Comparative Methods for Reconstructing Ancient Genome Organization

Yoann Anselmetti, Nina Luhmann, Sèverine Bérard, Eric Tannier, Cedric Chauve

Chapter 13

Comparative Methods for Reconstructing Ancient Genome Organization

Yoann Anselmetti, Nina Luhmann, Sèverine Bérard, Eric Tannier, and Cedric Chauve

Abstract

Comparative genomics considers the detection of similarities and differences between extant genomes, and, based on more or less formalized hypotheses regarding the involved evolutionary processes, inferring ancestral states explaining the similarities and an evolutionary history explaining the differences. In this chapter, we focus on the reconstruction of the organization of ancient genomes into chromosomes. We review different methodological approaches and software, applied to a wide range of datasets from different kingdoms of life and at different evolutionary depths. We discuss relations with genome assembly, and potential approaches to validate computational predictions on ancient genomes that are almost always only accessible through these predictions.

Key words Comparative genomics, Paleogenomics, Ancient genomes, Ancestral genomes

1 Introduction

Rearrangements were the first discovered genome mutations [1], long before the discovery of the molecular structure of DNA. Molecular evolutionary studies started with the reconstruction of the organization of ancient *Drosophila* chromosomes, from the comparison of extant ones [2]. However, it took almost 30 more years before the formal introduction of *paleogenetics*, as the field of reconstructing ancient genes [3]. Since then, the development of sequencing technologies and the availability of sequenced genomes has led to the introduction of *paleogenomics*, a field that aims at reconstructing ancient whole genomes using computational methods. The term paleogenomics can be understood in two ways: ancient genome sequencing [4], or the computational reconstruction of ancestral genome features, given extant sequences, offspring, and relatives [5]. We take it in the latter meaning, though we highlight several links between both interpretations.

João C. Setubal et al. (eds.), *Comparative Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1704, https://doi.org/10.1007/978-1-4939-7463-4_13,

Despite its early start as a molecular evolution problem, paleogenomics is still in its infancy. Whereas evolution by substitutions has been studied extensively from the 1960s, and has now well established mathematical and computational foundations, evolution by genome scale events such as rearrangements looks almost like a fallow field. Two reasons can be invoked. First, rearrangement studies require having fully assembled genomes, and genome assembly is still an extremely challenging problem, resulting in a small number of available genomes, compared to gene sequences for example. Second, the state space of sequence evolution is very small (4 possible nucleotides or 20 possible amino acids per ancestral locus), leading to computational problems that are much easier than the rearrangement ones, which work on the basically infinite discrete space of possible chromosomal organizations (gene orders for example). However, none of these reasons is biological, and recent progresses in technology and methodology are susceptible to quickly change this situation.

There have been tremendous methodological developments over the last 10–15 years. Standard and principled computational methods are now able to propose reconstructions of the organization of ancestral genomes over all kingdoms of life: mammals [6, 7], insects [8, 9], fungi [10], plants such as monocotyledons [11–13] (reviewed in [14]) and dicotyledons [13–16], bacteria [17–19]. Prospective ad hoc methods have attempted the reconstruction of more ancient animal proto karyotypes: amniotes [20–22], bony fishes [23–25], vertebrates [20, 21, 26], chordates [27], or even eumetazoa [28].

Here, we review some of the existing methods for reconstructing ancient gene orders, focusing on their methodological principles, strengths, and weaknesses. We detail the data preprocessing steps that are necessary to use these methods. We finally review the available software and give an insight on the possible validations of ancestral genomes.

2 Preliminaries: Material and Preprocessing

The starting material of comparative paleogenomics is composed of extant genome sequences and assemblies. These are often available in public databases such as Ensembl and the UCSC Genome Browser [29, 30]. A genome assembly is a set of linear or circular DNA sequences (we refer the reader to [31] for a recent review on genome assembly). Depending on the combination of the properties of the sequenced genomes (repeats in particular), the sequencing technology, and the assembly algorithm, the assembled sequences can be at various levels of completion, from full chromosomes (in which case the genome is said to be fully assembled) to scaffolds or contigs (fragmented assembly); for the sake of

exposition, we use here the term chromosome for an assembled contiguous DNA sequence. The fragmentation of extant genome assemblies has a significant impact on the quality of reconstructed ancestral genomes, which will be discussed in Subheading 3.4.

To reconstruct the organization of ancient genomes from the comparison of extant ones, it is first necessary to define sets of *markers* on extant genomes, that is, DNA segments defined by their coordinates on the genomes (chromosome or scaffold or contig, start position, end position, reading direction). Markers are clustered into families with the desired property that two markers in the same family are homologous over their whole length, and two markers from a different family show no or limited homology.

Gene families, available in some databases [29, 32], are good candidates for being markers, though intersecting genes and partial homologies can be a problem for certain methods. Markers can also be obtained by constructing synteny blocks from whole genome multiple alignments [33], Chapter 11, or by segmenting genomes according to pairwise alignments [34], or searching *ultra-conserved elements (UCEs)* [35] or virtual probes [36]. These methods are useful for example when considering genomes that exhibit low gene density.

Whether the considered markers are genes or other genomic markers, the identification of genomic marker families is both a fundamental initial step toward reconstructing ancestral genomes and a challenging computational biology problem, with links to sequence clustering, whole genome alignment, and phylogenetics, among others. There is currently no standard method or tool that is universally used and many applied works rely on ad hoc methods for this important preprocessing step.

Depending on combinatorial properties of the algorithms used to infer ancient genome organization from the comparison of extant genomes, several restrictions might need to be applied on families of genomic markers. Most methods require that no two markers overlap on a genome, as this might induce some ambiguity regarding their relative order along their chromosome. Other methods might also require that every genome contains at most one marker per family (*unique markers*) or at least one marker per family (*universal markers*), or both (unique and universal markers). Enforcing such constraints requires extra preprocessing of an initial marker set. Nevertheless, we consider now that we have obtained, for a set of extant genomes of interest, a dataset of genomic markers, that will serve as input to reconstruct the organization of one or several ancient genomes.

Eventually, a comparative approach requires phylogenetic information relating one or several ancestral species of interest to a set of extant species whose genome data are available. This information can range from a fully resolved *species phylogeny* with branch

lengths [6, 7], to a partition of the extant species in three non-empty groups that define a single ancestral species (two groups of descendant species and one group of outgroup species). So in the extreme case of considering a single ancestral species, a minimal dataset is composed of genome information for a set of three extant species, composed of two species whose last common ancestor is the ancestral species of interest and one outgroup [37].

3 Ancestral Reconstruction Methods

All methods consider a genome as a set of circular or linear orderings of markers, representing chromosomes or chromosomal segments. This implies that the exact markers' physical coordinates are transformed into a relative ordering of markers. It induces a loss of information which can have an influence on the result [38] but it is universally used. Then methods differ in their strategies: either they model the evolution of these arrangements of markers by evolutionary events such as duplications, losses, rearrangements, or they model the evolution of more local syntenic features/characters such as the physical proximity of sets of markers. In the following, we call *adjacency* (resp. *interval*) a pair of (resp. a set of at least three) markers that either occur contiguously along an extant genome or are assumed to occur contiguously along an ancestral genome.

The first strategy (evolution of whole genomes) quickly leads to computational tractability issues. The second strategy (evolution of local syntenic characters such as adjacencies and intervals) benefits from a standard evolutionary toolbox modeling the evolution of presence or absence of a character, and tractability issues are postponed to a final linearization step where local characters are assembled into chromosome scale arrangements of markers. Linearization procedures then benefit from standard algorithms originating from algorithms for computing physical maps of extant genomes [39].

3.1 Whole Genome Evolution

We first describe the approach that considers the evolution of genomes seen as sets of linear or circular orders of markers, i.e., roughly permutations that can possibly be separated into several chromosomes. Evolutionary events like inversions, translocations, transpositions, fissions, and fusions, all subsumed in the now standard Double Cut and Join (DCJ) model [40], are susceptible to alter these genomes. The reconstruction of ancestral genomes then aims, given marker orders representing extant genomes at the leaves of a species phylogeny, at assigning marker orders for all ancestral nodes, maximizing a mathematical criterion according to the chosen evolutionary model. Most of the time this criterion is the parsimony score, which is the minimum number of events

transforming a permutation into another [41], also called the distance, although some methods consider a likelihood criterion.

For most rearrangement models that do not include duplications, the distance between two genomes can be computed efficiently. But even the simplest non-pairwise ancestral genome reconstruction problem, the median problem reconstructing a genome minimizing the distance in a tree with only three leaves, is already NP hard [42]. Adding duplications makes all problems hard even for the comparison of two genomes [41]. Hence, with duplications considered, reconstructing rearrangement events that happened along the branches of a tree is not tractable either.

Heuristics for the ancestral genome reconstruction problem usually follow the strategy of assigning an initial genome arrangement to each internal node of the tree and then iteratively refining the solution by solving the median problem for internal nodes until no further improvement in the overall tree distance can be achieved. The implementation of GASTS [43] improves over previous methods applying this strategy by trying to find a good initial arrangement avoiding local optima. Using adequate subgraphs for heuristic assignment of the median, this method can handle multi chromosomal data with unique and universal markers. Another approach is based on the Pathgroup data structure [44] storing partially completed cycles in a breakpoint graph [41] for each branch in the phylogeny. Graphs are greedily completed and eventually form genomes at all internal nodes. This solution can be used as an initialization prior to local iterative improvements based on the median again using the Pathgroup approach. An interesting property of Pathgroup is that it can handle whole genome duplications. The method MGRA [45] on the other hand relies on a multiple breakpoint graph combining all extant genome organizations into one structure. MGRA then searches for breaks in agreement with the species tree structure transforming the breakpoint graph into an identity breakpoint graph. While MGRA requires unique and universal markers, it has recently been extended to handle unequal marker content [46]. More complex models of evolution have been considered, which include duplications for example [47, 48], but are tractable only under some specific condition, such as the hypothesis that rearrangement breakpoints are not re-used [47].

Some methods adopt a probabilistic point of view, like Badger [49], a software using Bayesian analysis under a model where circular genomes can evolve by reversals. It samples phylogenetic trees and rearrangement scenarios from the joint posterior distribution under this model by MCMC implementing different proposal methods in the Metropolis–Hastings algorithm. It is a similar local search to the heuristic on the minimization problem, but instead of giving a single solution without guarantee as an output, it provides a sample of solutions from a mathematically grounded

distribution. However, it faces the same tractability issues concerning the convergence time.

Finally, a simpler rearrangement distance is the Single-Cut-or-Join (SCJ) distance [50] that models cuts and joins of adjacencies. With this model the ancestral reconstruction becomes tractable. Ancestral genomes that minimize the SCJ distance can be computed efficiently using a variant of the Fitch algorithm [51] in polynomial time; however, constraints required to ensure linear or circular ancestral marker orders result in mostly fragmented ancestral genomes. In [52], a Gibbs sampler for sampling rearrangement scenarios under the SCJ model has been described. It starts with an optimal fragmented scenario obtained as described above and then explores the space of co-optima by repeatedly changing the scenarios of single adjacencies.

3.2 Genomes as Sets of Adjacencies and Intervals: Mapping Approaches

The linear or circular orders of markers can be seen as sets of adjacencies and intervals, instead of permutations. Then each adjacency or interval can be considered independently, as a separate syntenic feature, which evolves within the larger context of whole genomes. This independence assumption allows computing quickly ancestral states for adjacencies and intervals. The main problem is that the collection of ancestral adjacencies and intervals is not guaranteed to be compatible with a linear or circular ordering.

We describe here a family of approaches that focus on a single-ancestral genome and consist of two main steps, which are inspired by the methods initially developed to compute physical maps of extant genomes:

1. Genomes of related extant species are compared to detect common local syntenic features, such as marker adjacencies or intervals, that are then considered candidate ancestral features for the ancestral genome of interest. Common features are not necessarily conserved from an ancestor due to convergent evolution or assembly errors for example, so this method generates false positives. In some methods, each local syntenic feature is weighted, according to its pattern of presence/absence in extant species genomes, to represent a confidence measure in the hypothesis it is indeed an ancestral syntenic feature.
2. A maximum weight subset of the potentially ancestral local syntenic features (detected in the first step) is selected that is compatible with the genome structure of the considered ancestral species (linear/circular chromosomes, ancestral copy number of markers, etc.) and is then assembled into a more detailed ancestral genome map.

The case of unique markers: The initial applications [6, 7] of these physical mapping principles to ancestral genome organization reconstruction considered unique markers, i.e., markers that are

assumed to occur once and exactly once in the ancestral genome of interest.

In several methods [7, 10, 22, 53] step 1, the detection of common adjacencies and intervals and the inference of ancestral adjacencies and intervals, is implemented using a Dollo parsimony principle: any group of markers that are colocalized in two genomes of extant species whose evolutionary path in the species phylogeny contains the ancestral species of interest is deemed to be a potential ancestral syntenic feature. Here by *colocalized* we mean that the group of markers occur contiguously in both extant genomes regardless of their relative orders but without any other marker occurring in between; so the marker content of both occurrences of the colocalized group of markers in the extant genomes is identical while the marker orders can differ. Groups of two markers are adjacencies, while groups of more than two markers are intervals. Variations on the principle outlined above can be considered, such as relaxing the Dollo parsimony criterion or considering only adjacencies (*see* [6] for example) or considering probabilistic inference of ancestral adjacencies [54, 55].

Given a set of local ancestral syntenic groups, the second step aims at selecting a maximum weight subset of these groups that is compatible with the considered genome structure and does not contain any syntenic conflict, defined as a marker that is deemed adjacent to more than two other markers. Several methods such as Infercars [6] and MLGO [54] consider only marker adjacencies; these adjacencies define a graph whose vertices are markers and weighted edges represent adjacencies, and aim at computing a maximal set of weighted adjacencies that form a set of paths, each such path being then a linear order of markers called a *Contiguous Ancestral Region* (CAR). This problem is equivalent to a Traveling Salesman Problem (TSP) and is NP hard. It is addressed in [6] through a greedy heuristic and in [54] using a standard TSP solver. However, as shown in [56], if the linearity of CARs is relaxed and circular CARs are allowed, the optimization problem of selecting a maximum weight subset of adjacencies that forms a mix of linear and circular CARs is tractable and can be solved by reduction to a Maximum Weight Matching (MWM) problem.

When intervals are considered in addition to adjacencies, ancestral adjacencies and intervals can be encoded by a binary matrix, in the same way as hybridization experiments are encoded by binary matrices in physical mapping algorithms. The problem of extracting a conflict free maximum weight subset of adjacencies is then NP hard in all cases, i.e., even if a mix of circular and linear CARs is allowed. Traditionally, it is solved using either greedy heuristics or branch and bound algorithms (ensuring an optimal solution when they terminate). Moreover, when intervals are considered, CARs might not be completely defined and are represented using a PQ tree data structure that has been widely used in physical mapping

algorithms [39] and is related to the classical combinatorial concept of Consecutive Ones Property (C1P) (*see* [7] and references there). The software ANGES [53] and ROCOCO [57] are, so far, the only ancestral genome reconstruction methods that consider intervals of markers and encode CARs using PQ trees.

Last, when markers are assumed to be unique in the ancestral genome of interest but are subject to insertion or loss during evolution, the model of common adjacencies and intervals might be too stringent. In this case, the notions of *gapped adjacencies and intervals* were introduced that allows for some flexibility in the definition of conserved group of markers. However, this implies also that the C1P model is too stringent and needs to be relaxed into a *gapped C1P* model, in which optimization problems are NP hard [58, 59].

These approaches have been used on various datasets, including mammalian genomes [6, 7, 60], the amniote ancestor [22], fungi genomes [10], insect genomes [8], plant genomes [12, 16].

Non-unique markers. If markers exhibit varying copy numbers in extant genomes, they cannot be assumed to all occur once and only once in the considered ancestral genome. The first issue is then, for a given marker, to infer its ancestral copy numbers. This is a classical evolutionary genomics problem, for example to infer the gene content of an extinct genome. Given a model of gains and losses of markers, it is possible to compute a more likely ancestral content [61, 62], or content that minimizes the number of gains and losses [63], by a Dynamic Programming (DP) algorithm following the general pattern of the Sankoff-Rousseau algorithm [64].

Once copy numbers of ancestral markers, or bounds on such copy numbers, have been obtained, the two-steps approach outlined in the previous paragraphs can be applied: first, local syntenies (adjacencies and intervals) are detected using similar notions of adjacencies and intervals (we refer the reader to [65] for an overview of interval models when duplicated markers can occur) and are weighted according to their conservation pattern, and, in a second step, a maximum weight subset of local syntenies is computed that is compatible with the marker copy numbers. This second problem is known as the C1P with multiplicity (*mC1P*) and has been shown to be NP hard in general; the only tractable case requires considering only adjacencies and allowing an unbounded number of circular CARs [56, 66]. Moreover, when markers have a copy number higher than one and only adjacencies are considered, a conflict free set of adjacencies does not define unambiguously a set of CARs; this issue is similar to the well-identified problem of determining the location and context of repeats in genome assembly [67]. This issue can be addressed, at least partially, by considering intervals framed by non-repeats (*repeat spanning intervals*) as described in [68, 69]. Finally, when variation of copy numbers can be attributed

to Whole-Genome Duplications (WGD), specific methods based on a combination of gapped adjacencies and TSP algorithms have been proposed and applied to fungi and plant data [70].

3.3 Adjacency Evolution along Gene Phylogenies

We now discuss a variant of the approach described in the previous section, which still considers genomes as sets of adjacencies between markers, but assumes that evolutionary scenarios for marker families are also available and focuses on all ancestral genomes of the species phylogeny at once. Due to its similarity with traditional character-based phylogenetics, we rely on the standard phylogenetic vocabulary and call genomic markers *genes*. To summarize this approach, ancestral adjacencies are inferred, as previously, but using an optimization criterion and the available gene phylogenies both as a guide and a constraint.

Input: gene trees and adjacencies: This phylogeny-based approach requires as main input a fully binary rooted species phylogeny, and *reconciled phylogenies* for all gene families. This means that for all gene families, a rooted and annotated phylogenetic tree is required, depicting the whole history of the marker in ancestral and extant species in terms of speciations (S), duplications (D), transfers (T), or losses (L), where a transfer is the event of a species acquiring a genomic segment from another species (horizontal/lateral transfer). These reconciled gene trees can be obtained by several methods and software, depending on the set of evolutionary events one wants to consider (DTL or DL only), on the models and methods (parsimony or probabilistic approaches, joint or sequential reconstruction of the tree topology and reconciliation) [71–73]. Some databases also provide gene trees or reconciled gene trees [29, 32]. A reconciliation yields a presence pattern of ancestral genes in ancestral species. The leaves of these trees are the extant genes, and its internal nodes and events define ancestral genes.

The other information needed by the methods is the list of the gene adjacencies in the extant genomes. As defined above, we usually consider that two genes are adjacent if there is no other gene in that dataset between them, although here again relaxed notions of adjacencies can be considered.

Adjacency evolution: As genes and species, gene adjacencies also evolve. They can be gained, lost, duplicated, and transferred for example. The core element of the phylogeny-based methods we describe in this section is to infer the evolution of these adjacencies along the gene phylogenies, which themselves evolve within the species phylogeny. This leads to the inference of adjacencies between ancestral genes, i.e., ancestral adjacencies, and thus provides elements of the organization of genes in ancestral species. The currently available methods compute an evolutionary history of the adjacencies by either minimizing a discrete parsimony criterion or

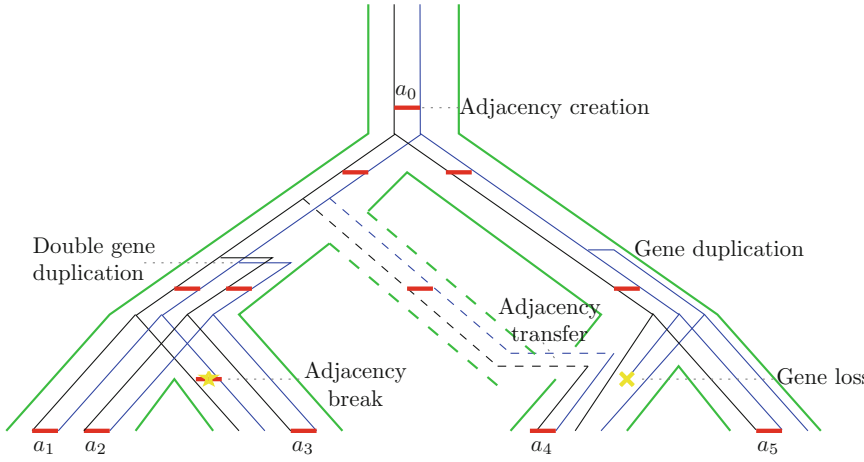

Fig. 1 Propagation of adjacencies (*red*) along gene phylogenies (*black* and *blue*) reconciled with a species phylogeny (*green*). This figure represents the evolutionary history of five extant adjacencies a_1 to a_5 sharing a common ancestor a_0 in agreement with the history of the genes present at their extremities. The double gene duplication on the *left side* induces an adjacency duplication, whereas the single-gene duplication on the *right side* does not. Events such as gene losses or rearrangements make adjacencies lost or broken

maximizing a likelihood within a probabilistic framework. The main difficulty in such methods is to infer adjacencies evolution scenarios that are consistent with the evolutionary history of the considered genes, encoded in their respective reconciled gene trees (*see* Fig. 1). The result of this approach, which considers each adjacency independently of all other adjacencies (like in the methods described in Subheading 3.2 but unlike those in Subheading 3.1), is a set of ancestral adjacencies for each ancestral species. As it is not guaranteed that these adjacencies are compatible with a linear structure, linearization methods such as [56] or global evolution methods such as [74] can be applied to infer valid ancestral gene arrangements, for each individual ancestor. This approach was followed in DupCAR [75], which imposes some constraints on the gene trees, and in the family of DeCo algorithms that we describe below.

DeCo algorithm family: The inference of adjacency histories is computed by Dynamic Programming techniques, implementing the rules of transmission of an adjacency from an ancestor to a descendant. As in the previous methods, at any point, a rearrangement can break or form an adjacency. But in addition, when a gene undergoes an event (Birth, Duplication, Loss, Transfer), an adjacency that has this gene as extremity necessarily changes: it can be gained, lost, duplicated, or transferred according to the evolutionary pattern of its extremities. The algorithm proceeds in three steps. A first step is to group adjacencies that may share a common ancestor in classes. Then, each class is examined independently using a DP algorithm that generalizes the Sankoff–Fitch parsimony

algorithms on binary alphabets; here the binary character is the presence or absence of an adjacency that evolves along pairs of reconciled gene phylogenies. Last, a backtrack step infers an unambiguous parsimonious evolutionary scenario for each adjacency.

This principle has been first implemented in parsimony when gene trees are reconciled in a duplication/loss model [76]. Following this initial model, several extensions have been proposed, which we outline briefly now. DeCoLT is an extension of DeCo that allows modeling the lateral transfers of genes between species, a frequent evolutionary event in bacterial evolution [18]. Two probabilistic extensions were recently introduced: in [9], the optimization criterion is a maximum likelihood criterion, while DeClone [77] implements a probabilistic approach to parsimony by allowing sampling evolutionary scenarios according to a Boltzmann-Gibbs probability distribution. Last, Art DeCo [78] has been introduced to handle fragmented extant genome assemblies (*see* the next section). DeCo and its variants all run in polynomial time allowing using them on large-scale datasets such as 69 eukaryotic genomes [76, 79].

3.4 Handling Fragmented Extant Genomes

Ideally, to reconstruct an accurate and complete organization of one (or several) ancestral genome(s) with a comparative approach, one would like to rely on the complete chromosomal organization of the considered related extant genomes. However, currently, most genome assemblies are incomplete and can even be highly fragmented[1]. This fact is due to the prevalence of sequencing technologies producing short and accurate reads that do not allow assembling repeated regions [67]. Recent improvements in sequencing technologies (for example long read sequencing protocols), as well as advances in processing methods (for example hybrid assemblies [80, 81] and gap closing methods [82–84]), make it possible to obtain the complete genome organization of microbial genomes [85]; however, the problem of genome assembly is still hard for large eukaryotic genomes [86].

Fragmented extant genome assemblies are characterized by the fact that chromosomes are split into several contigs or scaffolds, whose relative order and orientation is not known. This missing information on order and orientation of these scaffolds might hide conserved syntenies such as marker adjacencies, which leads to similarly fragmented ancestral genome organization. One can see the problem of reconstructing the organization of ancestral genomes as similar to genome mapping or scaffolding problems, in which case ancestral genome reconstruction and extant genome assembly can be considered a unique problem that consists in ordering genomic markers whether ancient or extant. The

[1] see the GOLD database for example https://gold.jgi.doe.gov/statistics.

algorithmic similarity between these two problems has been remarked [87] by noting a similarity between the breakpoint graph [41], used for the reconstruction of gene order in ancestral genomes, and the de Bruijn graph [88], used in genome assembly. This observation has led to the recent development of approaches aiming at improving extant genome assembly in an evolutionary framework that reconstructs jointly ancient genome organization.

This similarity was first exploited by Munoz et al. [89], to give an order and an orientation to scaffolds by contig fusion with the construction of the breakpoint graph of a reference genome and a target genome to assemble. The concept was taken further by Aganezov et al. [90]. They considered several related extant genomes (possibly at various levels of fragmentation) and applied simultaneous co-scaffolding of all extant genomes, under the hypothesis that fragmentation breakpoints are not the same (i.e., between the same markers) in all species and conserved syntenies can thus be detected, although with a weaker conservation signal. The core of their method is an extension of the classical breakpoint graph to more than two genomes [45, 46] and follows the parsimony principle on permutations (*see* Subheading 3.1). In consequence the method is limited to a small number of species (less than 10) and does not handle duplications.

Another alternative is an extension of the DeCo algorithm (*see* Subheading 3.3), called Art DeCo [77]. The method scaffolds several fragmented-related genomes by reconstructing gene adjacencies evolution. The method is based on a parsimony principle that considers gains and breaks of adjacencies, but also the cost of creating scaffolding adjacencies in extant genomes but is applied independently to each adjacency, thus avoiding the computational tractability issue of a parsimony approach on permutations. Art DeCo can handle a large number of species (several dozens) as well as gene duplications. The linearization issue however propagates to extant genomes: neither extant nor ancestral genomes are guaranteed to be compatible with a linear or circular structure, and linearization algorithms are needed as a post process.

3.5 Using Ancient DNA

In addition to extant genomes, ancient DNA (aDNA) extracted from archaeological or paleontological remains can provide direct evidence about the contents and structure of an ancient genome. Early works using aDNA concentrated on mitochondrial DNA not older than a few 1000 years, recovered for example from quagga [91], extinct moa [92], cave bears [93], or Neanderthal [94]. Later, advances in sequencing technologies and in aDNA recovery protocols [95] opened the way to the sequencing of nuclear aDNA in even older samples of bacteria like *Yersinia pestis* [96, 97] or mammals like the extinct woolly mammoth [98] or ancient horses [99].

However due to postmortem DNA decay and degradation by nucleases, only short fragments of aDNA can be recovered. Subsequently, the retrieved sequences are usually aligned to references and variants are identified keeping aDNA damage patterns in mind, precluding the analysis of more complex rearrangements between the ancient and extant genomes [100]. While a contig assembly based on such data can be expected to be quite fragmented, classical scaffolding approaches can often not be applied to aDNA data, due to the nature of the aDNA capture process for example. Hence, comparative phylogenetic methods following principles similar to the ancestral reconstruction methods described above have to be used to order and orient the obtained contigs. Combining aDNA sequencing data with comparative methods is therefore useful in two ways: scaffolding of a fragmented aDNA assembly while improving the reconstruction of other, probably older ancient genomes in the phylogeny. We outline this approach below.

Given sets of contigs from aDNA assemblies assigned to internal nodes of the species phylogeny, one first needs to define a common set of markers between the ancient contigs and extant genome sequences. Each family of markers should then consist of at least one ancient contig fragment and its occurrences in several extant genomes. An iterative segmentation approach based on mappings of ancient contigs to extant genomes is described in FPSAC [69] although other fragmentation or synteny blocks construction algorithms can also be applied [34, 101].

Once marker families have been obtained using aDNA and extant DNA data, the methods outlined in the previous sections can be applied directly. For example, the FPSAC method [69] computes copy numbers for markers using discrete parsimony, infers potential ancestral adjacencies using the Dollo parsimony principle, linearizes these adjacencies using the MWM algorithm introduced in [56], and clears ambiguities due to repeated markers using the algorithms of [68]. Moreover, as the set of markers is likely not covering the whole ancient genome, gaps between adjacent markers in scaffolds need to be filled. In FPSAC, the corresponding extant gaps are identified and their sequences are aligned. Then, for each column of the alignment, the parsimonious ancestral state is reconstructed with the Fitch algorithm [51]. This approach has been successfully applied to a set of aDNA contigs from the human pathogen *Yersinia pestis*, which was obtained from remains of victims of the Black Death pandemic in the fourteenth century [102].

3.6 Software

We review in Table 1 below the main existing software implementing the principles described in the previous sections.

3.7 Validation

Validation is a constant concern in evolutionary studies. Different hypotheses, different methods, and different types of data may lead to different results [103], and their quality is difficult to quantify.

Table 1
Main methods publicly available for ancient genome reconstruction

Name	Adjacencies Intervals Permutations	Parsimony (Pa) Probabilistic (Pr)	Insertions and losses	Duplications	Transfers	Exploration of alternative solutions and/or support of solutions
ANGES	A/I	Pa	Y	N	N	Y
FPSAC	A/I	Pa	Y	Y	N	N
DeCo*	A	Pa/Pr	Y	Y	Y	Y
DupCAR	A	Pa	Y	Y	N	N
ROCOCO	A/I	Pa	N	N	N	N
MGRA2	P	Pa	Y	N	N	N
MGLO	A	Pr	Y	Y	N	N
Badger	P	Pr	N	N	N	Y
GASTS	P	Pa	N	N	N	Y
Pathgroup	P	Pa	N	N	N	N
Infercars	A	Pa	N	N	N	Y

Col. 1 records the name of the method. Col. 2 indicates which type of method it implements, either genomes as permutations (Subheading 3.1), or genomes as sets of adjacencies and intervals (Subheadings 3.2 and 3.3). Col. 3 records if it uses a parsimony assumption or a probabilistic approach. Col. 4 indicates if the method allows unequal marker content in extant and ancestral species. Col. 5 indicates if the underlying evolution model considers gene duplication. Col. 6 indicates if the underlying evolution model considers gene transfers. Col. 7 indicate if alternative solutions can be provided (through sampling for example) or if there is a measure of support for features of the provided solution. References of the listed methods: ANGES [53], FPSAC [69], DeCo and variants [9, 18, 76–78], DupCAR [75], ROCOCO [57], MGRA2 [45, 46], MGLO [54], Badger [49], GASTS [43], Pathgroup [44], infercars [6]

Predictions concern events that can be up to 4 billion years old, and no DNA molecule is preserved, even in exceptional conditions, more than 1 million years. And even for the rare cases when ancient DNA is available, it is often not for ancestral genomes, and assembly issues make it hard to use it for validation purposes (*see* Subheading 3.5).

Theoretical considerations about the models and methods can help to assess the validity of the results. Agreement with widely accepted biological hypotheses, statistical consistency, computational complexity, clarity, and validity of the underlying hypotheses have to be discussed [104]. For example, a majority of the methods presented in this chapter are based on parsimony, which assumes that the possibility of convergence or reversion is negligible, while all statistical studies tended to show that it was not the case [105]. Models have to find a good balance between realism, consistency, and complexity. An important feature of a methodology is whether it is able to provide several alternative equivalent solutions

[43] (most of the time an optimal or a likely solution is not unique), or better, a sampling of possible solutions according to a likelihood [49]. At least, if this is not possible, statistical supports of local features such as ancestral adjacencies can provide a robustness [77] (*see* Col. 7 in Table 1).

Though it is not possible to travel in time, nor to replay the tape of evolution [106], it is possible to experimentally generate some lineages and test reconstruction methods on them [107, 108]. It has been realized for ancestral sequence reconstruction purposes, but it is very expensive, time consuming, and usually generates easy instances where all methods perform equally well. It has never been done for chromosome organization, although some experiments could theoretically be used as benchmarks [109].

Another validation technique is to compare the results with similar ones produced by independent data and techniques. For example, molecular evolutionary studies can compare their results with fossil data [110, 111]. Bioinformatics ancestral genome reconstructions have, for example, been compared with reconstructions from cytogenetics data [103]. But as for ancient sequences, each kind of protocol has caveats, and none can be considered as the truth.

The main validation tool remains simulation. Genome evolution can be simulated in silico for a much higher number of generations than in experimental evolution, at a lower cost. There are at least two issues that need to be considered for the simulation, where no general consensus exists: the set of operations applied, and the parameters (e.g., relative frequencies) of the different operations, if more than one type is used. Moreover, they are often designed by the team developing the inference method, and even if they are designed to be used by another team for inference [112, 113], they originate from a community interested in proving the validity of inference methods and are based on similar models that underly the reconstruction methods. Situations where the teams developing the inference methods and testing them are separated from the start are very rare [114] and, in their current state, existing testing schemes are not complex enough to be used for ancestral genome organization reconstruction yet. Nevertheless, this is likely an important aspect of ancient genome reconstruction methods that needs to be developed.

4 Conclusion—a Short User Guide

There has been an important effort, mostly over the last 10 years, in the development of computational methods for the reconstruction of ancestral genome organizations. Choosing a method among the many that are available requires considering several variables, such as the nature of available data, evolutionary properties of the considered lineages, computational infrastructures.

If a dataset is large (more than ~10 species), or if it contains many duplications that are deemed important to consider, it is better to look at methods that consider genomes as sets of adjacencies or intervals rather than permutations. The latter is appropriate for a small number of small genomes, provided duplicate markers can be ignored and a reasonable amount of computing power is available. In that case probabilistic methods as Badger should be preferred, because it proposes a sample of solutions based on grounded statistical principles, instead of a unique solution of a heuristic, but it is the most computationally intensive.

In all other cases, in our opinion, a local approach with adjacencies and intervals should be favored. If duplicates can be ignored (unique markers), ANGES is the most flexible tool, which allows retrieving most information (common intervals in addition to adjacencies). Otherwise, assuming duplicated markers are important and need to be considered, if good gene or marker phylogenies are available, the DeCo method and its variants are a natural choice providing the most comprehensive evolutionary scenarios. The choice of the variant depends if lateral transfers are considered, or the considered genomes are poorly assembled. In the absence of good reliable gene phylogenies, MGLO and FPSAC (used without aDNA data) are the only available methods.

Acknowledgment

C.C. is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant 249834. E.T., S.B., and Y.A. are funded by the French Agence Nationale pour la Recherche (ANR) through PIA Grant ANR-10-BINF-01-01 "Ancestrome". N.L. is funded by the International DFG Research Training Group GRK 1906/1.

References

1. Sturtevant AH (1921) A case of rearrangement of genes in drosophila. Proc Natl Acad Sci U S A 7:235–237
2. Dobzhansky T, Sturtevant AH (1938) Inversions in the chromosomes of drosophila pseudoobscura. Genetics 23:28–64
3. Pauling L, Zuckerkandl E (1963) Chemical paleogenetics. Acta Chem Scand 17:S9–S16
4. Poinar HN, Schwarz C, Qi J et al (2006) Metagenomics to paleogenomics: large–scale sequencing of mammoth DNA. Science 311:392–394
5. Muffato M, Roest Crollius H (2008) Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays 30:122–134
6. Ma J, Zhang L, Suh BB et al (2006) Reconstructing contiguous regions of an ancestral genome. Genome Res 16:1557–1565
7. Chauve C, Tannier E (2008) A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Comput Biol 4:e1000234
8. Neafsey DE, Waterhouse RM, Abai MR et al (2015) Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 anopheles mosquitoes. Science 347:1258522

9. Semeria M, Tannier E, Guéguen L (2015) Probabilistic modeling of the evolution of gene synteny within reconciled phylogenies. BMC Bioinformatics 16(Suppl 14):S5
10. Chauve C, Gavranovic H, Ouangraoua A et al (2010) Yeast ancestral genome reconstructions: the possibilities of computational methods II. J Comput Biol 17:1097–1112
11. Sankoff D, Zheng C, Wall PK et al (2009) Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. J Comput Biol 16:1353–1367
12. Murat F, Xu JH, Tannier E et al (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. Genome Res 20:1545–1557
13. Ming R, VanBuren R, Wai CM et al (2015) The pineapple genome and the evolution of CAM photosynthesis. Nat Genet 47:1435–1442
14. Salse J (2016) Ancestors of modern plant crops. Curr Opin Plant Biol 30:134–142
15. Murat F, Louis A, Maumus F et al (2015) Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol 16:262
16. Murat F, Zhang R, Guizard S et al (2015) Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. Genome Biol Evol 7:735–749
17. Wang Y, Li W, Zhang T et al (2006) Reconstruction of ancient genome and gene order from complete microbial genome sequences. J Theor Biol 239:494–498
18. Patterson M, Szöllősi G, Daubin V et al (2013) Lateral gene transfer, rearrangement, reconciliation. BMC Bioinformatics 14(Suppl 15):S4
19. Darling AE, Miklós I, Ragan MA (2008) Dynamics of genome rearrangement in bacterial populations. PLoS Genet 4:e1000128
20. Kohn M, Högel J, Vogel W et al (2006) Reconstruction of a 450–my–old ancestral vertebrate protokaryotype. Trends Genet 22:203–210
21. Nakatani Y, Takeda H, Kohara Y et al (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. Genome Res 17:1254–1265
22. Ouangraoua A, Tannier E, Chauve C (2011) Reconstructing the architecture of the ancestral amniote genome. Bioinformatics 27:2664–2671
23. Jaillon O, Aury JM, Brunet F et al (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto–karyotype. Nature 431:946–957
24. Woods IG, Wilson C, Friedlander B et al (2005) The zebrafish gene map defines ancestral vertebrate chromosomes. Genome Res 15:1307–1314
25. Catchen JM, Conery JS, Postlethwait JH (2008) Inferring ancestral gene order. Methods Mol Biol 452:365–383
26. Naruse K, Tanaka M, Mita K et al (2004) A medaka gene map: the trace of ancestral vertebrate proto–chromosomes revealed by comparative gene mapping. Genome Res 14:820–828
27. Putnam NH, Butts T, Ferrier DEK et al (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071
28. Putnam NH, Srivastava M, Hellsten U et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94
29. Herrero J, Muffato M, Beal K et al (2016) Ensembl comparative genomics resources. Database 2016:bav096. https://doi.org/10.1093/database/bav096
30. Speir ML, Zweig AS, Rosenbloom KR et al (2016) The UCSC genome browser database: 2016 update. Nucleic Acids Res 44: D717–D725
31. Nagarajan N, Pop M (2013) Sequence assembly demystified. Nat Rev Genet 14:157–167
32. Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perrière G (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10(Suppl 6):S3
33. Sankoff D, Nadeau JH (2003) Chromosome rearrangements in evolution: from gene order to genome sequence and back. Proc Natl Acad Sci U S A 100:11188–11189
34. M. Višnovská, T. Vinar, and B. Brejová (2013) DNA sequence segmentation based on local similarity. In: ITAT 2013 Proceedings, pp. 36–43
35. Dousse A, Junier T, Zdobnov EM (2016) CEGA–a catalog of conserved elements from genomic alignments. Nucleic Acids Res 44: D96–D100
36. M. Belcaid, A. Bergeron, A. Chateau, et al. (2007) Exploring genome rearrangements using virtual hybridization. In: APBC'07: 5th Asia–Pacific bioinformatics conference, Imperial College Press 2007, pp. 205–214

37. Kim J, Larkin DM, Cai Q et al (2013) Reference–assisted chromosome assembly. Proc Natl Acad Sci U S A 110:1785–1790
38. Biller P, Gueguen L, Knibbe C, Tannier E (2016) Breaking good: accounting for the fragility of genomic regions in rearrangement distance estimation. Genome Biol Evol 8 (5):1427–1439
39. Alizadeh F, Karp RM, Weisser DK et al (1995) Physical mapping of chromosomes using unique probes. J Comput Biol 2:159–184
40. Yancopoulos S, Attie O, Friedberg R (2005) Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21:3340–3346
41. Fertin G (2009) Combinatorics of genome rearrangements. MIT Press, Cambridge
42. Tannier E, Zheng C, Sankoff D (2009) Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics 10:120
43. Xu AW, Moret BME (2011) GASTS: parsimony scoring under rearrangements. In: Algorithms in bioinformatics. Springer, Berlin Heidelberg, pp 351–363
44. Zheng C, Sankoff D (2011) On the PATHGROUPS approach to rapid small phylogeny. BMC Bioinformatics 12(Suppl 1):S4
45. Alekseyev MA, Pevzner PA (2009) Breakpoint graphs and ancestral genome reconstructions. Genome Res 19:943–957
46. Avdeyev P, Jiang S, Aganezov S et al (2016) Reconstruction of ancestral genomes in presence of gene gain and loss. J Comput Biol 23:150–164
47. Ma J, Ratan A, Raney BJ et al (2008) The infinite sites model of genome evolution. Proc Natl Acad Sci U S A 105:14254–14261
48. Paten B, Zerbino DR, Hickey G et al (2014) A unifying model of genome evolution under parsimony. BMC Bioinformatics 15:206
49. D. Simon and B. Larget (2004) Bayesian analysis to describe genomic evolution by rearrangement (BADGER), version 1.02 beta, Department of Mathematics and Computer Science, Duquesne University
50. Feijao P, Meidanis J (2011) SCJ: a breakpoint–like distance that simplifies several rearrangement problems. IEEE/ACM Trans Comput Biol Bioinform 8:1318–1329
51. Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Biol 20:406–416
52. Miklós I, Smith H (2015) Sampling and counting genome rearrangement scenarios. BMC Bioinformatics 16(Suppl 14):S6
53. Jones BR, Rajaraman A, Tannier E et al (2012) ANGES: reconstructing ANcestral GEnomeS maps. Bioinformatics 28:2388–2390
54. Hu F, Zhou J, Zhou L et al (2014) Probabilistic reconstruction of ancestral gene orders with insertions and deletions. IEEE/ACM Trans Comput Biol Bioinform 11:667–672
55. J. Ma (2010) A probabilistic framework for inferring ancestral genomic orders. In: Bioinformatics and biomedicine (BIBM), pp. 179–184
56. Maňuch J, Patterson M, Wittler R et al (2012) Linearization of ancestral multichromosomal genomes. BMC Bioinformatics 13(Suppl 19): S11
57. Stoye J, Wittler R (2009) A unified approach for reconstructing ancient gene clusters. IEEE/ACM Trans Comput Biol Bioinform 6:387–400
58. Maňuch J, Patterson M, Chauve C (2012) Hardness results on the gapped consecutive–ones property problem. Discrete Appl Math 160:2760–2768
59. Maňuch J, Patterson M (2011) The complexity of the gapped consecutive–ones property problem for matrices of bounded maximum degree. J Comput Biol 18:1243–1253
60. Gavranović H, Chauve C, Salse J et al (2011) Mapping ancestral genomes with massive gene loss: a matrix sandwich problem. Bioinformatics 27:i257–i265
61. Csurös M (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912
62. De Bie T, Cristianini N, Demuth JP et al (2006) CAFE: a computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271
63. Csűrös M (2013) How to infer ancestral genome features by parsimony: dynamic programming over an evolutionary tree. In: Models and algorithms for genome evolution. Springer, London, pp 29–45
64. Sankoff D, Rousseau P (1975) Locating the vertices of a steiner tree in an arbitrary metric space. Math Prog 9:240–246
65. Bergeron A, Chauve C, Gingras Y (2008) Formal models of gene clusters. In: Bioinformatics algorithms. John Wiley & Sons, Inc, Hoboken, pp 175–202
66. Wittler R, Maňuch J, Patterson M et al (2011) Consistency of sequence–based gene clusters. J Comput Biol 18:1023–1039
67. Treangen TJ, Salzberg SL (2012) Repetitive DNA and next–generation sequencing:

computational challenges and solutions. Nat Rev Genet 13:36–46

68. Rajaraman A, Zanetti J, Manuch J et al (2016) Algorithms and complexity results for genome mapping problems. IEEE/ACM Trans Comput Biol Bioinform 14 (2):418–430. https://doi.org/10.1109/TCBB.2016.2528239
69. Rajaraman A, Tannier E, Chauve C (2013) FPSAC: fast phylogenetic scaffolding of ancient contigs. Bioinformatics 29:2987–2994
70. Gagnon Y, Blanchette M, El Mabrouk N (2012) A flexible ancestral genome reconstruction method based on gapped adjacencies. BMC Bioinformatics 13(Suppl 19):S4
71. Nakhleh L (2013) Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol 28:719–728
72. Szöllősi GJ, Tannier E, Daubin V et al (2015) The inference of gene trees with species trees. Syst Biol 64:42–62
73. Jacox E, Chauve C, Szöllősi GJ et al (2016) ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony. Bioinformatics 32(13):2056–2058. https://doi.org/10.1093/bioinformatics/btw105
74. Luhmann N, Thévenin A, Ouangraoua A et al (2016) The SCJ small parsimony problem for weighted gene adjacencies. In: Bioinformatics research and applications. Springer, Berlin Heidelberg
75. Ma J, Ratan A, Raney BJ et al (2008) DUPCAR: reconstructing contiguous ancestral regions with duplications. J Comput Biol 15:1007–1027
76. Bérard S, Gallien C, Boussau B et al (2012) Evolution of gene neighborhoods within reconciled phylogenies. Bioinformatics 28: i382–i388
77. Chauve C, Ponty Y, Zanetti J (2015) Evolution of genes neighborhood within reconciled phylogenies: an ensemble approach. BMC Bioinformatics 16(Suppl 19):S6
78. Anselmetti Y, Berry V, Chauve C et al (2015) Ancestral gene synteny reconstruction improves extant species scaffolding. BMC Genomics 16(Suppl 10):S11
79. Duchemin W, Anselmetti Y, Patterson M et al (2017) DeCoSTAR: reconstructing the ancestral organization of genes or genomes using reconciled phylogenies. Genome Biol Evol 9:1312–1319
80. Koren S, Schatz MC, Walenz BP et al (2012) Hybrid error correction and de novo assembly of single–molecule sequencing reads. Nat Biotechnol 30:693–700
81. Antipov D, Korobeynikov A, McLean JS et al (2015) hybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32:1009–1015
82. Paulino D, Warren RL, Vandervalk BP et al (2015) Sealer: a scalable gap–closing application for finishing draft genomes. BMC Bioinformatics 16:230
83. Salmela L, Sahlin K, Mäkinen V et al (2016) Gap filling as exact path length problem. J Comput Biol 23:347–361
84. English AC, Richards S, Han Y et al (2012) Mind the gap: upgrading genomes with Pacific biosciences RS long read sequencing technology. PLoS One 7:e47768
85. Koren S, Phillippy AM (2015) One chromosome, one contig: complete microbial genomes from long–read sequencing and assembly. Curr Opin Microbiol 23:110–120
86. Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289
87. Lin Y, Nurk S, Pevzner PA (2014) What is the difference between the breakpoint graph and the de Bruijn graph? BMC Genomics 15 (Suppl 6):S6
88. Compeau PEC, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nat Biotechnol 29:987–991
89. Muñoz A, Zheng C, Zhu Q et al (2010) Scaffold filling, contig fusion and comparative gene order inference. BMC Bioinformatics 11:304
90. Aganezov S, Sitdykova N, AGC Consortium et al (2015) Scaffold assembly based on genome rearrangement analysis. Comput Biol Chem 57:46–53
91. Higuchi R, Bowman B, Freiberger M et al (1984) DNA sequences from the quagga, an extinct member of the horse family. Nature 312:282–284
92. Cooper A, Lalueza-Fox C, Anderson S et al (2001) Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. Nature 409:704–707
93. Stiller M, Baryshnikov G, Bocherens H et al (2010) Withering away–25,000 years of genetic decline preceded cave bear extinction. Mol Biol Evol 27:975–978
94. Krings M, Stone A, Schmitz RW et al (1997) Neandertal DNA sequences and the origin of modern humans. Cell 90:19–30
95. Marciniak S, Klunk J, Devault A et al (2015) Ancient human genomics: the methodology behind reconstructing evolutionary pathways. J Hum Evol 79:21–34

96. Rasmussen S, Allentoft ME, Nielsen K et al (2015) Early divergent strains of Yersinia Pestis in Eurasia 5,000 years ago. Cell 163:571–582
97. Wagner DM, Klunk J, Harbeck M et al (2014) Yersinia Pestis and the plague of Justinian 541–543 AD: a genomic analysis. Lancet Infect Dis 14:319–326
98. Miller W, Drautz DI, Ratan A et al (2008) Sequencing the nuclear genome of the extinct woolly mammoth. Nature 456:387–390
99. Orlando L, Ginolhac A, Zhang G et al (2013) Recalibrating Equus evolution using the genome sequence of an early middle pleistocene horse. Nature 499:74–78
100. Peltzer A, Jäger G, Herbig A et al (2016) EAGER: efficient ancient genome reconstruction. Genome Biol 17:1–14
101. Minkin I, Patel A, Kolmogorov M et al (2013) Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In: Algorithms in bioinformatics. Springer, Berlin Heidelberg, pp 215–229
102. Bos KI, Schuenemann VJ, Golding GB et al (2011) A draft genome of Yersinia Pestis from victims of the black death. Nature 478:506–510
103. Froenicke L, Caldés MG, Graphodatsky A et al (2006) Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? Genome Res 16:306–310
104. Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol 17:839–850
105. Durrett R, Nielsen R, York TL (2004) Bayesian estimation of genomic distance. Genetics 166:621–629
106. Gould SJ (1990) Wonderful life: the burgess shale and the nature of history. Norton, New York
107. Hillis DM, Bull JJ, White ME et al (1992) Experimental phylogenetics: generation of a known phylogeny. Science 255:589–592
108. R.N. Randall (2012) Experimental phylogenetics: a benchmark for ancestral sequence reconstruction. https://smartech.gatech.edu/handle/1853/48998
109. Barrick JE, Yu DS, Yoon SH et al (2009) Genome evolution and adaptation in a long--term experiment with Escherichia Coli. Nature 461:1243–1247
110. Romiguier J, Ranwez V, Douzery EJP et al (2013) Genomic evidence for large, long--lived ancestors to placental mammals. Mol Biol Evol 30:5–13
111. Szöllosi 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 U S A 109:17513–17518
112. Beiko RG, Charlebois RL (2007) A simulation test bed for hypotheses of genome evolution. Bioinformatics 23:825–831
113. Dalquen DA, Anisimova M, Gonnet GH et al (2012) ALF–a simulation framework for genome evolution. Mol Biol Evol 29:1115–1123
114. Biller P, Knibbe C, Beslon G, Tannier E (2016) Comparative genomics on artificial life. In: Computability in Europe, to appear. Springer, Cham