



The Sources of Phylogenetic Conflicts

Dominik Schrempf, Gergely Szöllősi

► To cite this version:

Dominik Schrempf, Gergely Szöllősi. The Sources of Phylogenetic Conflicts. Scornavacca, Celine; Delsuc, Frédéric; Galtier, Nicolas. Phylogenetics in the Genomic Era, No commercial publisher | Authors open access book, pp.3.1:1–3.1:23, 2020. hal-02535482

HAL Id: hal-02535482

<https://hal.science/hal-02535482>

Submitted on 10 Apr 2020

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


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




Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Chapter 3.1 The Sources of Phylogenetic Conflicts

Dominik Schrempf

Department of Biological Physics, Eötvös Loránd University
Pázmány P. stny. 1A, H-1117 Budapest, Hungary
dominik.schrempf@gmail.com
 <http://orcid.org/0000-0001-8865-9237>

Gergely Szöllősi

MTA-ELTE “Lendület” Evolutionary Genomics Research Group
Department of Biological Physics, Eötvös Loránd University
Pázmány P. stny. 1A, H-1117 Budapest, Hungary
Evolutionary Systems Research Group, Centre for Ecological Research
Hungarian Academy of Sciences, 8237 Tihany, Hungary
ssolo@elte.hu
 <http://orcid.org/0000-0002-8556-845X>

Abstract

Recombination breaks up the evolutionary history between genomic regions and, as a result, the evolutionary history of different genomic regions may differ. In fact, conflicting phylogenetic signal between genes is commonplace. The reasons for conflicting signal may be statistical or systematic in nature. In order to avoid strongly supported but incorrect inferences driven by systematic error, use of appropriate phylogenetic methods accounting for these processes is of fundamental importance.

This chapter reviews possible causes of phylogenetic conflict between genes. Processes generating conflict, including gene duplication and loss, horizontal gene transfer, hybridization, and incomplete lineage sorting are presented using classic examples. In particular, we discuss compelling evidence for whole genome duplications in fish, as well as plants, and the role of horizontal transfer in the spread of antibiotic resistance. Finally, building on the material presented, we show how these processes lead to phylogenetic conflict, and how they can be described by phylogenetic models.

How to cite: Dominik Schrempf and Gergely Szöllősi (2020). The Sources of Phylogenetic Conflicts. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter No. 3.1, pp. 3.1:1–3.1:23. No commercial publisher | Authors open access book. The book is freely available at <https://hal.inria.fr/PGE>.

Funding DS and GJSz received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 714774 and the grant GINOP-2.3.2.-15-2016- 00057.

1 Introduction

The evolution of present-day life, that is, the causal events of the diversity of observed sequences, has been a complicated and intertwined process. Consequently, resolving the events and their order from observed hereditary sequences is challenging. This section sets out to describe some of these evolutionary processes such as transmission of genes from parents to offspring during reproduction, gene duplication and loss or horizontal gene





© Dominik Schrempf and Gergely Szöllősi.

Licensed under Creative Commons License CC-BY-NC-ND 4.0.

Phylogenetics in the genomic era.

Editors: Celine Scornavacca, Frédéric Delsuc and Nicolas Galtier; chapter No. 3.1; pp. 3.1:1–3.1:23

 A book completely handled by researchers.

 No publisher has been paid.

3.1:2 The sources of Phylogenetic Conflicts

transfer, and continues to corroborate that they affect phylogenetic inference. In fact, if these processes are ignored, they can lead to false inferences with strong statistical support (see Chapter 2.1 [Simion et al. 2020]).

Before moving on, we have to reflect on what phylogenetic conflicts actually are. The concept of *homology* (Greek; *homos*, same; *logos*, relation) plays a pivotal role in answering this question. Evolutionary characters are defined to be either homologous or analogous (Greek; *ana*, up to, upwards), if they have shared or no shared ancestry, respectively. The definition of homology appears simple but harbours several layers of hidden complexity. For example, it is possible that ultimately all characters have a single ancestor. The decision about whether two sequences are homologous or not involves a statistical test where the null hypothesis is that the observed similarity simply arose by chance.

Consequently, not single DNA or amino acid characters are compared, but sequences of characters. For this reason we focus on the larger, abstract phylogenetic unit of a *gene*. We think of a gene as a set of characters, mostly but not necessarily adjacent along the sequence, which is inseparably transmitted together from parent to offspring, or, more abstract, from donor to recipient. Recombination reshuffles the genetic sequence during the production of offspring, and thereby separates different genomic fragments so that their evolutionary histories are partially independent. We assume that recombination acts only between and not within genes. The assumption that genes are passed on as a whole is reasonable because recombination within genes is likely to break function, and so, is likely to be deleterious.

Usually, pair-wise distances between sequences are determined using BLAST (Altschul et al., 1997). The resulting distance matrix can be interpreted as a completely connected graph. The vertices are the characters, and the edge weights correspond to BLAST scores. Typically, a threshold on edge weights is applied to produce a sparser graph without edge weights. Then, graph clustering algorithms and other criteria such as sequence overlap are used to identify sub-graphs representing tightly connected homologous clusters (Miele et al., 2011; Li et al., 2003; Enright et al., 2002). Admittedly, the threshold discriminating between sequences being part of the same or of different homology clusters depends on user-defined parameters.

The outcome of homology search are distinct sets of homologous genes which are termed *gene families*. Genes are sorted into different families if we have no means of finding common ancestry, either because the genes are not similar enough, or because intermediate homologous genes have not been sampled. After all, it is possible that all genes belong to a single gene family. Gene families are a treasure trove of information to reconstruct ancestry but they bear some enigmas. Given that we have decided that two genes belong to the same gene family, we can track their histories back in time until they merge. At the merging point, different events may have happened. First, and probably most frequently, the corresponding event may be associated with a cell division during reproduction. In this case, the ancestor and the two descendants are all members of the same species, and the two descending genes are at the same genomic position, which we call *locus*. Loosely, a locus is the genomic location where a gene resides (Gillespie, 2004). As a result of speciation, the two descendant genes may end up in two different species (Section 2.4). We call genes related by coalescence or speciation *orthologs* (Greek; *orthos*, in a straight line, true, correct, see Chapter 2.4 [Fernández et al. 2020]). Second, the merging point may correspond to some form of gene duplication (Section 2.2). The two descending genes co-exist at different loci in the genome of the same individual. Genes related by duplication are termed *paralogs* (Greek; *para*, besides). Whole genome duplication is a special, concerted form of gene duplication; genes related by whole genome duplication are termed *ohnologs* (in honor of

Susumu Ohno). Third, the merging point may correspond to some form of horizontal gene transfer (Section 2.2). The ancestral and one of the descendant genes are of the same species, which is termed *donor*. The other descendant gene is part of the genome of an individual belonging to a different species, the *recipient*. In this case, we speak of *xenologs* (Greek; *xenos*, foreign). Finally, genes brought together by inter-species hybridization (Section 2.3) are called *homeologs* (Greek; *homeo*, similar).

Biological mechanisms of duplication and transfer will be discussed in the next section. The evolutionary history, or phylogenetic relationships, of all genes within a gene family can be depicted in a *gene tree*. We use the term tree to refer to a tree-like object including topology and branch lengths. Similar to gene trees, relationships of species are depicted in a *species tree*. Increasingly sophisticated models of sequence evolution (see Chapter 1.1 [Pupko and Mayrose 2020]) have allowed accurate reconstruction of gene and species trees.

Importantly — and this brings us back to our original problem of defining phylogenetic conflict — recombination breaks up the histories of different genes within a gene family and, all the more, between gene families. Within a few generations linkage is erased and genes evolve independently of each other. Hence, phylogenetic methods usually assume free recombination between genes (see Chapter 3.4 [Bryant and Hahn 2020]). In the light of recombination, the presence of paralogs, ohnologs, xenologs, and homeologs — that is, gene duplication together with loss, horizontal gene transfer, and hybridization — can cause disagreement between a gene tree and the species tree. As a matter of fact, it can even cause disagreement about the species tree within a single gene tree, for example, when histories of two ancient paralogs are different. Disagreement of a gene tree with the species tree or disagreement within a gene tree about the species tree is the definition of *phylogenetic conflict*. However, we usually have no *a priori* knowledge about the type of homology, and so do not know if two genes are orthologs, paralogs, ohnologs, xenologs, or homeologs.

Further, genes within a gene family have different, independent coalescent times. Consequently, orthologous genes not coalescing before the previous speciation event can lead to phylogenetic conflict (see Chapter 3.3 [Rannala et al. 2020]). Moreover, free recombination between genes also implies that gene families have independent histories, aside from evolving along the same species tree. It follows that phylogenetic conflict will also be manifested as disagreement between the gene trees themselves. As a matter of fact, phylogenetic conflict is mostly observed as disagreement between gene trees because the species tree is unknown, as it cannot be reconstructed directly.

For example, the cumulative effects of gene duplication, gene loss and horizontal gene transfer result in severe conflict between six gene trees obtained from protein coding sequences (Philippe and Forterre, 1999). Further, all gene trees conflict with the famous three domain tree previously obtained from ribosomal RNA by Woese et al. (1990). Even within Eukaryotes, gene trees reconstructed from different protein sequences exhibit conflicts, and the order of emergence of basal groups depends on the rate of evolution of the protein used.

It seems that we are confronted with an unsolvable conundrum at the heart of which are recombination, gene duplication and loss, horizontal gene transfer, and hybridization. In this chapter we aim to elucidate how independent coalescence, gene duplication and loss, horizontal gene transfer, and hybridization can cause phylogenetic conflict. Given that we understand the causes and processes responsible for phylogenetic conflict, it is our task to develop methods that can assess the likelihood of gene trees supporting a given species tree (see Chapter 3.2 [Boussau and Scornavacca 2020]). In the end, most phylogeneticists are interested in resolving the best-supported species tree, and elucidate the origin and evolution of life.

2 Processes causing phylogenetic conflict

Evolutionary change of genetic sequences is introduced by a series of mutations which may spread in a population, either by chance or because they provide selective advantages. Mutations affecting more than one DNA base are called structural mutations. The size of structural mutations ranges from a few bases to billions of bases. Structural mutations which remove parts from the genome and add parts to the genome are called deletions and insertions, respectively. Many biological processes can cause the creation of new gene structures within the genome of an individual (for a review, see Long et al., 2003). Some of these processes create perfect to near-perfect gene copies, others may only copy parts of genes or combine/extend different genes (e.g., exon-shuffling). Further processes include retroposition, where a gene is inverted by reverse transcription with subsequent insertion back into the genome of the individual, insertion of mobile elements into genes, or fusion and fission of genes.

Even though the creation of new gene structures involves complex biological processes, their quantification in mathematically tractable models requires significant simplification. To our knowledge, most phylogenetic models only consider perfect copying of genes, either in the same individual (gene duplication), or copy/cut and paste into a coexisting individuals from a different species (horizontal gene transfer). In accordance with current phylogenetic models, we distinguish in the following between (1) gene origination, which is the formation of a novel genetic sequence establishing a new gene family and (2) gene birth and death, which are the addition of a gene to a gene family and the removal of a gene from a gene family, respectively. Further, inter-species hybridization and incomplete lineage sorting are discussed.

2.1 Gene origination

Across all three domains of life, up to 30 percent of the genes within an organism have no known homologs, and are presumably of recent origin (Tautz and Domazet-Lošo, 2011; Dehal, 2001). First, such *orphan genes* can emerge through the creation of new genetic material for example by gene duplications due to errors during recombination, the acquisition of extrinsic genes, or by transposition mechanisms. The generation of new genetic material is followed by a phase of fast adaptive evolution leading to divergence beyond the threshold of homology searches.

Second, gene families may originate due to *de novo* evolution. Random sequences from non-coding regions may form cryptic functional sites that could subsequently come under regulatory control. In this case, the creation of *de novo* genes may be as simple as having a mutation at a single base that activates transcription of a downstream stretch of DNA. The activated sequence may fortuitously code for a protein enhancing fitness. Initially, *de novo* evolution was deemed highly unlikely, but several cases with compelling evidence have been observed and reported. For example, several human specific genes have been detected which correspond to non-transcribed regions in other primates (Knowles and McLysaght, 2009). The question is: “Is more likely that these genes have been switched on in humans, than switched off in all other primates”? The patterns of the observed population data provide evidence that the open reading frames are functional.

2.2 Gene birth and death

Changes in gene copy number in a gene family are a frequent mutational event (Reams and Roth, 2015). However, the exact mechanisms are difficult to discover, because they vary with genomic position. Further, determination of gene birth rates, that is, the number of events happening per unit time, is hard. Furthermore, the mechanisms of gene birth and death are distinct, and therefore, different rates can arise. Conceptually, gene birth and gene death denote the increase and decrease of gene copy number, respectively. Below, we give a brief summary of gene birth through gene duplications, horizontal gene transfer, and hybridization, and gene death which is synonymously called gene loss.

Gene duplication

Gene duplication (Ohno, 1970), which is the emergence of a heritable copy of a gene within a genome of an individual, is an integral constituent of the evolution of biological complexity. Gene duplication is prevalent across the tree of life and has greatly shaped the hereditary material of present-day organisms (Kondrashov et al., 2002). In fact, gene duplication is the most common source of new genes in Eukaryotes. The presence of a gene copy can have detrimental as well as beneficial effects. Often, gene duplications are present in some but not all individuals of natural populations resulting in a situation called copy number variation. Although we have fundamental knowledge about gene duplications on the functional and the genomic level (Conant and Wolfe, 2008), knowledge about their emergence, and maintenance is incomplete. In the following, common biological mechanisms leading to gene duplication are outlined.

1. Unequal crossing over is an event where the breaks during recombination happen at different positions on the chromosomes which are then misaligned during meiosis.
2. During replication slippage, the DNA polymerase erroneously changes position and copies a part of the chromosome twice.
3. Retrotransposition is a process where messenger RNA is reverse transcribed to DNA which is integrated back into the genome.
4. Repetition of protein domains is a pattern observed within many genes. Repeated domains can be caused by exon-shuffling, which is the duplication of exons as a result of recombination between non-homologous sequences.
5. Transposition of active mobile elements (Long et al., 2003; Lynch, 2007) may also lead to novel genes partially containing duplicated genetic material.
6. The nucleus of polyploid organisms contains more than two sets of chromosomes. Whole genome duplication is an extreme type of mutation where the gamete produced during meiosis carries the entire diploid genome rather than the haploid one. Whole genome duplication results in so-called tetraploidy, a condition where each chromosome is present four times in the nucleus. Repeated whole genome duplication leads to octaploidy, although copies might be lost between the consecutive whole genome duplications resulting in a different number of chromosome copies. Whole genome duplication is more common in plants than animals, but see the discussion below about fish and jawed vertebrates.
7. In a similar manner, the fusion of two genomes during hybridization of two species results in an allo(tetra)polyploid and is discussed in Section 2.3.

In prokaryotes, the process of gene duplication is less well understood, but several methods allowing assessment of duplication rate are now available (Reams and Roth, 2015). Often, the methods involve beneficial multi-step processes.

3.1:6 The sources of Phylogenetic Conflicts

Most newly created genes will be removed by genetic drift within a few generations. Only a very small fraction of novel genetic material will rise in frequency and eventually become fixed in a population. In the case of gene duplications, the two gene copies may be redundant and consequently under relaxed selection. As a result, one gene copy may again be lost by a deletion or the accumulation of loss-of-function mutations (see below). More interesting, one gene copy may evolve a novel biological function; a process termed neofunctionalization. The two gene copies may also subfunctionalize, and perform separate functions which, in concert, fulfill all original functions and may provide more (see Chapter 4.2 [Robinson-Rechavi 2020]).

Whole genome duplications have been a major confounding factor in phylogenetic analyses (Van de Peer et al., 2017). For example, two rounds of whole genome duplications (2R) in the last common ancestor of all jawed vertebrates were the cause of a fourfold increase of vertebrate genes (Sidow, 1996). The 2R hypothesis was intensely debated, but is now widely accepted (Meyer and Van de Peer, 2005). Even more, a third round of whole genome duplication has happened in fish (Meyer and Schartl, 1999; Brunet et al., 2006), but evidence was not always conclusive. For example, Robinson-Rechavi et al. (2001) analyze the evolutionary history of 35 gene families in fish. Seven gene families follow a pattern consistent with an ancestral whole genome duplication, whereas eleven gene families show duplications that had most likely happened after the divergence of the considered fish species, and 19 gene families do not show any signs of duplications.

Whole genome duplications are even more common in plants. For example, there is evidence that the genome of the common ancestor of angiosperms was duplicated three times. In fact, this may be the reason for the morphological and ecological diversification of angiosperms. Polyploidy increases biological complexity and the amount of genetic material subject to natural selection. The resulting phenotypic variation, mostly caused by overall differences in expression levels between diploid and polyploid individuals may lead to selective advantages in polyploids. For example, polyploidy has been repeatedly discussed in the context of major evolutionary transitions, and hybrid vigor. Further, it has been hypothesized that polyploidy is a major driver of adaptive radiation of species (De Bodt et al., 2005).

Inference of ancient whole genome duplications is difficult due to saturation of synonymous distance (Tiley et al., 2018). Recently, probabilistic methods have been developed to infer whole genome duplications (Zwaenepoel and Van de Peer, 2019) employing amalgamated likelihood estimation (ALE, Szöllősi et al. 2013a; Chapter 3.2 [Boussau and Scornavacca 2020]).

ALE is based on conditional clade probabilities, which roughly correspond to the observed frequency distribution of clades. These probabilities can be calculated from a collection of gene trees, or from trees yielded during a bootstrap analysis, or during an MCMC analysis. Importantly, gene tree uncertainty is accounted for, and also unobserved gene tree topologies can be evaluated.

Horizontal gene transfer

In contrast to vertical inheritance of genes from parents to offspring, organisms can also incorporate foreign genes, or variant copies of foreign genes from distant relatives through a process termed horizontal gene transfer. A successful horizontal gene transfer event requires the genetic material to be successfully released by the donor, transported to the recipient, acquired and incorporated into the genome of the recipient, and expressed in a way that benefits the recipient. Several processes enable the required succession of hereditary events.

1. Transduction (Latin; *trans* - across, beyond; *duco*, to lead, to conduct) is the import of viral DNA through agents such as bacteriophages, probably even as a result of infection.
2. Conjugation by plasmids (Latin; *con*, together; *jugum*, yoke; “yoke together”) is horizontal transfer involving direct cell-to-cell contact, for example, through surface appendages with transfer of plasmids.
3. Transformation (Latin; *forma*, to shape, to form, to direct) is the uptake of free, extracellular DNA.
4. Gene transfer agents are bacteriophage-like elements integrated in the donor genome (Gogarten and Townsend, 2005; Stewart, 2013; Soucy et al., 2015). Gene transfer agents are sometimes under regulatory control by the donor. They package random DNA fragments from the donor and transport them to a recipient. Horizontal transfer by gene transfer agents differs from transduction in that, unlike bacteriophages, gene transfer agents are unable to transport all required genes to reproduce themselves.

Knowledge about the different forms of transfer is important because the ranges, and, consequently, their signatures differ.

Most interestingly, horizontal gene transfer played a key role in the identification of DNA as the molecular basis of inheritance. The famous experiment by Griffith (1928) showed that a non-virulent bacterial strain can incorporate genetic material of a heat-killed virulent strain, and subsequently cause disease. Avery et al. (1944) identified DNA to be a substance transferring genetic information by transformation. When antibiotic resistance spread unexpectedly fast across many different enteric strains (Davies and Davies, 2010; Davies, 1996), it was widely appreciated that horizontal gene transfer can not only be induced in the lab, but is of general importance in the evolution of bacterial genomes.

Direct observation of horizontal gene transfer happens rarely, and so, evidence of its occurrence needs to be collected independently from the traces that are left behind in the molecular sequences themselves. Of course, we expect that a horizontally transferred gene exhibits high resemblance between donor and recipient, and that the gene will be limited to the descendants of the initial donor and recipient. Especially if donor and recipient are distantly related, unduly high levels of resemblance between restricted sub-groups of otherwise unrelated species should capture the attention of methods detecting horizontal gene transfer.

Early studies collecting evidence for horizontal gene transfer analyzed the nucleotide compositions, and patterns of codon usage bias (Ochman et al., 2000). Genes with sequence characteristics departing significantly from the rest of the considered genome were classified as recent transfers. The detected amount of transferred DNA varied greatly between virtually 0 to nearly 17 percent in the 19 bacterial and archaeal genomes analyzed. This result is most likely an underestimation, because transfers from species with similar sequence characteristics cannot be detected. In fact, there is growing evidence, that horizontal gene transfer has played a major evolutionary role and has integrally shaped bacterial and archaeal genomes, as well as their diversification and speciation patterns. In fact, a significant proportion of bacterial and archaeal genetic diversity has been acquired through horizontal gene transfer (Abby et al., 2012; Lerat et al., 2005). The high frequency of horizontal gene transfer among prokaryotes can lead to phylogenetic relationships that are more net-like than tree-like. The importance of horizontal transfer in eukaryotes was still a topic of dispute until recently, when suspected examples of horizontally transferred genes were detected in metazoans including sponges, cnidarians, rotifers, nematodes, molluscs, arthropods (Boto, 2014), and even humans (Crisp et al., 2015). Additionally, significant amount of transfer was observed across the kingdom of fungi (Szöllösi et al., 2015). For reviews, refer to Daubin

3.1:8 The sources of Phylogenetic Conflicts

and Szöllősi (2016) or Husnik and McCutcheon (2017).

Finally, a transfer without recognizable homologs may be interpreted as gene origination. Possible reasons can be, for example, rapid divergence following the transfer event, gene loss in the donor species, erroneous homology search, or incomplete sampling and sequence availability.

Gene loss

Gene loss is the removal of existing genes from a gene family. On the one hand, gene loss can be a sudden mutational event caused, for example, by unequal crossing over during meiosis, or transposition of mobile elements. On the other hand, gene loss can be a slow process. Nonsense mutations creating truncated proteins or frameshifts, as well as missense mutations affecting crucial amino acid positions lead to the initial inactivation of a gene. The so-called *pseudogenization* is followed by a progression of deletion events with small fitness effect. Non-functional genes, may they be the result of inactivation, or non-functional duplicates are called pseudogenes. The number of pseudogenes can be large. For example, the human genome has nearly as many pseudogenes as functional genes (Lynch, 2007).

Gene loss greatly influences the gene content of genomes and contributes to the divergence of related species, next to other processes such as mutation. The potentially caused phenotypic diversity indicates that gene loss may be an adaptive evolutionary change (Albalat and Cañestro, 2016). Especially in bacteria and archaea, genome size is a strong fitness determining factor (*less-is-more* hypothesis). Likewise, unexpected high levels of intra-species variation of gene losses have been observed.

2.3 Hybridization

Occasionally, species evolve from hybrid crosses between two different ancestral species, an event termed inter-species hybridization. Inter-species hybridization is especially common in plants. The resulting genomic and phenotypic features reveal the two ancestral sources. Species histories exhibiting inter-species hybridization have rejoining branches, and as such, are not tree-like but correspond to more general networks called directed acyclic graphs. The respective gene trees can take various topologies depending on which gene copies are kept or lost.

Bread wheat is one of our most important staple crops and has been cultivated for more than ten thousand years. The evolution of bread wheat comprises several hybridization events (Pont et al., 2019). The genome of bread wheat consists of three closely related subgenomes, which are usually denoted (AABBDD). A first hybridization between wild *Triticum urartu* (AA) and a species of the *Aegilops speltoides* lineage (BB) produced an allotetraploid species, of which Durum wheat is a direct descendant. Subsequently, a second hybridization event with the wild *Aegilops tauschii* species (DD) formed the present day allohexaploid genome.

Further, analysis of transcriptome data revealed a pervasive ancient hybridization event in relatives of bread wheat (Glémin et al., 2019). Detection of hybridization events was performed using a hybridization index (Meng and Kubatko, 2009). The hybridization index measures the likelihood that a hybridization event, and not incomplete lineage sorting, is the cause of the observed phylogenetic conflicts. The analysis revealed that the overlooked wild species *Aegilops mutica* was involved in the first hybridization event leading to an allotetraploid (AABB) which further developed into bread wheat.

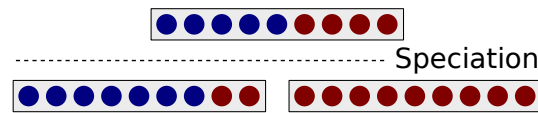


Figure 1 Incomplete sorting of alleles during a speciation event of a population of 9 individuals (circles). The blue and the red allele coexist at a specific locus in the population during the speciation and are incompletely sorted. The left daughter species contains two individuals with the red allele. At the considered locus, these two individuals are more closely related to the individuals in the right species than to the individuals containing the blue allele in their own species.

An ancient interspecies hybridization was also detected in the baker's yeast lineage (Marcet-Houben and Gabaldón, 2015). The authors propose that the resulting hybrid was forced to undergo a subsequent whole genome duplication to regain fertility.

Another form of hybridization is reassortment with viruses. Host cells simultaneously infected by two virus strains may assemble new viral particles whose origins are mixed. Some genetic material may originate from the first strain, other genetic material from the second.

2.4 Incomplete lineage sorting

In this section, we will describe a process fundamentally different from gene duplication and loss, horizontal gene transfer, or hybridization because it operates on the population level, and leads to conflict between orthologous genes. To facilitate the following exposition, we introduce the term allele which is a variant of a genetic sequence at a specific locus. Here, we mean different orthologous genes of the same species, but in a different context, alleles can also be different nucleotides, amino acids, or even whole chromosomes. The word allele is a short form of *allelomorph* (Gree; *allelo*, mutual, each other; *morph*, form) and emphasizes that we can discriminate between one or the other variant of an entity at the same locus in the genome of the considered individuals.

Different alleles can coexist in a population potentially for a long period of time spanning speciation events. Conceptually, a binary speciation sorts the alleles at a specific locus into the first or the second daughter species. Alleles with more variants fully participating in the speciation, maybe because they are partly causing the speciation, are completely sorted and no allele is present in both daughter species. In contrast, and because of recombination, coexisting alleles can be incompletely sorted during a speciation event such that they are present in both daughter species (Figure 1). For example, the individuals containing the red allele in the first daughter species are more closely related to individuals in the second daughter species than to individuals in their own species carrying the blue allele. Of course, because of recombination, the misleading affinity is only observed at this specific locus. When analyzing more loci, individuals within a species will be closer related to each other than to individuals of the other species (see Chapter 3.4 [Bryant and Hahn 2020]).

The process described above is referred to as *incomplete lineage sorting*. Incompletely sorted alleles coexisting over multiple consecutive speciation events may lead to phylogenetic conflict in that the corresponding gene tree supports a different topology than the species tree. For example, Scally et al. (2012) estimated that up to 30 percent of the genome of humans, chimpanzees, and gorillas has higher support for either one or the other of the two wrong species trees, and not for the correct species tree. That is, if we turn the argument around, only 70 percent of the genome of humans, chimpanzees, and gorillas supports the correct species tree.

3.1:10 The sources of Phylogenetic Conflicts

What are the factors determining the prevalence of incomplete lineage sorting? First, recombination is a prerequisite. Further, at a specific locus, incomplete sorting of alleles is likely when alleles coexist in a population, that is, when more alleles than a single one have higher frequency. A well-known measurement of genetic variation is the heterozygosity H of a locus, which is the probability of sampling two different alleles (e.g., Gillespie, 2004). For a haploid population with neutral variation, the heterozygosity is reduced per generation by *genetic drift*

$$\Delta H_D \approx -\frac{1}{N}H, \quad (1)$$

where N is the size of the considered population, and increased per generation by *mutations* happening with rate u per locus and generation.

$$\Delta H_M \approx +2u(1 - H). \quad (2)$$

Consequently, incomplete lineage sorting is prevalent if the population size and the mutation rate are large. Only then is sufficient variation generated compared to the rate at which it is removed by genetic drift. Moreover, incompletely sorted alleles need to coexist over multiple speciation events, in order to cause disagreement between a gene tree with the species tree. In this case, it is not enough that the population size is large, but also that the number of generations between speciation events is low. In particular, shorter internal branches of the species tree when measured in number of generations indicate higher prevalence of incomplete lineage sorting.

The name incomplete lineage sorting originates from the term *lineage* which denotes the line of descent from a common ancestor. The concept of a lineage is especially important when viewing the process backwards in time. Then, incomplete lineage sorting is manifested by lineages of alleles in the same species which do not coalesce within that species but only in an ancestor. For this reason, the term *deep coalescence* is often used to describe incomplete lineage sorting. Deep coalescence only leads to phylogenetic conflict, if the coalescent event involves a lineage ultimately leading to a different species.

3 Phylogenetic description

Development of appropriate models is key to understanding observations and collecting evidence for hypotheses. First, we remind ourselves that phylogenetic conflict is only an issue if genes are not co-transmitted from parents to offspring. For example, mitochondrial genomes are passed without recombination simplifying phylogenetic analysis. However, exclusive analysis of the mitochondrion is unsatisfactory, because the mitochondrion only contains a limited amount of genes and statistic errors are to be expected. In contrast, the eukaryotic nuclear genome contains a vast amount of informative sites but is continuously broken up by recombination.

Similarly, data sets including bacteria and archaea cannot be analyzed, because they also show recombination as a result of horizontal gene transfer combined with homologous recombination. The resulting presence of genes not being orthologs is a pervasive problem (e.g., Page, 2000). One approach to analyze data including genes not being co-transmitted is to concatenate putative orthologs and hope the corresponding phylogenetic signal outweighs the one of spuriously utilized paralogs, xenologs, and so on. Nevertheless, much data is ignored with this procedure. Full exploitation of data including recombining genes requires methods appropriate for analysis of paralogs (e.g., Page, 2000), ohnologs, xenologs, and homeologs (e.g., Szöllösi et al., 2012).

3.1 Homologous group sizes

Historically, search for homologous genes was only performed within species. In this section, the term *homologous group* will be used to denote a set of homologous genes within a species. We refrain from using the term gene family because gene families usually have members in several species. Early methods sought to describe the frequency distribution of the homologous group sizes which follows a power law characterized by long tails (Huynen and van Nimwegen, 1998). In particular, homologous groups with ten or more genes are more abundant than what would be expected from analyzing the frequency of homologous groups of low to moderate size (Szöllősi and Daubin, 2012). Remarkably, the distribution is very similar across bacteria, archaea, and eukaryotes, indicating shared universal features of the underlying processes responsible for the creation and removal of genetic material.

We can improve our understanding about the origins of the observed power law in the frequency distribution of homologous group sizes using stochastic linear birth and death processes (Yule, 1925).

► **Definition 1.** A linear birth and death process is a stochastic process that describes the evolution of an integer state variable. The considered system is a number $N_t \in \{0, 1, 2, \dots\}$ of units at time t evolving according to the following rules (Kendall, 1949; Bartholomay, 1958).

1. The sub-units generated by a unit develop in complete independence of each other.
2. A unit existing at time t multiplies by binary fission with probability $\lambda dt + o(dt)$, and dies with probability $\mu dt + o(dt)$ in the following time-interval of length dt .
3. All units in the population exhibit the same *birth rate* λ and the same *death rate* μ .

In detail, a birth event adds one unit to the population, and a death event removes one unit from the population. Waiting times until the next birth or death event are independently and identically distributed with exponential distributions with predefined birth and death rates λ and μ , respectively. The birth and death rates are shared across all units and independent of the number of units N_t . Linear birth and death processes are particular in that the waiting time until the next event necessarily becomes smaller the more units are present in the population.

In the context of our discussion, the units of the birth and death process are genes within a homologous group, birth events correspond to gene duplication or horizontal gene transfers events, and death events correspond to gene loss events. Of course, as described above, the instantaneous creation of identical gene copies in birth and death processes is only a rough approximation of biological gene birth and death. First, death rates are usually estimated to be higher than birth rates. Second, and more important, mathematical analysis shows that we cannot explain the observed power law with linear birth and death processes alone. That is, there is no set of birth and death parameters that can possibly explain the long tails of the frequency distribution of homologous group sizes. Rather, we need to relax the assumption of independence and employ general birth and death processes where the rates may depend on the total number of units in the population. We can only explain the long tails of the observed frequency distribution of homologous group sizes when letting the death rate, which is usually larger than the birth rate, decrease with the homologous group size so that it approaches the birth rate for large homologous groups (Karev et al., 2002).

The power law can also be obtained as the stationary distribution of a birth and death process with origination of homologous groups (Reed and Hughes, 2004). Specifically, this model is a superposition of two layers of stochastic processes. First, a stochastic process similar to a pure birth process, which has a death rate of zero, describes the origination of

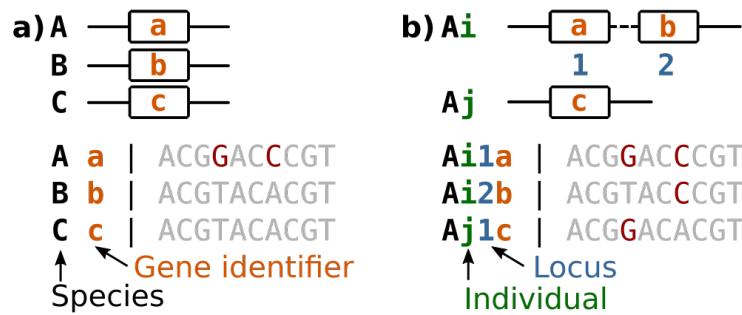
3.1:12 The sources of Phylogenetic Conflicts

homologous groups. Thereby, homologous groups randomly duplicate at a given rate. Naturally, the number of homologous genes may differ between homologous groups. Hence, this process involves non-identical units and is not a classical, birth and death process. Second, for each homologous group, a separate linear birth and death process with parameters λ and μ describes the number of genes in the respective homologous group. The stationary distribution of this two-layered process exhibits a stretched exponential if $\lambda \leq \mu$, and a power law if $\lambda > \mu$. A stationary distribution exists, because homologous groups are removed when the last gene dies. In conclusion, we fail to obtain the observed power law when treating genes within homologous groups independently. However, the frequency distribution of homologous group sizes can be described by relatively simple stochastic processes such as general birth and death processes, or using a two layered stochastic process modeling origination of homologous groups. Notice that both models treat different homologous groups independently.

At the same time, it is evident that the evolutionary histories of homologous groups of the same species but different gene families are correlated because of common decent along a shared species tree. Further, homologous groups of the same gene family and of closely related species are expected to have more similar sizes than homologous groups of the same gene family but from distantly related species. In this respect, a *gain-duplication-loss* model has been developed which uses a fixed species tree inferred from the concatenated alignments to assess the likelihood of homologous group sizes within gene families (Csűrös and Miklós, 2009; Csűrös, 2010). In the gain-duplication-loss model, gain is the pendant to horizontal gene transfer, but because only gene counts are considered, and not the gene trees themselves (see below), the origin of transferred genes is unknown. and all we can observe, is a gain in gene count. Ignoring the gene tree, may lead to biases in inference of gene transfer (Szöllősi et al., 2015). Rate variation can be accounted for with a discrete gamma distribution similar to the treatment with substitution models. Also, branch-wise gain, duplication and loss rates can be used. An application to archaea and bacteria shows that birth and death rates are similar across the two analyzed domains of life and that death rates seem to be larger than birth rates (Szöllősi and Daubin, 2012).

3.2 Multi-sequence alignments

More recently, advances in sequencing technologies and improvement of clustering and alignment algorithms used in homology search have led to the identification of numerous gene families with available sequence data. The analysis of the hereditary sequences themselves is much more intriguing than the analysis of their mere quantity. The sequence data is usually prepared in form of multi-sequence alignments (Figure 2) which contain a vast amount of information. In general, there is one multi-sequence alignment per gene family, and each sequence is labeled with the corresponding species and a label of their own, since there can be more genes per species. We are not concerned with species delimitation (Rannala and Yang 2013; Yang and Rannala 2010; Chapter 5.5 [Rannala and Yang 2020]) in this chapter. Typically, the alignment does not contain information about the homology relationships of the genes, and thus, about the loci. Here, we use loci in form of integers solely to encode the type of homology, and not relative position. That is, locus 1 is not necessarily left of or close to locus 2. However, genes at the same locus are assumed to be orthologous, and genes at different loci are assumed to be paralogous, ohnologs, xenologous, or homeologs. Further, common multi-sequence alignments of gene families only provide information from a reference genome and not about the genetic variation of genes. In detail, they do not contain sequences from different individuals. As a side note, the term *genome* has been used



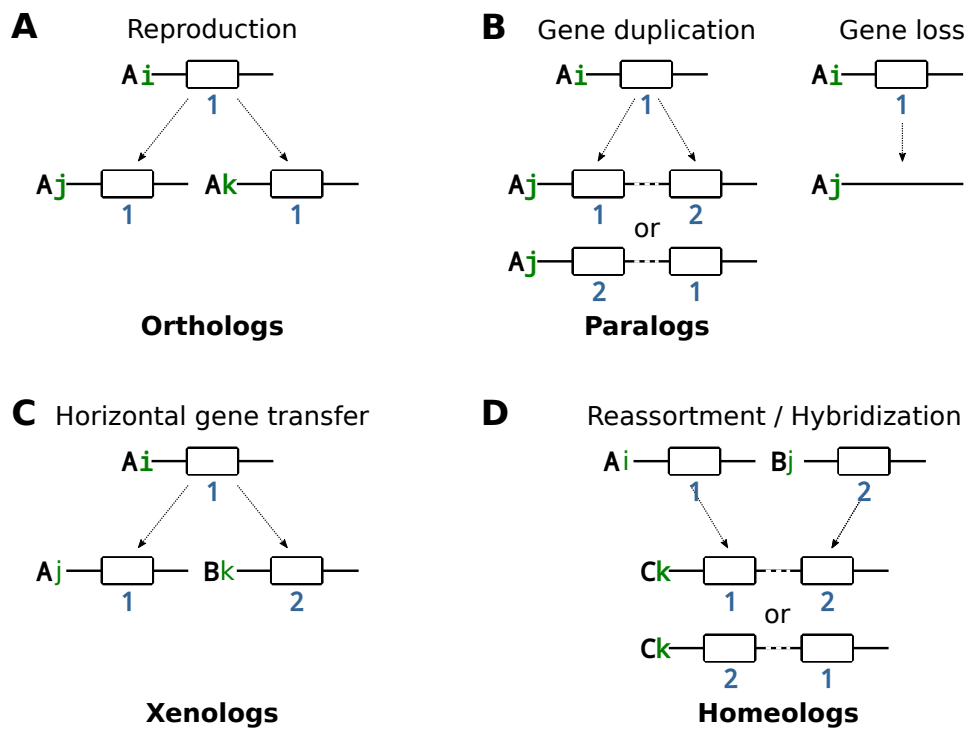
■ **Figure 2** Toy observation of a gene family with corresponding nucleotide multi-sequence alignment containing three genes of length ten. The horizontal lines represent genomes of sampled individuals and rectangles symbolize genes. The dashed line indicates that the genes may not be close to each other (in fact, they may be on different chromosomes). The genomes of the individuals and the genes are labeled with their corresponding species (capital black letter). The genes additionally have labels (orange lowercase letter). Nucleotide variants are emphasized in red. a) The alignment contains no information about the homology relationship of the genes. b) Here, species A contains three genes from two individuals, (green letter), as well as positional information about the locus (blue number). Usually, gene family alignments neither contain population data nor positional information.

in ambivalent ways. First, genome can refer to the complete genetic material present in an individual. Second, and in a more abstract way, genome can refer to the genetic material of a species, a set of species, or more generally, a set of individuals; for example, “the genome of humans” or “archaeal genomes”. Even though we are aware of this homonym, we were not able to completely avoid it. However, we aim to be explicit when describing the sources of phylogenetic conflicts in this section. That is why we focus on individuals themselves, for example, we examine “genes present in an individual” (and not in a genome).

For a given multi-sequence alignment of a gene family, we seek to coherently describe possible evolutionary histories. This task involves the construction of a gene tree, and most importantly, the event types at the nodes of the tree. We have already briefly discussed some types of events that can happen at gene tree nodes and which are accounted for by current evolutionary methods. For example, we can classify gene tree nodes as gene duplications or horizontal gene transfers, and add unobserved nodes corresponding to gene losses so that the observed gene tree agrees with the species tree; this process is called gene tree species tree reconciliation. Most gene tree, species tree reconciliation approaches involve two steps: (1) the reconstruction of gene trees with maximum likelihood methods, and (2) the reconciliation thereof using maximum parsimony methods minimizing the total number of gene duplications and losses (see Chapter 3.2 [Boussau and Scornavacca 2020]).

We will now present details on how gene duplication and loss, horizontal gene transfer, and hybridization affect the genes and genomes under consideration (Figure 3). Cell division during reproduction, or coalescence when viewed backwards in time, is the creation of two new individuals of the same species both of which contain the original gene. The locus of the genes remains unchanged. Gene duplication adds a copy of the same gene into the one produced daughter individual. The novel gene copy is inserted at a new locus. The genetic sequences of the copies are identical, and so we do not know which of the genes is to be found at the new locus. We may have to take this combinatorial fact into account during model design. Whole genome duplication corresponds to a concerted, massive gene duplication but can otherwise be described in the same way as simple gene duplication.

3.1:14 The sources of Phylogenetic Conflicts



■ **Figure 3** Depiction of how phylogenetic methods model cell division during reproduction, or coalescence when moving backwards in time, gene duplication and loss, and horizontal gene transfer. The horizontal lines represent genomes of considered individuals and the rectangles symbolize genes. The dashed line indicates may not be close to each other. The sequences of the genes are unaffected by the events. The individuals have a label (green letter) and are assigned to a species (black capital letter). The labels of the genes have been left blank. A hypothetical locus number is written below the genes in blue.

Horizontal gene transfer inserts a gene copy into a coexisting but foreign genome of a different species at a new locus. It has been argued that a horizontally transferred gene can take over the function of a previously existing gene in the recipient. In this case, the absence of purifying selection on the preexisting gene copy can induce divergence or loss of the preexisting gene copy. The result is an event called replacement transfer. There are models that exclusively allow replacement transfers because of their biological relevance and because a replacement transfer corresponds to a specific topological move called subtree pruning and regrafting (Hasić and Tannier, 2019).

Similar to the remark about whole genome duplications above, inter-species hybridization can be modeled like a concerted, massive horizontal gene transfer event combining two ancestral species. Whole genome duplications, and inter-species hybridization events are visible not only on the gene tree, but also on the species tree.

Finally, gene loss simply removes a gene from the genome. Gene loss is not directly observable in gene trees that we reconstruct, because branches leading to loss events are pruned from the tree. However, gene loss is an important constituent of phylogenetic models because, as we will see below, it can lead to phylogenetic conflict either together with the other discussed processes, or on its own.

A consequence of the considerations above is the following thought: If we had information about the loci of the genes in a given multi-sequence alignment, we could greatly reduce

the number of possible evolutionary histories explaining the data. Orthologs must be from different genomes but the same locus, gene duplication and loss involves genes from the same genome but different loci, and horizontal gene transfer involves genes from different species and possibly different loci. Note that turning this argument around, probabilistic models accounting for the events discussed above are informative about the homology relationships of the genes in the given multi-sequence alignment. Next to elucidating the gene tree, the detection of orthologs, paralogs, and xenologs is an important application (see Chapter 3.2 [Boussau and Scornavacca 2020]). Further, we only know of one model describing the actual synteny of loci, that is, the physical co-localization of loci (Delabre et al., 2018). Thereby, gene duplication and loss can affect segments spanning more than one locus.

4 Summary and discussion

In summary, the term gene loosely denotes a stretch of hereditary sequence passed on as a whole. Two genes with detectable shared ancestry are homologous to each other. A set of homologous genes is called a gene family. A gene family usually spans many species and can have more than one gene per species. The phylogenetic history of a gene family can be depicted in a gene tree. The lineage of a gene is the path from the gene, to the root of the gene tree. The type of homology of two considered genes is defined by the event happening when the lineages of the two genes join in the past. The types of homology we have discussed are (1) orthologous genes related by cell division during reproduction, (2) paralogous genes related by gene duplication, (3) ohnologous genes related by whole genome duplication, (4) xenologous genes related by horizontal gene transfer, and (5), homeologous genes related by inter-species hybridization.

Phylogenetic conflict is the complete or partial disagreement of a gene tree with the species tree. The species tree is usually not known, and so, phylogenetic conflict is either observed as disagreement about the species tree within a single gene tree, or disagreement between different gene trees. All discussed homology relations can cause phylogenetic conflict. For example, reproduction combined with recombination and mutation can lead to incomplete lineage sorting which is manifested by deep coalescing lineages maybe suggesting a misleading topology.

A prerequisite of conflict between gene trees is recombination. In contrast, co-transmitted genes will not show conflict. For this reason, genomic architecture such as chromosomes, plasmids or nuclear vs cytoplasmic compartments is an important factor. Doubts about exclusive usage of genes as phylogenetic units have been raised (Springer and Gatesy, 2018). For example, smaller units such as exons could be used. Further, knowledge about recombination patterns can help in discriminating between phylogenetic reconstruction errors and truly different gene trees (Reddy et al., 2017). As previously mentioned, conflicting histories between mitochondrial genes are unexpected.

The correct description of homology relationships in phylogenetic analyses is imperative, yet, the abundance of actual phylogenetic conflict is a matter of dispute. The importance of incomplete lineage sorting has been a topic of dispute for many years. After all, Scally et al. (2012) estimated that only 70 percent of the genomes of humans, chimpanzees and gorillas follow the correct species tree. High levels of incomplete lineage sorting are also reported in birds (Jarvis et al., 2014). It has been argued that phylogenetic conflict caused by incomplete lineage sorting may be wrongly estimated by the assumption that genes are passed on as a whole and by disregarding exon and intron structure (Springer and Gatesy, 2018; Mendes et al., 2019). In mammals, phylogenetic conflict caused by incomplete lineage sorting is minor

3.1:16 The sources of Phylogenetic Conflicts

when observing whole genes but conflicting histories are more pronounced when observing exons in a separate way (Scornavacca and Galtier, 2017). Further, it was shown that for species tree aware methods, the number of inferred gene duplications and horizontal gene transfers depends strongly on the used species tree (Szöllősi et al., 2013a). Although this is expected, caution is warranted when interpreting inferences involving phylogenetic conflict. In general, identification of systematic error as well as statistical error is difficult. For instance, it was postulated that phylogenetic conflict between orthologous mitochondrial genes may mostly be caused by statistical or systematic error (Richards et al., 2018).

Another issue is the relative importance of incomplete lineage sorting, gene duplication and loss, and horizontal gene transfer in causing phylogenetic conflict. The probability of incomplete lineage sorting is high when the number of generations between consecutive speciation events is low. If the average branch length measured in number of generations decreases from the root of the species tree towards the present, incomplete lineage sorting is more prevalent close to the present. If we assume that the species tree evolves according to a pure birth model (Yule tree, Yule, 1925), this assumption is not met since the average branch length on the tree is independent of the position of the branch on the tree (Stadler and Steel, 2012). There are no analytical solutions for the distribution of branch lengths for trees originating from a linear birth and death process (e.g., see Paradis, 2016). However, there is evidence from simulations that internal branch lengths increase compared to terminal branch lengths when increasing the death rate from zero towards values closer to the birth rate. This effect is more pronounced the closer the death rate is to the birth rate. Zhaxybayeva and Gogarten (2004) assume that the tree of life evolved according to a coalescent model. The coalescent model assumes that the total population size is constant and that the time to the next coalescence (moving back in time towards the root) is exponentially distributed with parameter $\binom{n}{2}$, where n is the number of sampled species. As a result, the average branch length increases towards the root. In this case, as well as for the birth and death process, the relative importance of incomplete lineage sorting decreases from the leaves to root of the species tree. Of course, we can only hypothesize about the distribution of branch lengths since we do not know the tree of life.

Spurious phylogenetic conflict can also arise if the reconstruction method suffers from systematic error (see Chapter 2.1 [Simion et al. 2020]). For example, across-site or across-gene composition heterogeneity (Lartillot and Philippe, 2004) can cause topological errors. In general, saturation of sequence distance can lead to long branch attraction artifacts (Felsenstein, 1978). Furthermore, decisions made during homology search may induce spurious phylogenetic conflict and greatly influence the identification of gene losses. In particular, if gene origination is not correctly detected, the lack of genes pertaining to the new gene family in neighboring species not affected by the gene origination, may be incorrectly attributed to massive genes loss. Similarly, undetected gene copies can be misinterpreted as gene losses (Page, 2000).

Another aspect related to this topic is the fact that in all phylogenetic analyses, most species remain unsampled. For this reason, trees originating from birth and death processes including the probability of incomplete sampling at the leaves of the tree have been analyzed (Stadler, 2009). In essence, the sampling probability corresponds to a transformation of the birth and death rates assuming complete sampling. Interestingly, a relatively simple calculation shows that we should expect the donors of most horizontally transferred genes in a considered data set to be members of either extinct or unsampled species (Szöllősi et al., 2013b).

The last example displays a big advantage of using birth and death processes in phylo-

genetic inference. The mathematical study of birth and death processes has a long tradition (Yule, 1925; Kendall, 1949; Bartholomay, 1958; Thompson, 1975; Gernhard, 2008; Stadler and Steel, 2019), and many properties have been derived analytically. For example, the expected number of units with time and the corresponding variance are known. Further, the probability density of birth events and the distribution of the time of origin of reconstructed trees have been derived (Gernhard, 2008). The knowledge can be summarized strikingly in so-called lineage through time plots, which show the average number of lineages of a tree evolving under the birth and death process with time. Further, the probability of death and the probability of no change within a given time can be used to calculate, for example, the probability of a gene tree evolving under the birth and death process within a constraining species tree. However, it is still difficult to simulate aforementioned gene trees conditioned on a specific number of genes per species and possibly also on the time of origin of the process. We can employ a forward process combined with rejection sampling, but computation times are immense for larger trees. The problem becomes even more difficult when accounting for horizontal gene transfer. The main reason is that the assumption of independence of the birth and death process is not met anymore. That is, the sub-units created by a birth event do not evolve in complete independence.

In any case, the combined treatment of incomplete lineage sorting, gene duplication and loss, and horizontal gene transfer is demanding but the unification of phylogenetic and population genetic models promises to improve both our understanding of, and our ability to reconstruct, the tree of life. In general, genetic change traverses three stages, (1) origin in a genome of an individual through mutation, (2) fixation in the respective population, and (3) maintenance in the population. For example, the dynamics of gene duplications differ slightly, in that the fates of the copies are tightly linked and determined by changes accumulated during, or after the fixation phase. Innan and Kondrashov (2010) argue that in order to understand these processes we have to examine the genetic variation of gene copies on the population level. Phylogenetic models, combining the description of gene duplication and loss, and horizontal gene transfer with the description of the evolution of variation in populations such as the multi-species coalescent model could play an important role in this respect. Especially, since they will allow more precise identification of the type of homology, which is an important piece of information that could, for instance, help resolve the complete species tree of animals (Pett et al., 2019).

In this respect, the three tree model by Rasmussen and Kellis (2012) has been seen as a considerable methodological advance (Du and Nakhleh, 2018). Additionally, it is now possible to infer hybridization events (Du et al., 2019; Elworth et al., 2019). Hybridization is when individuals of already diverged species successfully have offspring. The hybrids may be founders of a new, separate species transforming the species tree into a phylogenetic network. Hybridization can also be interpreted as a massive horizontal gene transfer which affects large parts of the genome. On the other hand, we have introduced a different three tree model that is more consistent but assumes no recombination between homologous genes on the same haplotype. Both three tree models use ultra-metric, time-like trees but rate modifiers can be used to account for the different molecular clocks. Also, the three tree models do not account for gene origination on the species tree. This may be problematic with respect to gene families limited to a few species. For example, consider two genes from the same homologous gene family that are present in Human and Elephant, but absent in all other mammals. Is this not already phylogenetic conflict caused solely by gene loss (or horizontal gene transfer)? Another type of event not discussed in this chapter is allopolyploidization. Allopolyploidization is the retention of both parental genomes in the offspring, an event that

is especially likely in plant species.

Phylogenetic trees and networks are the basis of a wide range of outstanding probabilistic methods. Even so, they can be accompanied by methods inspired by the machine learning community. For example, a K -means clustering algorithm was applied to data simulated using the three tree model of Rasmussen and Kellis (2012) to create a classifier that can identify orthologous regions (Knowles et al., 2018). Note that the classifier is not an independent method because it is trained on data simulated using the discussed probabilistic methods. Altogether, it is natural to expect that different gene families tell different stories about the species tree. It will be exciting to see how the treatment of the three most important causes of phylogenetic conflict can be combined in a successful and conclusive way.

References

- Abby, S. S., Tannier, E., Gouy, M., and Daubin, V. (2012). Lateral gene transfer as a support for the tree of life. *Proceedings of the National Academy of Sciences*, 109(13):4962–4967.
- Albalat, R. and Cañestro, C. (2016). Evolution by gene loss. *Nature Reviews Genetics*, 17(7):379–391.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–402.
- Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii. *Journal of Experimental Medicine*, 79(2):137–158.
- Bartholomay, A. F. (1958). On the linear birth and death processes of biology as markoff chains. *The Bulletin of Mathematical Biophysics*, 20(2):97–118.
- Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences*, 281(1777):20132450.
- Boussau, B. and Scornavacca, C. (2020). Reconciling gene trees with species trees. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.2, pages 3.2:1–3.2:23. No commercial publisher | Authors open access book.
- Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V., and Robinson-Rechavi, M. (2006). Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Molecular Biology and Evolution*, 23(9):1808–1816.
- Bryant, D. and Hahn, M. W. (2020). The concatenation question. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.4, pages 3.4:1–3.4:23. No commercial publisher | Authors open access book.
- Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics*, 9(12):938–950.
- Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A., and Micklem, G. (2015). Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, 16(1).
- Csűrös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912.
- Csűrös, M. and Miklós, I. (2009). Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Molecular Biology and Evolution*, 26(9):2087–2095.

- Daubin, V. and Szöllősi, G. J. (2016). Horizontal gene transfer and the history of life. *Cold Spring Harbor perspectives in biology*, 8(4):a018036.
- Davies, J. (1996). Origins and evolution of antibiotic resistance. *Microbiologia (Madrid, Spain)*, 12(1):9–16.
- Davies, J. and Davies, D. (2010). Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews*, 74(3):417–433.
- De Bodt, S., Maere, S., and Van de Peer, Y. (2005). Genome duplication and the origin of angiosperms. *Trends in Ecology & Evolution*, 20(11):591–597.
- Dehal, P. (2001). Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science*, 293(5527):104–111.
- Delabre, M., El-Mabrouk, N., Huber, K. T., Lafond, M., Moulton, V., Noutahi, E., and Castellanos, M. S. (2018). Reconstructing the history of syntenies through super-reconciliation. *Lecture Notes in Computer Science*, pages 179–195.
- Du, P. and Nakhleh, L. (2018). Species tree and reconciliation estimation under a duplication-loss-coalescence model. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB 18*.
- Du, P., Ogilvie, H. A., and Nakhleh, L. (2019). Unifying gene duplication, loss, and coalescence on phylogenetic networks. bioRxiv <https://www.biorxiv.org/content/10.1101/589655v1>.
- Elworth, R. A. L., Ogilvie, H. A., Zhu, J., and Nakhleh, L. (2019). Advances in computational methods for phylogenetic networks in the presence of hybridization. *Computational Biology*, pages 317–360.
- Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584.
- Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Biology*, 27(4):401–410.
- Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.
- Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769–778.
- Gillespie, J. H. (2004). *Population Genetics - A Concise Guide*. JHU Press, second edition edition.
- Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., and Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. *Science Advances*, 5(5):eaav9188.
- Gogarten, J. P. and Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9):679–687.
- Griffith, F. (1928). The significance of pneumococcal types. *The Journal of Hygiene*, 27(2):113–59.
- Hasić, D. and Tannier, E. (2019). Gene tree reconciliation including transfers with replacement is np-hard and fpt. *Journal of Combinatorial Optimization*, 38(2):502–544.
- Husnik, F. and McCutcheon, J. P. (2017). Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology*, 16(2):67–79.
- Huynen, M. A. and van Nimwegen, E. (1998). The frequency distribution of gene family sizes in complete genomes. *Molecular Biology and Evolution*, 15(5):583–589.

- Innan, H. and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics*, 11(2):97–108.
- Jarvis, E. D., Mirarab, S., [...], Gilbert, M. T. P., and Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.
- Karev, G. P., Wolf, Y. I., Rzhetsky, A. Y., Berezhovskaya, F. S., and Koonin, E. V. (2002). Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evolutionary Biology*, 2(1):18.
- Kendall, D. G. (1949). Stochastic processes and population growth. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):230–282.
- Knowles, D. G. and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, 19(10):1752–1759.
- Knowles, L. L., Huang, H., Sukumaran, J., and Smith, S. A. (2018). A matter of phylogenetic scale: Distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. *American Journal of Botany*, 105(3):376–384.
- Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I., and Koonin, E. V. (2002). Selection in the evolution of gene duplications. *Genome Biology*, 3(2):research0008.1.
- Lartillot, N. and Philippe, H. (2004). A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21(6):1095–1109.
- Lerat, E., Daubin, V., Ochman, H., and Moran, N. A. (2005). Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology*, 3(5):e130.
- Li, L., Stoeckert Jr., C. J., and Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.
- Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11):865–875.
- Lynch, M. (2007). *The origins of genome architecture*. Sinauer Associates Sunderland, MA.
- Marcet-Houben, M. and Gabaldón, T. (2015). Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker’s yeast lineage. *PLoS Biology*, 13(8):e1002220.
- Mendes, F. K., Livera, A. P., and Hahn, M. W. (2019). The perils of intralocus recombination for inferences of molecular convergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1777):20180244.
- Meng, C. and Kubatko, L. S. (2009). Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theoretical Population Biology*, 75(1):35–45.
- Meyer, A. and Schartl, M. (1999). Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Current opinion in cell biology*, 11(6):699–704.
- Meyer, A. and Van de Peer, Y. (2005). From 2r to 3r: evidence for a fish-specific genome duplication (fsgd). *Bioessays*, 27(9):937–945.
- Miele, V., Penel, S., and Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with silix. *BMC Bioinformatics*, 12(1).
- Ochman, H., Lawrence, J. G., and Groisman, E. A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304.
- Ohno, S. (1970). *Evolution by Gene Duplication*. Springer Science & Business Media.

- Page, R. D. (2000). Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, 14(1):89–106.
- Paradis, E. (2016). The distribution of branch lengths in phylogenetic trees. *Molecular Phylogenetics and Evolution*, 94:136–145.
- Pett, W., Adamski, M., Adamska, M., Francis, W. R., Eitel, M., Pisani, D., and Wörheide, G. (2019). The role of homology and orthology in the phylogenomic analysis of metazoan gene content. *Molecular Biology and Evolution*, 36(4):643–649.
- Philippe, H. and Forterre, P. (1999). The rooting of the universal tree of life is not reliable. *Journal of Molecular Evolution*, 49(4):509–523.
- Pont, C., Leroy, T., Seidel, M., Tondelli, A., Duchemin, W., Armisen, D., Lang, D., Bustos-Korts, D., Goué, N., Balfourier, F., et al. (2019). Tracing the ancestry of modern bread wheats. *Nature Genetics*, 51(5):905.
- Pupko, T. and Mayrose, I. (2020). A gentle introduction to probabilistic evolutionary models. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.1, pages 1.1:1–1.1:21. No commercial publisher | Authors open access book.
- Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.
- Rannala, B. and Yang, Z. (2013). Improved reversible jump algorithms for bayesian species delimitation. *Genetics*, 194(1):245–253.
- Rannala, B. and Yang, Z. (2020). Species delimitation. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 5.5, pages 5.5:1–5.5:18. No commercial publisher | Authors open access book.
- Rasmussen, M. D. and Kellis, M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765.
- Reams, A. B. and Roth, J. R. (2015). Mechanisms of gene duplication and amplification. *Cold Spring Harbor Perspectives in Biology*, 7(2):a016592.
- Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K.-L., Harshman, J., Huddleston, C. J., Kingston, S., et al. (2017). Why do phylogenomic data sets yield conflicting trees? data type influences the avian tree of life more than taxon sampling. *Systematic Biology*, 66(5):857–879.
- Reed, W. J. and Hughes, B. D. (2004). A model explaining the size distribution of gene and protein families. *Mathematical Biosciences*, 189(1):97–102.
- Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., and Thomson, R. C. (2018). Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic Biology*, 67(5):847–860.
- Robinson-Rechavi, M. (2020). Molecular evolution and gene function. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 4.2, pages 4.2:1–4.2:20. No commercial publisher | Authors open access book.
- Robinson-Rechavi, M., Marchand, O., Escriva, H., and Laudet, V. (2001). An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Current Biology*, 11(12):R458–R459.
- Scally, A., Dutheil, J. Y., [...], Tyler-Smith, C., and Durbin, R. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388):169–175.
- Scornavacca, C. and Galtier, N. (2017). Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66(1):112–120.

- Sidow, A. (1996). Gen(om)e duplications in the evolution of early vertebrates. *Current Opinion in Genetics & Development*, 6(6):715–722.
- Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.
- Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8):472–482.
- Springer, M. and Gatesy, J. (2018). Delimiting coalescence genes (c-genes) in phylogenomic data sets. *Genes*, 9(3):123.
- Stadler, T. (2009). On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *Journal of Theoretical Biology*, 261(1):58–66.
- Stadler, T. and Steel, M. (2012). Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *Journal of Theoretical Biology*, 297:33–40.
- Stadler, T. and Steel, M. (2019). Swapping birth and death: Symmetries and transformations in phylodynamic models. *Systematic Biology*, 68(5):852–858.
- Stewart, F. J. (2013). Where the genes flow. *Nature Geoscience*, 6(9):688–690.
- Szöllősi, G. J., Boussau, B., Abby, S. S., Tannier, E., and 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.
- Szöllősi, G. J. and Daubin, V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. In Anisimova, M., editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*, volume 856 of *Methods in Molecular Biology*, pages 29–51. Springer.
- Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V., and Boussau, B. (2015). Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 370(1678).
- Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. (2013a). Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912.
- Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. (2013b). Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397.
- Tautz, D. and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, 12(10):692–702.
- Thompson, E. A. (1975). *Human Evolutionary Trees*. Springer.
- Tiley, G. P., Barker, M. S., and Burleigh, J. G. (2018). Assessing the performance of ks plots for detecting ancient whole genome duplications. *Genome Biology and Evolution*.
- Van de Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nature Reviews Genetics*, 18(7):411–424.
- Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579.
- Yang, Z. and Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20):9264–9269.
- Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of dr. j. c. willis, f.r.s. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(402-410):21–87.

- Zhaxybayeva, O. and Gogarten, P. J. (2004). Cladogenesis, coalescence and the evolution of the three domains of life. *Trends in Genetics*, 20(4):182–187.
- Zwaenepoel, A. and Van de Peer, Y. (2019). Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Molecular Biology and Evolution*, 36(7):1384–1404.