to allow to focus on fine-grained changes within the syntenic blocks. This assumption is commonly made, for example when investigating homologous recombination in the core genome (Didelot et al. 2010).

Another option is to map the relative position of smaller evolving units (usually gene families) in each genome of the dataset. Such a relative map can be represented by a matrix of presence or absence of a direct adjacency between genes in a given genome, contemporary or ancestral. This more abstract representation allows the use of incomplete data, such as draft genome assemblies, where the physical linkage of sequences is not fully or not unambiguously documented. The evolution of gene neighbourhood is modelled by invoking events of creation and breakage of adjacencies between neighbour genes, thereby modelling any insertion, deletion and rearrangement. Ancestral state reconstruction (see below) is then undertaken, by estimating a genome map at each ancestral node of a species phylogeny (Fig. 3d) (Bérard et al. 2012; Patterson et al. 2013; Duchemin et al. 2017). These models are, however, quite heavy computationally and may become overwhelmed by large structural diversity in the dataset.

# 4 Methodological Approaches to the Reconstructing Pangenome Microevolution

4.1 Ancestral State Reconstruction

The inference of ancestral genomes and corresponding gene gain and loss scenarios can be a complex and computationally intensive task, but it can also be simplified to the point that it becomes almost straightforward if the research questions are relatively simple. For example, using profiles of gene presence/absence in genomes and a phylogenetic tree as only input, ancestral state reconstruction can be applied to infer in which internal nodes of the tree the genes were present, and therefore on which branches the genes were gained and lost. For a general review on ancestral state reconstruction, see Joy et al. (2016). One of the simplest and most popular approach is to perform a parsimonious reconstruction, where the number of gain and loss events is minimised without the need to estimate any parameter (Mirkin et al. 2003). In practice, this is more or less equivalent to performing maximum likelihood inference under a model in which gain and loss happen at the same small rates. However, probabilistic modelling of state evolution has the interesting property to integrate over several possible scenarios. Even a maximum likelihood point estimate

of the presence of a gene at a given ancestral node will therefore consist of a non-binary probability, a nuanced result allowing to consider the uncertainty in the ancestral reconstruction (Pagel 1999). A similar Bayesian approach is stochastic character mapping (Huelsenbeck et al. 2003), which consists in sampling gain and loss histories from their posterior probability distribution via a Monte Carlo method.

Ancestral state reconstruction is particularly well suited to analyses focused on specific genes rather than the whole pangenome, for example analysing the gain and loss of pathogenicity genes (Dingle et al. 2014) or resistance genes (Ward et al. 2014). It can also be applied more generally to all genes in a pangenome, and the rates of gain and loss would typically be assumed to be equal meaning that the genome size is at equilibrium (Touchon et al. 2009). Alternatively, the reconstruction can be based on genomic elements known to be gained and lost in one block, such as bacteriophages, plasmids, and integrative conjugational elements (Zhou et al. 2013). This represents one simple way of dealing with the linkage of genes mentioned previously, although at the cost of potentially losing information about the gene content evolution of the genomic elements assumed to be perfectly linked. At the other end of the spectrum, the reconstruction can be based on smaller elements than genes, for example k-mers, but in this case it becomes vital to relax the assumption of a fixed clock on gain and loss, for example using a local clock model (Didelot et al. 2009) as illustrated in Fig. 4. This technique has been applied to the pangenomes of *Escherichia coli* (Didelot et al. 2012) and *Campylobacter jejuni* (Sheppard et al. 2013a), showing in both cases a strong relationship between evolution of the accessory genome via gain and loss events and evolution of the core genome via homologous recombination.

An important drawback of ancestral state reconstruction methods is that they ignore the nature (recombination or duplication) and origins (recombination donor) of gene gain events, which can yield partial and inaccurate scenarios when the true history is complex, especially with many recombination events. In particular, the exploitable signal from a profile of gene presence/absence in extant genomes are quickly saturated when several gene copies coexist in a genome, and likely descend from separate past events. This issue can sometimes be tackled by defining strict families of orthologs, where every gene is present in one copy or none, but at the cost of losing the information on evolution of homologues. Ancestral state reconstruction could also in principle be applied to data on family of homologues, where each genome can contain zero, one or more copies of a gene. This would require to fit a ladder model similar to the ones used when analysing microsatellite data (Ohta and Kimura 1973; Wilson and Balding 1998). This approach is difficult in practice because bacterial accessory gene families of interest have often too complex histories to reliably infer orthologous groups and have high gain and/or loss rates that quickly saturate signals. It has, however, been applied successfully in studies where genomes of single representatives from fairly distant species were compared, thus ignoring the 'messy' variation introduced by within-population evolution (Csűrös and Miklós 2009).
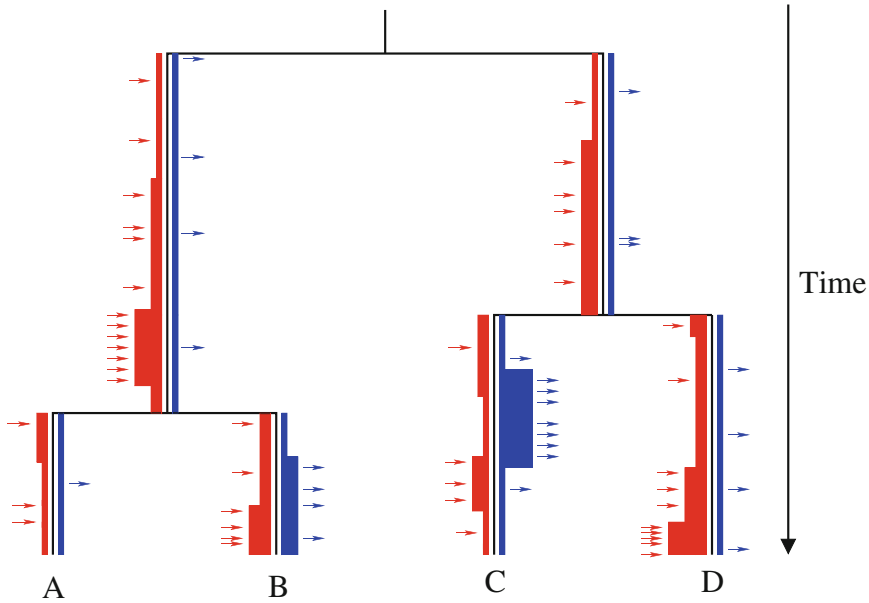
Fig. 4 Illustration of a pangenome gain and loss model with local clock. The clonal genealogy is shown in black. The width of the red block on the left of the branches is proportional to the rate of gain. Similarly, the blue block on the right of each branch represents the rate of loss. Both the gain and loss rates occasionally change over time. Individual gain events are represented by red arrows, and individual loss events are represented by blue arrows

4.2 Phylogenetic Reconciliation

To deliver more informative scenarios of evolution, it is necessary to know the origin of gene gains, which effectively means to know the relationship between observed genes. Gene tree versus species tree reconciliation methods compare the topologies of phylogenetic trees built from individual gene sequences against a reference species tree (Maddison 1997). In the context of pangenome analysis, the species tree is a phylogenetic tree reconstructed from the whole of the core genome. Species and gene trees often have inconsistent topologies, which could happen by chance, especially since the gene tree typically has limited statistical support, or may be the result of evolutionary events affecting the history of the gene relative to the clonal history. Reconciliation methods intend to explain the significant topological discords by events of gene duplication, transfer, or loss (Szollosi et al. 2015). Figure 3c illustrates the principles behind reconciliation methods. Practically, both trees are annotated with the inferred events, such that there is a full agreement on the course of events, from the root of the gene lineage to the contemporary distribution of genes in genomes—thus reconstructing the path of evolution and diversification of genes in the clonal frame of genome evolution. As a result, this approach allows to explicitly determine the donors and recipients of transferred genes, or the ancestor in which a

gene was duplicated. Methods for pangenome reconciliation analysis have been proposed based on parsimonious reconstruction (Abby et al. 2010; Jacox et al. 2016) and probabilistic models (Szollosi et al. 2012, 2013).

The ancestral state reconstruction approach and the reconciliation approach have a lot in common, and the latter can be thought of as a natural extension of the former when observation is not limited to presence or absence or number of copies of a gene, but also includes the phylogenetic relationships between genes from separate genomes. Reconciliation methods are therefore superior in the sense that they exploit more of the available signal, but they are also much more challenging to implement computationally and have so far been limited to analysis of a handful of genomes. Ancestral state reconstruction methods are currently more popular but we predict that reconciliation methods will become increasingly widespread in the near future with the development of more effective statistical methods. Beyond the study of the atomic events whereby the pangenome evolves, both methods allow to infer ancestral states in hypothetical ancestors, or in other words to reconstruct ancestral genomes. Doing so, one can derive the expected phenotypic traits of the ancestors—antimicrobial resistance, metabolic activities, even ecological lifestyle. These inferred traits can then be compared to historical records of Earth evolution or pathogen epidemic spread to try and find causal relations between biological activity and the course of events (David and Alm 2011; Holden et al. 2013), or be considered in support of further ancestral reconstruction, such as scenarios of ecological niche colonisation (Lassalle et al. 2017).

References

Abby SS, Tannier E, Gouy M, Daubin V (2010) Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. BMC Bioinformatics 11:324. https://doi.org/10.1186/1471-2105-11-324

Altenhoff AM, Dessimoz C (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.1000262

Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27:334–342

Aziz RK, Bartels D, Best AA et al (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics 9:75. https://doi.org/10.1186/1471-2164-9-75

Bankevich A, Nurk S, Antipov D et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021

Bérard S, Gallien C, Boussau B et al (2012) Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts374

Boussau B, Karlberg EO, Frank AC et al (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci. https://doi.org/10.1073/pnas.0400975101

Buckee C, Jolley K, Recker M et al (2008) Role of selection in the emergence of lineages and the evolution of virulence in Neisseria meningitidis. Proc Natl Acad Sci USA 105:15082–15087. https://doi.org/10.1073/pnas.0712019105

Castillo-Ramírez S, Harris SR, Holden MTG et al (2011) The impact of recombination on dN/dS within recently emerged bacterial clones. PLoS Pathog 7:e1002129. https://doi.org/10.1371/journal.ppat.1002129

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205. https://doi.org/10.1038/nrg2526

Collins C, Didelot X (2018) A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol 14: e1005958. https://doi.org/10.1101/140798

Croucher NJ, Page AJ, Connor TR et al (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196

Csurös M (2008) Ancestral reconstruction by asymmetric Wagner parsimony over continuous characters and squared parsimony over distributions. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)

Csurös M, Miklós I (2009) Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol 26:2087–2095. https://doi.org/10.1093/molbev/msp123

Darling AE, Miklós I, Ragan MA (2008) Dynamics of genome rearrangement in bacterial populations. PLoS Genet 4:e1000128. https://doi.org/10.1371/Citation

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. https://doi.org/10.1371/journal.pone.0011147

David LA, Alm EJ (2011) Rapid evolutionary innovation during an Archaean genetic expansion. Nature. https://doi.org/10.1038/nature09649

Didelot X, Maiden MCJ (2010) Impact of recombination on bacterial evolution. Trends Microbiol 18:315–322. https://doi.org/10.1016/j.tim.2010.04.002

Didelot X, Wilson DJ (2015) ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11:e1004041. https://doi.org/10.1371/journal.pcbi.1004041

Didelot X, Achtman M, Parkhill J et al (2007) A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination? Genome Res 17:61–68. https://doi.org/10.1101/gr.5512906.1

Didelot X, Darling AE, Falush D (2009) Inferring genomic flux in bacteria. Genome Res 19:306–317. https://doi.org/10.1101/gr.082263.108.clearly

Didelot X, Lawson DJ, Darling AE, Falush D (2010) Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 186:1435–1449. https://doi.org/10.1534/genetics.110.120121

Didelot X, Méric G, Falush D, Darling AE (2012) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13:256. https://doi.org/10.1186/1471-2164-13-256

Didelot X, Walker AS, Peto TE et al (2016) Within-host evolution of bacterial pathogens. Nat Rev Microbiol 14:150–162. https://doi.org/10.1038/nrmicro.2015.13

Dingle KE, Elliott B, Robinson E et al (2014) Evolutionary history of the *Clostridium difficile* pathogenicity locus. Genome Biol Evol 6:36–52. https://doi.org/10.1093/gbe/evt204

Donnelly P, Tavare S (1995) Coalescents and genealogical structure under neutrality. Annu Rev Genet 29:401–421

Duchemin W, Anselmetti Y, Patterson M et al (2017) DeCoSTAR: reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. Genome Biol Evol. https://doi.org/10.1093/gbe/evx069

Everitt RG, Didelot X, Batty EM et al (2014) Mobile elements drive recombination hotspots in the core genome of *Staphylococcus aureus*. Nat Commun 5:3956. https://doi.org/10.1038/ncomms4956

Fisher RA (1931) XVII—the distribution of gene ratios for rare mutations. Proc R Soc Edinburgh. https://doi.org/10.1017/S0370164600044886

Griffiths RC, Marjoram P (1997) An ancestral recombination graph. Prog Popul Genet Hum Evol (Minneapolis, MN, 1994) 87:257–270

Griffiths R, Tavare S (1994) Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc B Biol Sci 344:403–410

Hedge J, Wilson DJ (2016) Practical approaches for detecting selection in microbial genomes. PLoS Comput Biol 12:e1004739. https://doi.org/10.1371/journal.pcbi.1004739

Holden MTG, Hsu L-Y, Kurt K et al (2013) A genomic portrait of the emergence, evolution and global spread of a methicillin resistant *Staphylococcus aureus* pandemic. Genome Res 23:653–664

Huelsenbeck JP, Nielsen R, Bollback JP (2003) Stochastic mapping of morphological characters. Syst Biol. https://doi.org/10.1080/10635150390192780

Jacox E, Chauve C, Szöllősi GJ et al (2016) ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics 32:2056–2058. https://doi.org/10.1093/bioinformatics/btw105

Jolley KAA, Maiden MCJ (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11:595. https://doi.org/10.1186/1471-2105-11-595

Joy JB, Liang RH, Mccloskey RM et al (2016) Ancestral reconstruction. PLoS Comput Biol 12: e1004763. https://doi.org/10.1371/journal.pcbi.1004763

Kingman JFC (1982) The coalescent. Stoch Process Appl 13:235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12:32. https://doi.org/10.1186/1471-2164-12-32

Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. Philos Trans R Soc B Biol Sci 361(1475):1929–1940

Kurtz S, Phillippy A, Delcher AL et al (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12

Lassalle F, Planel R, Penel S et al (2017) Ancestral genome estimation reveals the history of ecological diversification in agrobacterium. Genome Biol Evol 9:3413–3431. https://doi.org/10.1093/gbe/evx255

Lawrence J (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. Curr Opin Genet Dev 9(6):642–648

Li L, Stoeckert CJJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13(9):2178–2189. https://doi.org/10.1101/gr.1224503.candidates

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536. https://doi.org/10.1017/CBO9781107415324.004

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature. https://doi.org/10.1038/351652a0

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol. https://doi.org/10.1186/1471-2148-3-2

Moran PAP (1958) Random processes in genetics. Math Proc Camb Philos Soc 54:60–71

Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop M, Cannings C (eds) Handbook of statistical genetics. Wiley, Hoboken, NJ

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304. https://doi.org/10.1038/35012500

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electropho-retically detectable alleles in a finite population. Genet Res (Camb) 22:201–204. https://doi.org/10.1017/S0016672308009531

Oliveira PH, Touchon M, Cury J, Rocha EPC (2017) The chromosomal organization of horizontal gene transfer in bacteria. Nat Commun. https://doi.org/10.1038/s41467-017-00808-w

Page AJ, Cummins CA, Hunt M et al (2015) Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421

Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401:877–884. https://doi.org/10.1038/44766

Patterson M, Szöllosi G, Daubin V, Tannier E (2013) Lateral gene transfer, rearrangement, reconciliation. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-14-S15-S4

Pepperell CS, Casto AM, Kitchen A et al (2013) The role of selection in shaping diversity of natural M. tuberculosis populations. PLoS Pathog 9:e1003543. https://doi.org/10.1371/journal.ppat.1003543

Petersen L, Bollback JP, Dimmic M et al (2007) Genes under positive selection in *Escherichia coli*. Genome Res 17:1336–1343. https://doi.org/10.1101/gr.6254707

Rosenberg NA, Nordborg M (2002) Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat Rev Genet 3:380–390. https://doi.org/10.1038/nrg795

Seemann T (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153

Sheppard SK, Didelot X, Jolley KA et al (2013a) Progressive genome-wide introgression in agricultural Campylobacter coli. Mol Ecol 22:1051–1064. https://doi.org/10.1111/mec.12162

Sheppard SK, Didelot X, Meric G et al (2013b) Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. Proc Natl Acad Sci USA 110:11923–11927. https://doi.org/10.5061/dryad.28n35.

Sims GE, Kim S-H (2011) Whole-genome phylogeny of *Escherichia coli*/Shigella group by feature frequency profiles (FFPs). Proc Natl Acad Sci USA 108:8329–8334. https://doi.org/10.1073/pnas.1105168108

Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988

Szollosi GJ, Boussau B, Abby SS et al (2012) Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc Natl Acad Sci 109:17513–17518. https://doi.org/10.1073/pnas.1202997109

Szollosi GJ, Rosikiewicz W, Boussau B et al (2013) Efficient exploration of the space of reconciled gene trees. Syst Biol 62:901–912. https://doi.org/10.1093/sysbio/syt054

Szollosi GJ, Tannier E, Daubin V et al (2015) The inference of gene trees with species trees. Syst Biol 64:e42–e62. https://doi.org/10.1093/sysbio/syu048

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science. https://doi.org/10.1126/science.278.5338.631

Thomas CM, Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat Rev Microbiol 3:711–721. https://doi.org/10.1038/nrmicro1234

Touchon M, Hoede C, Tenaillon O et al (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344. https://doi.org/10.1371/journal.pgen.1000344

Vallenet D, Labarre L, Rouy Z et al (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res. https://doi.org/10.1093/nar/gkj406

Vaughan TG, Welch D, Drummond AJ et al (2017) Inferring ancestral recombination graphs from bacterial genomic data. Genetics 205:857–870. https://doi.org/10.1534/genetics.116.193425

Vos M (2011) A species concept for bacteria based on adaptive divergence. Trends Microbiol 19:1–7

Vos M, Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199–208. https://doi.org/10.1038/ismej.2008.93

Vos M, Wolf AB, Jennings SJ, Kowalchuk GA (2013) Micro-scale determinants of bacterial diversity in soil. FEMS Microbiol Rev 37(6):936–954

Ward MJ, Gibbons CL, McAdam PR et al (2014) Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of *Staphylococcus aureus* clonal complex 398. Appl Environ Microbiol 80:7275–7282. https://doi.org/10.1128/AEM.01777-14

Wilson IJ, Balding DJ (1998) Genealogical inference from microsatellite data. Genetics 150:499–510

Wilson DJ, McVean G (2006) Estimating diversifying selection and functional constraint in the presence of recombination. Genetics 172:1411–1425. https://doi.org/10.1534/genetics.105.044917

Wiuf C, Hein J (2000) The coalescent with gene conversion. Genetics 155:451–462

Wright S (1931) Evolution in Mendelian populations. Genetics 16:97–159

Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13:303–314. https://doi.org/10.1038/nrg3186

Yang C, Cui Y, Didelot X et al (2018) Why panmictic bacteria are rare. bioRxiv. https://doi.org/10.1101/385336

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829. https://doi.org/10.1101/gr.074492.107

Zhou Z, McCann A, Litrup E et al (2013) Neutral genomic microevolution of a recently emerged pathogen, *Salmonella enterica* Serovar Agona. PLoS Genet 9:e1003471. https://doi.org/10.1371/journal.pgen.1003471