Review

Genomes as documents of evolutionary history

Bastien Boussau and Vincent Daubin

Université de Lyon; université Lyon 1; CNRS; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, F-69622 Villeurbanne, France

Genomes conceal a vast intricate record of their carriers' descent and evolution. To disclose this information, biologists need phylogenetic models that integrate various levels of organization, ranging from nucleotide sequences to ecological interactions. Rates of duplication and horizontal gene transfer, organism trees and ancestral population sizes can all be inferred through statistical models of gene family evolution and population genetics. Similarly, phylogenomics combined with other fields of natural sciences can reveal the nature of ancient phenotypes and paleoenvironments. These computationally intensive approaches now benefit from progress in statistics and algorithmics. In this article, we review the recent advances and discuss possible developments towards a comprehensive reconstruction of the history of life.

From sequences to life's history

Since Zuckerkandl and Pauling established DNA as a document of evolutionary history [1], many approaches have been devised to transform the information enclosed in genomes into knowledge of their history. Comparative genomics and phylogenetics have emerged as the keys to many biological questions, such as the identification of functional sequences in complete genomes, the principles of genome architecture, the reconstruction of life's evolution and the inference of ancestral phenotypes and population characteristics. The traditional approach to studying genomes, starting from raw sequences, goes through largely independent steps: annotation and inference of sequence homology based on similarity, alignment of homologous genes or genome segments, gene phylogeny and deduction of the underlying phylogeny of organisms (Figure 1). However, all these levels of analysis depend on each other and the failure to model this dependence can result in the accumulation of errors. For instance, gene trees are inferred from sequence alignments, but an alignment is actually the result of a gene history, and is usually built using a rough guide tree that can affect all subsequent results. Hence, designing methods that couple high quality alignment and phylogeny is necessary. Similarly, gene histories are generally inferred independently from each other, although for many biological reasons, some genes are expected to have related histories. The development of new statistical approaches, which couple inferences at different levels of analysis, allows explicit modeling of the influence of several evolutionary processes on the structure of data. In this article, we will specifically review these recent breakthroughs, from the assignment of sequence homology and the inference of organisms' phylogenies to mechanisms of genome evolution and the influence of the environment. Furthermore, we will anticipate and discuss further developments that are needed to reconstruct a comprehensive history of life.

Alignment and the hypothesis of homology

Comparative genomics has proved to be a very powerful tool for genome annotation. The first step in any comparison of genomes is the identification of homologous sequences, which relies on a search for sequence similarity. Although this step and the subsequent definition of gene

Glossary

Bayesian inference: Likelihood is the probability of the data given the model; Bayesian inference instead deals with the probability of the model given the data, also named 'posterior probability'. This posterior probability of a model is proportional to the product of the likelihood and of a 'prior probability'. Such a prior probability permits incorporation of exterior knowledge into an analysis; for instance, one could assume that the prior probability over the transition/transversion ratio in a particular dataset follows a uniform distribution on [1,10]. Contrary to Maximum Likelihood inference, the common practice in Bayesian inference is not to return parameter values of the highest posterior probability; instead, whole distributions of parameter values are returned. To obtain these distributions, MCMC techniques are often used.

Coalescence: The coalescence of two genes is their last common ancestor.

Heuristic: Contrary to an exact algorithm, a heuristic is an algorithm that has not been proved to provide the exact result. A heuristic is often faster than an exact algorithm.

Hidden Markov Model (HMM): Probabilistic model used to describe a succession of states by associating hidden states with observed ones; a Markov Model is used to describe transitions between these hidden states. Such hidden states can be 'intron', 'intergenic' or 'exon' for models predicting gene structure, 'slow' or 'fast' for models predicting evolutionary rate, or different tree topologies for models predicting gene trees or recombination.

Incomplete lineage sorting: Observed discrepancy between a gene tree and the organism tree, due to the conservation of ancestral polymorphisms in different species (trans-specific polymorphisms).

Markov Chain Monte Carlo (MCMC): Algorithm used to sample from a probability distribution, by building a Markov model whose equilibrium distribution is the desired probability distribution. This means that when the chain has been run for a sufficiently long time, each state is visited with a frequency equal to its probability.

Maximum Likelihood inference (ML): For a given probabilistic model with specific parameters and particular data, the Maximum Likelihood values of these parameters correspond to the values under which it is most probable that the model has generated the data.

Markov Model: Probabilistic model of a process in which the state at time $t+1$ only depends on state at time t, not at time $t-1$. Models of substitution assume the substitution process is Markovian: a substitution $x \rightarrow y$ does not depend on the state preceding x.

Trans-specific polymorphisms (TSPs): The sharing among species of alleles inherited from an ancestor. These alleles have diverged prior to speciation, so that gene trees reconstructed using these genes can be different from the organisms tree.

Corresponding author: Daubin, V. (daubin@biomserv.univ-lyon1.fr).

0169-5347/$ – see front matter doi:10.1016/j.tree.2009.09.007 Available online 31 October 2009

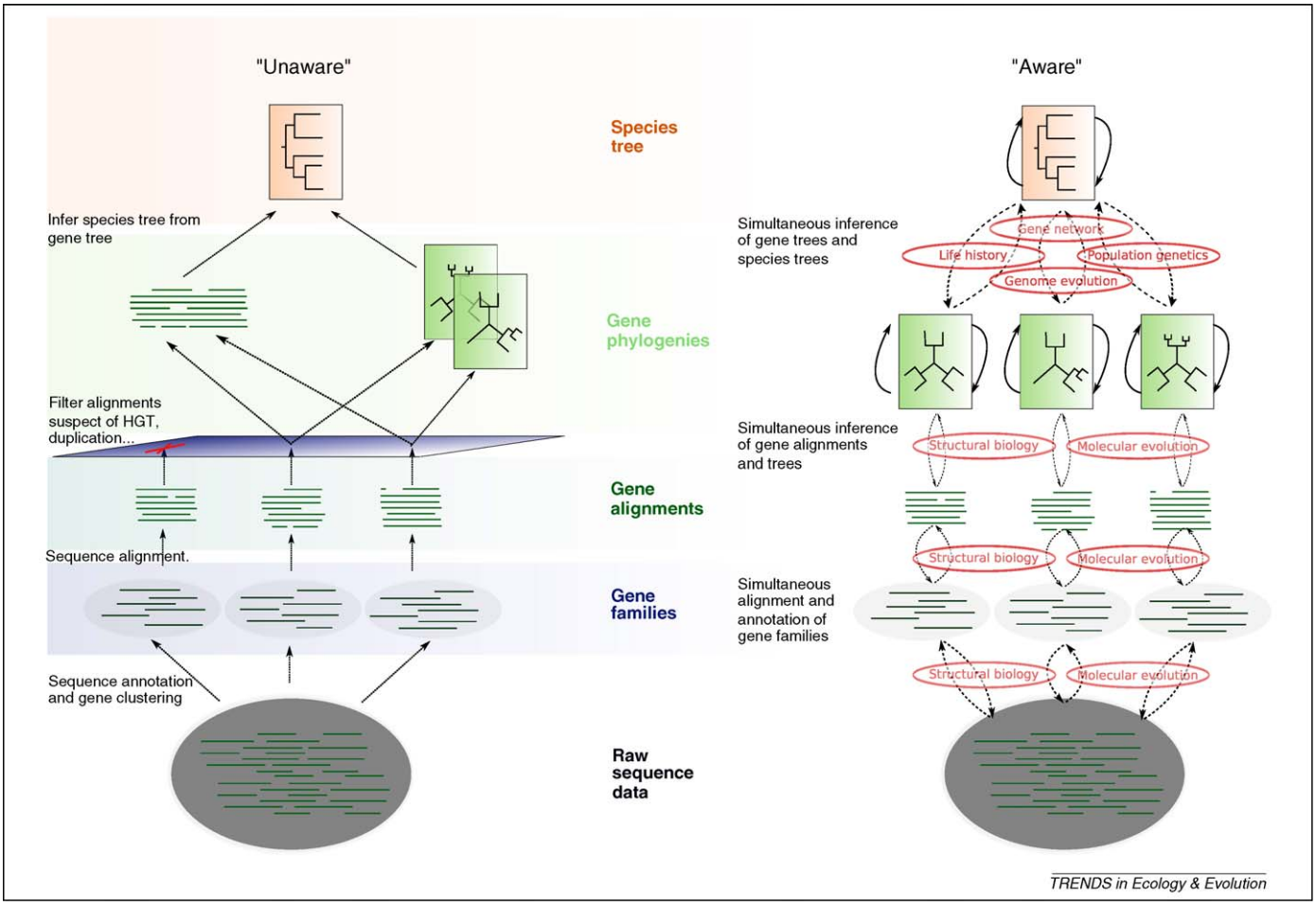

Figure 1. Phylogenetic awareness: the two paths from sequences to an organism tree. In the 'unaware' path (the traditional way of inferring species phylogenies) each stage of the phylogenetic inference is essentially independent from the steps upstream and downstream. In addition, sequence alignments have to pass different filters in order to make gene trees readily understandable as organism trees (absence of duplicates, lateral gene transfer (LGT), etc.). In contrast, the 'aware' path models the dependency and degree of complexity between each step using knowledge from different fields of biology (red ellipses, the list is not exhaustive). Alignments can be statistically estimated simultaneously with gene trees using models of sequence evolution that incorporate insertion and deletion events; and models of gene family evolution incorporating LGT, duplication and/or incomplete lineage sorting specify the dependency between gene trees and organism tree. Two-way arrows represent these dependencies, and solid arrows represent gene tree and organism tree searches. The dependency between gene family annotation, alignment and phylogenetics has not yet been explored, but could theoretically be modeled (see text for discussion). The schematic representation of the synchronous search for organism trees, gene trees, gene alignments and others suggests an obvious architecture for parallelizing this search.

families are not trivial (Box 1), one can usually find relatively coherent sets of homologous sequences to align. Alignment is the procedure by which the hypothesis of homology, defined at the level of the whole sequence, is refined to identify homologous sites by placing gaps at sites where insertions or deletions have occurred since the last common ancestor. This step is necessary both to identify sites under particular constraints and to reconstruct gene phylogenies. Usually, alignment is performed prior to other analysis and never questioned afterwards. However, the definition of homologous characters depends upon a description of phylogenetic relationships among sequences and because such a description is not available *a priori*, alignment algorithms first use 'quick-and-dirty' methods to obtain a low-quality phylogenetic tree. This tree influences all subsequent steps of the genomic analyses. Most alignment programs use heuristics (see Glossary) to place gap characters into sequences and the optimality of gap placement is assessed with respect to an arbitrary score, which differently penalizes gap insertions, gap extensions and substitutions. In the end, the alignment is the best estimate of the true alignment, according to arbitrary penalties that can be unrealistic for the data under study and to particular heuristics (e.g. [2]), which even if the penalties were perfectly tuned to the data, might not find the optimal alignment. Still, even if the optimal alignment is found, there is no guarantee that it is the true alignment.

These sobering considerations have long been known, and the limitations inherent in relying on a single alignment to analyze genomes and genes are now well accepted [3]. A statistically sound approach to this problem needs to: first, use a probabilistic model of insertion and deletion events combined with classical substitution matrices, so that parameters can be finely tuned to the data under study; and second, simultaneously estimate sequence alignments and phylogenetic trees, so that homology relationships are no longer based on a low-quality phylogenetic tree. A probabilistic model of pairwise alignment [4] can be seen as relying on Hidden Markov Models (HMMs) (or on the closely related transducers [5]), in which the possible states are 'match', 'insertion', and 'deletion'. Associating pairwise alignments based on HMMs to each

Box 1. Recombination and homology

The inference of homology relationships is usually done after genome annotation so that similar genes are grouped in the same family. However, considering gene families as the unbreakable bricks of phylogenetic reconstruction is incorrect. The shuffling of genetic material, through the processes of recombination and gene fusion frequently produces genes with mixed phylogenetic signals or even heterologous parts. Homologous recombination, the replacement of part of a sequence by a related sequence will result in gene alignments with contradictory phylogenetic signals over their length. Gene fusion will have a deeper impact, affecting the early steps of inferred gene homology, as only portions of protein sequences can be considered homologous. In any case, the reconstruction of a gene family history based on its entire length can be at best, partial, or worst, completely wrong. Although the only truly irreducible homologous character is the nucleotide, these events can comprise relatively long stretches of sequence, and the conflicting signals can be identified.

Many approaches have been devoted to identifying events of homologous recombination in multiple gene alignments, and recent models can simultaneously search for segment boundaries and histories in an alignment (e.g. [60]). Although this step has not been undertaken yet, an organism tree reconstruction model using multiple genes could be devised which includes these models of gene recombination.

Gene fusion and domain shuffling will probably be more complicated to model as they have an impact on the primary hypothesis of homology, i.e. the attribution of a protein to a family. Protein homology is typically inferred from overall similarity and several public databases propose automatically reconstructed gene families based on this criterion. Although the clustering method used to group proteins can vary, local similarities are usually dismissed, and protein sequences sharing homologous segments can be typically placed in different (heterologous) families. The high significance of this protein modularity, which can be viewed as a 'level of homology' problem, can be gauged from the fact that 19% of eukaryotic exons have undergone recombination with a non-homologous portion of the genome [61]. Interestingly, compared to entire proteins, domains should provide information on deeper phylogenetic relationships. An ideal way of dealing with this issue would be to couple the processes of homology assignment, sequence alignment and phylogenetic reconstruction into a model able to reconstruct and combine trees at different levels of homology.

branch of a phylogenetic tree permits easy computation of the likelihood of a multiple alignment [6], but integrating over the distribution of probable alignments and trees is computationally very intensive. Several recent algorithms address this issue by sampling both gene alignments and gene trees through a Bayesian Markov Chain Monte Carlo (MCMC) procedure (e.g. [7,8]). Alternatively, a maximum likelihood (ML) approach has also been proposed, and achieved both good accuracy and high speed [9] (for a list of available software, see Table 1). However, this approach considered gaps as missing data, and thus discarded the information that insertion–deletion events can provide. Using this algorithm jointly with the model of single site insertion or deletion proposed by Rivas and Eddy [10] might result in a program able to analyze hundreds of sequences in a reasonable amount of time, and making good use of the informational content of the sequences.

Bayesian joint estimations of sequence alignments and phylogenetic trees offer the possibility to better characterize the process of sequence evolution, as probabilities of insertions and deletions can be simultaneously estimated and compared with substitution probabilities. In contrast to most commonly used phylogenetic software, programs that simultaneously estimate alignments and phylogenies do not treat gaps as unknown characters, but can use insertions and deletions as phylogenetically informative events, which has been shown to improve phylogenetic reconstruction in a group of viruses [11]. As the rate of insertion and deletion is believed to be lower than substitution rates, their incorporation into phylogenetic reconstruction can also help resolve ancient divergences. Moreover, as a deleted character cannot be reinserted (otherwise it is not considered homologous), insertion–deletion events can impose a direction on a phylogenetic tree, and therefore point to its root, information otherwise difficult to extract from sequences [12].

In addition to phylogenetic reconstruction, other sequence-based inferences can benefit from averaging out the alignment. For instance, the detection of sites under positive selection has been shown to depend on alignment methods [3], and the reliability of protein structure prediction has been correlated to alignment quality [13]. Similar conclusions have been drawn for phylogenetic footprinting techniques that benefit from the comparison of several genomes to detect putatively functional regions, i.e. portions of sequences that are more conserved than expected and therefore must be under purifying selection [14]. However, relying on a single alignment affects the quality of inferences: when two different alignment algorithms were used to annotate transcription factor binding sites in 12 Drosophila genomes, less than 60% of the binding sites were predicted by both [15]. Consequently, Satija *et al.* [16] devised an algorithm to detect slowly evolving regions on distributions of alignments, which significantly improved binding site detection when compared to single alignments, especially when binding sites were not perfectly conserved.

Integrating gene trees and organism tree reconstruction

If sequence alignments benefit from being considered the result of sequence evolution along a gene phylogeny, the reconstruction of gene histories can also benefit from being considered the result of gene family evolution along an organism's phylogeny. Gene family evolution includes processes acting at the population as well as molecular levels (Box 2), such as trans-specific polymorphisms, gene duplication and loss, and lateral gene transfer (LGT). Models of gene family evolution allow a joint reconstruction of organism and gene phylogenies, yielding a better organism tree, better gene trees, in addition to estimates of ancestral population sizes, duplication, loss and transfer rates, and other insights [17,18].

This joint reconstruction can be achieved through a hierarchical structure, on top of which an organism tree is inferred from gene trees through models of gene family evolution. The gene trees themselves are inferred from sequence alignments through models of sequence evolution (Figure 1). The relationship between an organism tree and gene trees is bi-directional: given an organism tree some gene trees are more likely than others, while gene trees also inform the organism tree.

Table 1. Soft-aware

Software name	Description	Web link	Ref.
Alignment and phylogeny			
BAli-Phy (Bayesian Alignment and Phylogeny estimation)	Bayesian program to reconstruct alignments and phylogenetic trees.	http://www.biomath.ucla.edu/msuchard/bali-phy/index.php	[9]
StatAlign	Bayesian program to reconstruct alignments and phylogenetic trees.	http://phylogeny-cafe.elte.hu/StatAlign/	[69]
SimulFold	Bayesian program to reconstruct RNA structural alignment as well as phylogenetic trees.	http://www.cs.ubc.ca/~irmtraud/simulfold/	[70]
SAPF (Statistical Aligner, Phylogenetic Footprinter)	Bayesian program that samples alignments of non-coding sequences given a phylogenetic tree and predicts functional regions, i.e. regions that are particularly well conserved.	http://www.stats.ox.ac.uk/~satija/SAPF/	[16]
Dart (DNA, Amino and RNA Tests)	Software package to build and analyze alignments and phylogenetic trees through transducers notably, for sequences as well as RNA secondary structures.	http://biowiki.org/DART	[5]
Prank (Probabilistic Alignment Kit)	Phylogenetic-aware tool permitting the alignment of multiple sequences given a phylogenetic tree. Contrary to classical heuristics, it distinguishes insertions from deletions and thus has shown higher alignment accuracy.	http://www.ebi.ac.uk/goldman-srv/prank/prank/	
SATé (simultaneous alignment and tree estimation)	An automated method to quickly and accurately estimate both DNA alignments and trees with the maximum likelihood criterion [9].	http://www.cs.utexas.edu/~kliu/public/sate_journal.html	[2]
Species and gene trees			
Best (Bayesian Estimation of Species Tree):	Bayesian program to reconstruct species trees from gene alignments accounting for trans-specific polymorphisms.	http://www.stat.osu.edu/~dkp/BEST/	[22]
Bucky (Bayesian Untangling of Concordance Knots)	Bayesian program permitting analysis of several gene families simultaneously, accounting for some correlations between gene histories through gene-to-trees maps.	http://www.stat.wisc.edu/~larget/bucky.html	[39]
Prime	Set of software applications that can be used to analyze gene families in the presence of duplications and losses given a known species tree.	http://prime.sbc.su.se/	[25]
Inversions and phylogeny			
Badger (Bayesian Analysis to Describe Genomic Evolution by Rearrangement)	Badger is a Bayesian program to analyze genomic evolution through inversions.	http://badger.duq.edu/	[41]
Character evolution			
Sifter (Statistical Inference of Function Through Evolutionary Relationships)	Sifter predicts the function of genes in a gene family based on a model of function evolution and a phylogenetic tree of the gene family.	http://sifter.berkeley.edu/	[26]
BayesTraits	Bayesian program allowing one to analyze the evolution of discrete or continuous characters on a distribution of phylogenies.	http://www.evolution.reading.ac.uk/BayesTraits.html	[57]
Ape (Analysis of Phylogenetics and Evolution):	Package of functions to use in the R statistical software. Ape notably permits analyzing the evolution of discrete or continuous characters on a phylogeny, or studying shapes of phylogenies.	http://ape.mpl.ird.fr/	[71]

Several proposed models of gene family evolution focus on trans-specific polymorphisms, gene duplication and loss, or gene transfers (see Table 1). The relevance of these models depends on the organisms under study (e.g. closely related vs. distant organisms; high vs. low probability of LGT, etc.) but they can all be used for the joint reconstruction of gene and organism phylogeny as discussed below.

Gene family evolution and population genetics

Processes acting at the population level influence gene phylogenies through trans-specific polymorphisms (TSP) (Box 2). Recent theoretical and simulation studies using the coalescent model have shown that, in some conditions of population size and divergence time, most gene trees differ from the organism tree under which they were generated, and that simply using these as an estimate of the organism's tree is misleading [19]. More precisely, if the number of generations separating two speciations is not very large compared to the size of the populations between these two speciation events, it becomes likely that the coalescence of genes present in two species is more ancient than the previous speciations, which results in a

Box 2. The myth of 'orthologous gene families'

Coined by Walter Fitch [62], the term 'ortholog' designates genes that are related through speciation events, as opposed to 'paralogs', which are the result of duplications. Therefore, to reconstruct a phylogeny of species, one could use orthologous genes. However, the identification of orthologous genes is not always unequivocal (Figure I). First, phylogeneticists usually rely on the absence of duplicated copies in the datasets under study, but duplications could have occurred during the history of a gene family without leaving obvious traces. This is particularly dramatic in the event of reciprocal losses, when two species lose different copies of an ancestrally duplicated gene. The impact of this phenomenon, known as a hidden paralogy, is difficult to estimate on a large scale, but reciprocal losses have been shown to be frequent after whole genome duplications in yeasts and fish [63]. Second, lateral gene transfer (LGT) has been shown to be pervasive throughout the history of life. Therefore, it is unsafe to assume *a priori* that the history of a gene is devoid of such events, whatever its function. Third, even genes that would be considered genuine orthologs might not retrace the history of species; the persistence of different allelic forms of a gene during long periods of time relative to the lapse between speciation events, a phenomenon known as trans-specific polymorphisms (TSP) [20], can result in differences among gene trees (incomplete lineage sorting) even in the absence of paralogy or LGT. The assemblage of these processes makes it difficult to expect that a single gene history would faithfully mirror a tree of species throughout several billion years of evolution. In addition to these biological problems, even the most advanced phylogenetic methods are often unable to accurately model the evolution of biological sequences, which can result in the inference of erroneous trees. There is no, and will never be, a perfect dataset, devoid of lateral gene transfer, incomplete lineage sorting, hidden or apparent paralogies, convergent gene losses or systematically biased or accelerated evolutionary rates. As the impact of most of these processes is only expected to increase with more data, it is necessary to exploit the evolutionary significance of these events rather than discard them.

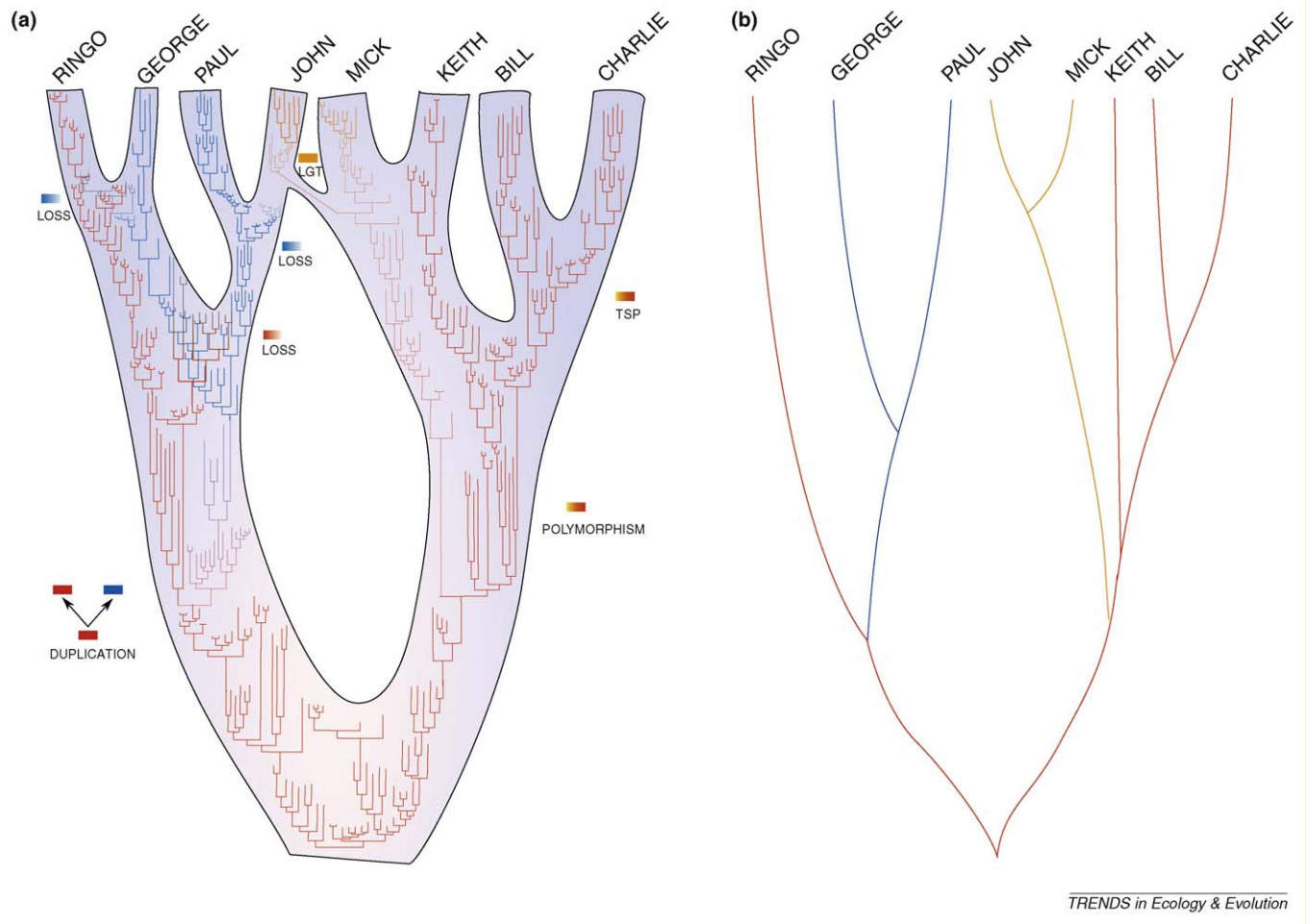

Figure I. Various processes can generate discordance between organism and gene trees. **(a)** A tree depicting the relationships of eight species. Ringo and George, Paul and John are on one side of the root, and Mick and Keith, Bill and Charlie on the other. The history of a gene family is depicted within the bounds of this organism tree, and processes acting at the genome level (duplication, loss, gene transfer) as well as population level (polymorphism) are shown. **(b)** The gene tree reconstructed from this gene family shows a topology that conflicts with the organism tree. Following a duplication and losses, George and Paul are grouped together, a gene transfer groups John and Mick, and trans-specific polymorphism leads to Keith being clustered with Bill and Charlie. Processes from population genetics and from genome dynamics both affect gene histories; models of gene family evolution could help reconstruct gene phylogenies, organism history and genome evolution.

gene tree different from the organism tree. Models of gene family evolution have been developed that use the amount and types of gene tree and organism tree incongruence to infer divergence times and ancestral population sizes; in some cases, they might be the only way to get a correct organism tree [19,20].

The first models incorporating TSP assumed a known species phylogeny and used gene trees to estimate divergence times and ancestral population sizes [21]. More recent models can also estimate the species phylogeny [22,23], with better results than classical methods such as gene concatenation. For a detailed discussion of these

issues, we direct the reader to a recently published review in TREE [20].

Gene family evolution, duplication and loss

The combined action of gene duplication and gene loss considerably complicates gene trees. Even a gene family with only one representative per species can harbor events of duplication and loss, yielding a gene tree different from the organism tree (Box 2). Models of gene family evolution taking into account duplications and losses have been proposed by Lagergren's group [24,25]. The evolution of a gene family is modeled by a birth–death process running along an organism tree: 'birth' corresponds to gene duplication, and 'death' to gene loss. In addition, a model of sequence evolution is used, so that given an organism tree and gene alignments, likelihoods of gene trees can be computed. This combined model has been implemented in a program that can estimate gene trees given an organism tree, through Bayesian MCMC integration, resulting in gene trees that are more biologically meaningful than if they had been inferred based on sequence alignments alone (see Table 1). This model also provides posterior probabilities of orthology and paralogy for each pair of genes, which could further aid functional prediction, as it is often assumed that function is better conserved between orthologs than between paralogs [26]. However, to date, these models use a single duplication rate and a single loss rate for all branches of an organism tree, even though it is known that different lineages undergo different rates of duplication and loss. Future models should cope with this heterogeneity of the evolutionary process to properly depict gene family evolution. Moreover, relying on a known organism tree for building accurate gene trees is certainly optimistic in several cases: just as the program of Akerborg *et al.* [25] integrates over scenarios of duplications and losses with respect to a given organism tree, a better statistical estimation of gene trees might be obtained by integrating over the distribution of organism trees. Such a model would provide an organism tree built using more than the genes that happen to be single-copies in most genomes, helping to resolve some difficult phylogenies, and at the same time clarifying the dynamics of genome expansion or shrinking over the entire tree.

Gene family evolution and lateral gene transfer

When modeling lineage sorting and duplication, it seems reasonable to consider that there exists an underlying organism tree, i.e. a tree depicting the history of vertical inheritance in the genome. However, the ability of prokaryotes to acquire genes by LGT has led to questioning whether the concept of an organism tree applies in such organisms [27–29]. Indeed, most, if not all, gene families have had time, during their history, to be transferred even among distant organisms, so that gene trees are different from the organism tree. However, at a given time, a gene can be inherited vertically along the organism tree, or laterally through LGT. Integrative models that account for both an organism tree and LGT would probably help to quantify the relative contributions of both vertical descent and lateral transfer to genome evolution.

Suchard [30] reached a first stage in the elaboration of a procedure that simultaneously searches for an organism tree and hundreds of gene trees under an LGT model. In his approach, gene trees can be produced from an organism tree through topological rearrangements mimicking LGT. These gene trees are then evaluated with respect to precomputed gene alignments and thus inform on the likelihood of the organism tree. The whole model was implemented in the Bayesian framework with MCMC sampling. This model converged on a unique organisms tree and confirmed previously reported results that informational genes, involved in the processing of genetic information, tend to undergo relatively low rates of transfer. However, the dimension of the topological space that can be reached through LGT simulation from a single organisms tree limited the applicability of the method to six species. Other approaches using fast algorithms for reconciling gene and organisms trees under a LGT model could help tackle this

Box 3. Computational challenges and numbers of parameters

It can be unreasonable to devise a model accounting for all the processes that contribute to the evolution of genomes all at once. First, the computational task would be immense and second, only a limited quantity of parameters can be estimated from a finite amount of data [64]. However, today's better algorithms and powerful statistical methods provide means for tackling integrative models.

For instance, up to now, the only described model that inferred an organism tree and multiple gene trees at the same time, only did so for a very modest number of species [30], as the computational task is enormous. Searching for a gene tree based on an alignment is already an intimidating task; searching for both an organism tree and several gene trees is even more difficult because for each organism tree, one needs to estimate and sample corresponding gene trees. There is however an obvious way to parallelize computations through architecture based on a server and several clients (Figure 1): a server node would search for an organism tree, while client nodes would search corresponding gene trees for each organism tree. Such parallelization would be necessary to compute trees based on whole genomes and could require hundreds or thousands of computers running for several days.

In addition to multiplying processors, it is also possible to make a better use of the resources available in the average personal computer: Suchard and Rambaut [65] recently developed new algorithms to use graphical cards when building phylogenetic trees. Thanks to the highly-parallel structure of the graphical processing units, they achieved up to 90-fold speed increases.

Bayesian MCMC (Markov Chain Monte Carlo) techniques can tolerate larger numbers of parameters than Maximum Likelihood (ML) approaches as MCMC integrates over the distributions of parameters when ML only uses point estimates: in this latter case, any small error over the value of a variable can snowball to drastically affect the accuracy of other estimates. However, different models will most likely have to be used depending on the question under scrutiny, where only the most relevant parameters are included: when there are too few parameters then estimates are biased, whereas too many parameters with large variances prohibit any conclusion. In this respect, issues of model selection will be particularly pressing. Recently, several works estimated the values of parameters, but also their numbers, through techniques such as Dirichlet Process priors [66], reversible jump MCMC methods [67], or Poisson processes [68]. Such techniques that auto-regulate their number of parameters will be necessary to use complex models with large amounts of data.

problem [31,32]. The research on LGT would strongly benefit from such a statistical framework, which does not take gene trees as error-less data, but as statistical estimates from gene sequences themselves. In addition, gene transfer events constitute informative characters for phylogenetic reconstruction [33], and provide relative dates for nodes of an organism tree: if a descendant from node A gives a gene to an ancestor of node B, this means that node B is more recent than node A. Considering the immense difficulty of dating nodes in the prokaryotic tree of life where fossils are scant and at best difficult to relate to extant species, a relative dating would certainly be highly valuable. Lastly, accurately reconstructing gene trees would also offer new possibilities to study the evolution of genome contents: up to now, ancestral genome content was obtained through comparisons of numbers of related genes in extant genomes [34,35], discarding all sequence information relative to gene histories. Using gene trees instead would probably greatly improve inferences.

A model of gene family evolution incorporating trans-specific polymorphisms, LGT, duplication and loss with an appropriate number of parameters (Box 3) could still be refined to account for dependencies between genes. For instance, two neighboring genes, from a bacterial operon or from a eukaryotic chromosome, as well as genes that interact functionally are more likely to share a similar history than genes from different regions of the genome [36,37]. Accounting for effects of spatial proximity on gene histories can be achieved through HMMs, as recently used by Hobolth *et al.* [38] to infer recombination hotspots, ancestral population sizes and divergence times. Another more general approach to model coevolution could use gene-to-tree maps [39]. Gene-to-tree maps are objects that associate genes with distributions of trees: two genes that have coevolved for a part of their history will show partially similar tree topologies. The degree of topological similarity will depend on the type of coevolution, which can be inserted into a statistical model through prior probabilities; for instance, two interacting genes are *a priori* more likely to share tree topologies than genes that are part of two separate pathways.

Genome dynamics

Reconstructing the evolution of genomes is not merely reconstructing the history of their genes. Events such as inversions, tandem duplications and chromosome fission and fusion also affect genomes, and therefore need to be modeled to properly depict genome evolution. However, their inclusion poses considerable computational challenges, and consequently no complete probabilistic model has been proposed so far. Using only inversions as possible rearrangements, Larget *et al.* [40] devised a Bayesian program able to estimate both organism phylogeny and ancestral genome arrangements on datasets containing 87 mitochondrial genomes, or eight genomes of closely related bacterial strains [41]. In a parsimony setting, Ma *et al.* [42,43] found an efficient and exact algorithm to reveal the most parsimonious scenario depicting genome evolution along an organism tree using events of speciation, deletion, insertion, duplication and a special type of rearrangement. However, this efficient algorithm can only be applied under the hypothesis that genomes contain an infinite number of sites; when applied to real finite data, approximations must be made, and only heuristics are available to find the most parsimonious scenario.

Devising a probabilistic model of genome evolution taking into account inversions, duplications, losses, and chromosome fission and fusion would likely offer tremendous insight into genome evolution, but seems very difficult; further incorporating substitutions, insertions and deletions would build a model able to align whole genomes in a statistical framework (Figure 1), and should result in an improved understanding of genome dynamics, as it is known that the genomic environment influences the substitution pattern [44]. If issues of algorithmic complexity and potential overparametrization (Box 3) can be solved, this can help test theories of speciation by chromosomal rearrangements [45], help detect and characterize whole genome duplications [37], pave the way for the reconstruction and study of ancestral genomes [46,47], and help test whether genome complexity originates in small effective population sizes [48]. Additionally, it could provide a useful platform for genome annotation; if averaging out sequence alignments improves robustness and accuracy for a range of estimates [3,13,16], one can expect that averaging out genome alignment will improve annotation accuracy, as annotation is already known to benefit from a comparative approach [26,49].

Ancestral phenotypes and ecosystems

Beyond information on molecular evolution and phylogeny, genomes conceal the footprints of ancient functions,

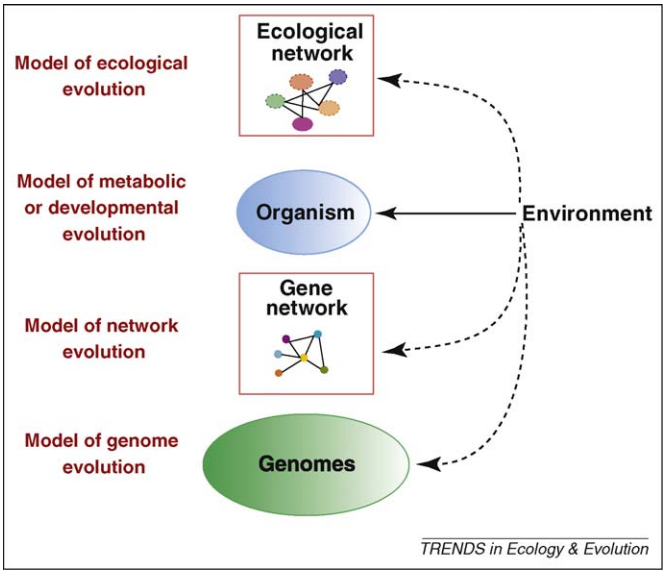

Figure 2. Models of genome evolution: from raw genome sequences to ecosystems. Organisms can be described at different levels of organization, and for each level different models of evolution can be devised. Using these models of evolution in combination can benefit the reconstruction of characteristics of ancient organisms. The environment in which organisms live operates a direct selective pressure on the organism (continuous arrow), which has repercussions on each organization level (dashed arrows). Genomes can adapt to the environment by shifting the composition of their genes, gene networks through the loss or acquisition of new genes and new interactions between genes, and ecological networks by the appearance or disappearance of new connections between organisms or altogether new organisms. Thus, there is a relation between the way organisms function and their environment, which can be input into models of evolution for more accuracy.

environmental constraints, and selection pressures. Such footprints permit reconstructing ancient ecosystems and the encoded phenotype of ancestral organisms based on the analysis of the genomes of extant individuals. It is possible to infer parts of an ancient organism phenotype by reconstructing and characterizing one of its genes (e.g. [50]) by statistically predicting gene functions with a model of function evolution [26], or by analyzing relevant genome characteristics such as structural RNA or protein content [51]. Such approaches could be combined and extrapolated to the entire gene repertoire, thus improving previous approaches that inferred ancestral phenotype based on gene content [35]. In addition to reconstructing ancestral states, which may be misleading because genomes are never at equilibrium with their environments, molecular evolution models can capture tendencies through the estimation of equilibrium values (e.g. [52]).

To reconstruct the inner workings of million-year-old cells, one would have to infer signaling pathways and metabolic cycles from genomic data (Figure 2). The reconstruction of regulation and metabolic networks has been shown to greatly benefit from an explicit evolutionary model [53–55]. Superimposed on models of genome evolution, such models of network evolution would offer opportunities to better infer ancestral metabolism and thus ancient ecologies [56]. Obviously, issues of overparametrization would again be a concern (Box 3).

Such genome-based inferences can be compared to results from other natural science disciplines through statistical models; for instance, a model could be built that infers ancient phenotypes as explained above, and simultaneously assesses the correlation with environmental temperature or atmospheric oxygen as inferred from geology. This would thus permit identifying lineages under the influence of the prevailing conditions at the surface of the Earth at a given time, and which lineages escaped these conditions. Similarly, models of continuous or discrete character evolution [57] or models of phylogeography [58] can be coupled to models of genome and network evolution to better analyze genotype–phenotype coevolution. Such integrative approaches should enable statistical tests of the significance of correlations between environmental conditions and biological phenomena, and will enlighten how the Earth and its biosphere shaped each other during billions of years of coexistence.

Conclusion

'The past is never dead. It's not even past' [59] – the chronicles of life resonate in extant genomes and we have only started to exploit the historical potential of macromolecules. New integrative statistical models of evolution, that fully exploit substitutions, gene duplication, loss and transfer, insertion–deletions and rearrangements can not only yield a better resolved history of life, but also a thorough representation of the whole process. Furthermore, through correlations between genomic properties and non-genomic variables, genomes document the history of interactions of organisms with each other and their environment. We can now envision the reconstruction of the history of genomes as well as metabolic and signaling networks, phenotypes and environments.

Acknowledgement

This work has been supported by the Agence Nationale de la Recherche grant ANR-08-EMER-011-03 "PhylAriane" and the GIP ANR JC05_49162. We would like to thank Colin Havenar-Daughton, Christine Pohl, Mathilde Paris and all members of the LBBE for help and comments.

References

1 Zuckerkandl, E. and Pauling, L. (1965) Molecules as documents of evolutionary history. *J. Theor. Biol.* 8, 357–366
2 Löytynoja, A. and Goldman, N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 320, 1632–1635
3 Wong, K.M. *et al.* (2008) Alignment uncertainty and genomic analysis. *Science* 319, 473–476
4 Thorne, J.L. *et al.* (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33, 114–124
5 Bradley, R.K. and Holmes, I. (2007) Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics* 23, 3258–3262
6 Holmes, I. and Bruno, W.J. (2001) Evolutionary hmms: a Bayesian approach to multiple alignment. *Bioinformatics* 17, 803–820
7 Lunter, G. *et al.* (2005) Bayesian coestimation of phylogeny and sequence alignment. *BMC Bioinformatics* 6, 83
8 Redelings, B.D. and Suchard, M.A. (2005) Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* 54, 401–418
9 Liu, K. *et al.* (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* 324, 1561
10 Rivas, E. and Eddy, S.R. (2008) Probabilistic phylogenetic inference with insertions and deletions. *PLoS Comp. Biol.* 4, e1000172
11 Redelings, B.D. and Suchard, M.A. (2007) Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC Evol. Biol.* 7, 40
12 Huelsenbeck, J.P. *et al.* (2002) Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51, 673–688
13 Miklós, I. *et al.* (2008) How reliably can we predict the reliability of protein structure predictions? *BMC Bioinformatics* 9, 137
14 Duret, L. *et al.* (1993) Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic Acids Res.* 21, 2315–2322
15 Stark, A. *et al.* (2007) Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature* 450, 219–232
16 Satija, R. *et al.* (2008) Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics* 24, 1236–1242
17 Maddison, W.P. (1997) Gene trees in species trees. *Syst. Biol.* 46, 523–536
18 Suchard, M.A. *et al.* (2003) Hierarchical phylogenetic models for analyzing multipartite sequence data. *Syst. Biol.* 52, 649–664
19 Degnan, J.H. and Rosenberg, N.A. (2006) Discordance of species trees with their most likely gene trees. *PLoS Genet.* 2, e68
20 Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332–340
21 Beerli, P. and Felsenstein, J. (1999) Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* 152, 763–773
22 Liu, L. and Pearl, D.K. (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.* 56, 504–514
23 Carstens, B.C. and Knowles, L.L. (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. *Syst. Biol.* 56, 400–411
24 Arvestad, L. *et al.* (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. In *RECOMB04*
25 Akerborg, O. *et al.* (2009) Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Natl. Acad. Sci. U. S. A.* 106, 5714–5719
26 Engelhardt, B.E. *et al.* (2005) Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.* 1, e45

27 Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science* 284, 2124–2129
28 Kurland, C.G. *et al.* (2003) Horizontal gene transfer: a critical view. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9658–9662
29 Lerat, E. *et al.* (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biol.* 1, E19
30 Suchard, M.A. (2005) Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics* 170, 419–431
31 Addario-Berry, L. *et al.* (2003) Towards identifying lateral gene transfer events. In *Pacific Symp. Biocomputing*, pp. 279–290
32 Hallett, M. *et al.* (2004) Simultaneous identification of gene duplication and horizontal transfer events. In *RECOMB*
33 Huang, J. and Gogarten, J.P. (2006) Ancient horizontal gene transfer can benefit phylogenetic reconstruction. *Trends Genet.* 22, 361–366
34 Snel, B. *et al.* (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25
35 Boussau, B. *et al.* (2004) Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 101, 9722–9727
36 Barker, D. and Pagel, M. (2005) Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput. Biol.* 1, e3
37 Sémon, M. and Wolfe, K.H. (2007) Consequences of genome duplication. *Curr. Opin. Genet. Dev.* 17, 505–512
38 Hobolth, A. *et al.* (2007) Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* 3, e7
39 Ané, C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.* 24, 412–426
40 Larget, B. *et al.* (2005) A Bayesian approach to the estimation of ancestral genome arrangements. *Mol. Phylogenet. Evol.* 36, 214–223
41 Darling, A.E. *et al.* (2008) Dynamics of genome rearrangement in bacterial populations. *PLoS Genet.* 4, e1000128
42 Ma, J. *et al.* (2008) Dupcar: reconstructing contiguous ancestral regions with duplications. *J. Comput. Biol.* 15, 1007–1027
43 Ma, J. *et al.* (2008) The infinite sites model of genome evolution. *Proc. Natl. Acad. Sci. U. S. A.* 105, 14254–14261
44 Eyre-Walker, A. and Hurst, L.D. (2001) The evolution of isochores. *Nat. Rev. Genet.* 2, 549–555
45 Rieseberg, L. (2001) Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* 16, 351–358
46 Sturtevant, A.H. and Dobzhansky, T. (1936) Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. *Proc. Natl. Acad. Sci. U. S. A.* 22, 448–450
47 Muffato, M. and Crollius, H.R. (2008) Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. *Bioessays* 30, 122–134
48 Lynch, M. (2007) *The Origins of Genome Architecture*, Sinauer Assoc Inc
49 Dewey, C. *et al.* (2004) Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Res.* 14, 661–664
50 Gaucher, E.A. *et al.* (2008) Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* 451, 704–707
51 Boussau, B. *et al.* (2008) Parallel adaptations to high temperatures in the Archaean eon. *Nature* 456, 942–945
52 Boussau, B. and Gouy, M. (2006) Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* 55, 756–768
53 Wiuf, C. *et al.* (2006) A likelihood approach to analysis of network data. *Proc. Natl. Acad. Sci. U. S. A.* 103, 7566–7570
54 Ratmann, O. *et al.* (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Comput. Biol.* 3, e230
55 Pinney, J.W. *et al.* (2007) Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *Proc. Natl. Acad. Sci. U. S. A.* 104, 20449–20453
56 Price, N.D. *et al.* (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* 2, 886–897
57 Pagel, M. *et al.* (2004) Bayesian estimation of ancestral character states on phylogenies. *Syst. Biol.* 53, 673–684
58 Kozak, K.H. *et al.* (2008) Integrating GIS-based environmental data into evolutionary biology. *Trends Ecol. Evol.* 23, 141–148
59 Faulkner, W. (1951) *Requiem for a Nun*, Penguin
60 Minin, V.N. *et al.* (2005) Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics* 21, 3034–3042
61 Long, M. *et al.* (2003) The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875
62 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113
63 Sémon, M. and Wolfe, K.H. (2007) Reciprocal gene loss between tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* 23, 108–112
64 Steel, M. (2005) Should phylogenetic models be trying to "fit an elephant"? *Trends Genet.* 21, 307–309
65 Suchard, M.A. and Rambaut, A. (2009) Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25, 1370–1376
66 Lartillot, N. and Philippe, H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.* 21, 1095–1109
67 Suchard, M.A. *et al.* (2001) Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18, 1001–1013
68 Huelsenbeck, J.P. *et al.* (2000) A compound Poisson process for relaxing the molecular clock. *Genetics* 154, 1879–1892
69 Novák, A. *et al.* (2008) Statalign: an extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* 24, 2403–2404
70 Meyer, I.M. and Miklós, I. (2007) Simulfold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.* 3, e149
71 Paradis, E. *et al.* (2004) Ape: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20, 289–290