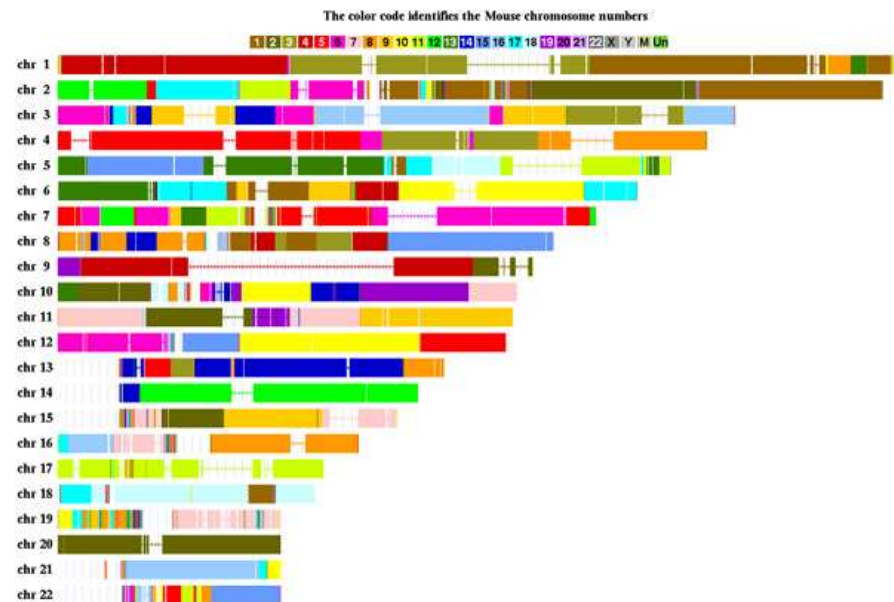
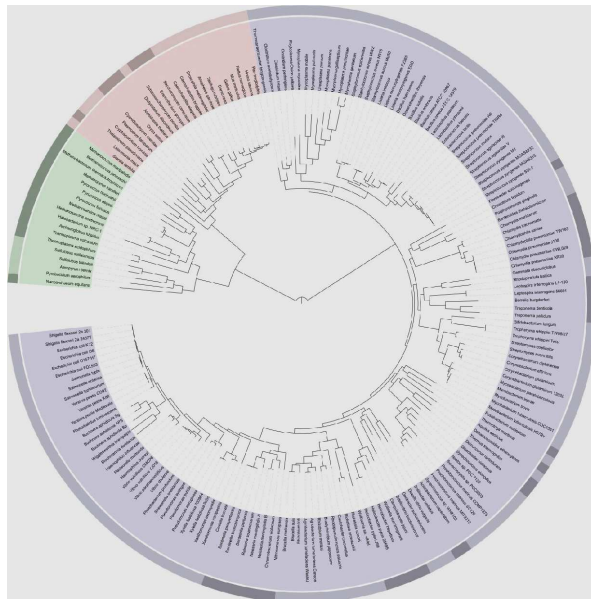


# Comparative Genomics and the Tree of Life

Bernard M.E. Moret

Laboratory for Computational Biology and Bioinformatics

EPFL



# Overview

- **The genome and its evolution**
- **Genomic events in phylogenetic reconstruction**
- **Models and algorithms for two genomes**
- **Comparative genomics: multiple genomes**
- **Medians, ancestral reconstruction, and alignment**
- **Conclusion**

# Comparative genomics and phylogeny

## Comparative genomics benefits from phylogeny:

*Direct comparisons fare poorly on pairs of highly divergent taxa, but a phylogenetic framework provides intermediate references. A phylogeny gives a basis for much richer annotations than possible in a pairwise comparison.*

## Phylogenetic inference benefits from comparative genomics:

*Rare genomic events such as rearrangements and duplication/losses give insight into the distant past of evolution. Enabling full-genome computations may help resolve disputed parts of the tree (e.g., primates/rodents/carnivores). Model-based inference enables the reconstruction of past events and of ancestral genomes.*

# Evolution of the genome

## Evolutionary events that affect the genome:

**nucleotide-level:** *“classical” sequence evolution  
(mutations and indels)*

**genomic rearrangements:** *inversions, transpositions, translocations,  
and chromosomal fusion and fission*

**duplication:** *gene retrotransposition, tandem duplication, segmental  
duplication, whole-genome duplication*

**loss:** *point mutation, segmental deletion, neofunctionalization*

**recombination:** *meiotic recombination, hybridization,  
lateral gene transfer*

# Evolutionary models

And how well understood are they?

**nucleotide-level**: *well established models with good statistics,  
but large variations within a population (SNPs, copy numbers)*

**genomic rearrangements**: *enormous work in the last 10 years,  
but still parameter-poor*

**duplication/loss**: *established work in lineage sorting,  
much attention to whole-genome duplication,  
just starting work on segmental duplications*

**recombination**: *established work in population genetics,  
much work on identifying lateral gene transfer,  
just starting detailed work on recombination*

# Rearrangements and phylogeny

- 1930s-1940s:** *Sturtevant and Dobzhansky use rearrangements found through chromosome banding to reconstruct a small phylogeny of fruit flies.*
- 1970s-1980s:** *Jeff Palmer and various colleagues document and study gene transfer from organelles to the nucleus and rearrangements in chloroplast and mitochondria, using the former to reconstruct phylogenies.*
- 1990s:** *Systematists and evolutionary biologists advocate using genome rearrangements as a source of phylogenetic information.*
- 2000:** *David Sankoff convenes DCAF, devoted to genomic rearrangements and featuring work in models, algorithms, and phylogenetic reconstruction.*
- 2001–2005:** *Huge progress on the computational aspects of rearrangements. GRAPPA and Badger reconstruct chloroplast phylogenies; DCM-GRAPPA reconstructs bacterial phylogenies; MGR compares vertebrate genomes.*
- Last few years:** *Rearrangements and duplication/loss models come of age. In use in biological groups; produce new biological results.*

# Some recent results

*The “genolevure” (yeasts) group (a French-led European consortium): fully automated ancestral yeast genome reconstruction equalling results from a year’s bench efforts in Ken Wolfe’s lab.*

*Darling, Miklós, and Ragan (PLoS Genetics 2008): studies of rearrangement patterns within *Y. pestis*, with new findings regarding frequency and scope of inversions around the origin of replication.*

*Chauve and Tannier (PLoS Comp. Bio. 2008): using gene clusters and conserved syntenies to reconstruct genomic segments in mammalian phylogenies.*

*Ma et al. (JCB 2008): DUPCAR, software to reconstruct ancestral regions with large-scale (segmental) duplications.*

*Swenson, Arndt, Tang, and Moret (APBC’08): rearrangement phylogenies with fully automated handling of highly unequal gene content for bacterial genomes ( $\gamma$ -proteobacteria).*

# Genomic rearrangements

Rearrangements alter the order and strandedness of genomic regions, from subsequences through genes to syntenic blocks.

*To study rearrangements, genomes are represented by ordered sequences of signed indices, each index representing a gene or syntenic block.*

*Rearrangements can be characterized by*

- *their outcome: breakpoints*
- *their mechanism: inversions, transpositions, translocations*
- *a mathematical model: permutations, the Nadeau-Taylor model, double-cut-and-join (DCJ)*

*In any framework, they present challenging algorithmic questions.*

# Two is easy, more is hard

An optimization problem that can be solved efficiently for two objects often becomes *intractable* for three or more objects.

*Examples include Satisfiability of Boolean formulae in conjunctive form (easy for clauses of 2 variables, hard for clauses of 3 variables), matching (easy for matching pairs, hard for matching triples), problems on graphs of fixed degree (trivial on graphs of degree 2, often hard on graphs of degree 3), etc.*

In comparative genomics, the two basic problems exhibit this same behavior.

- *Sequence alignment, and hence also genomic alignment, is easy for two sequences, hard for more.*
- *Finding the rearrangement median between genomes is easy for two genomes, hard for more.*

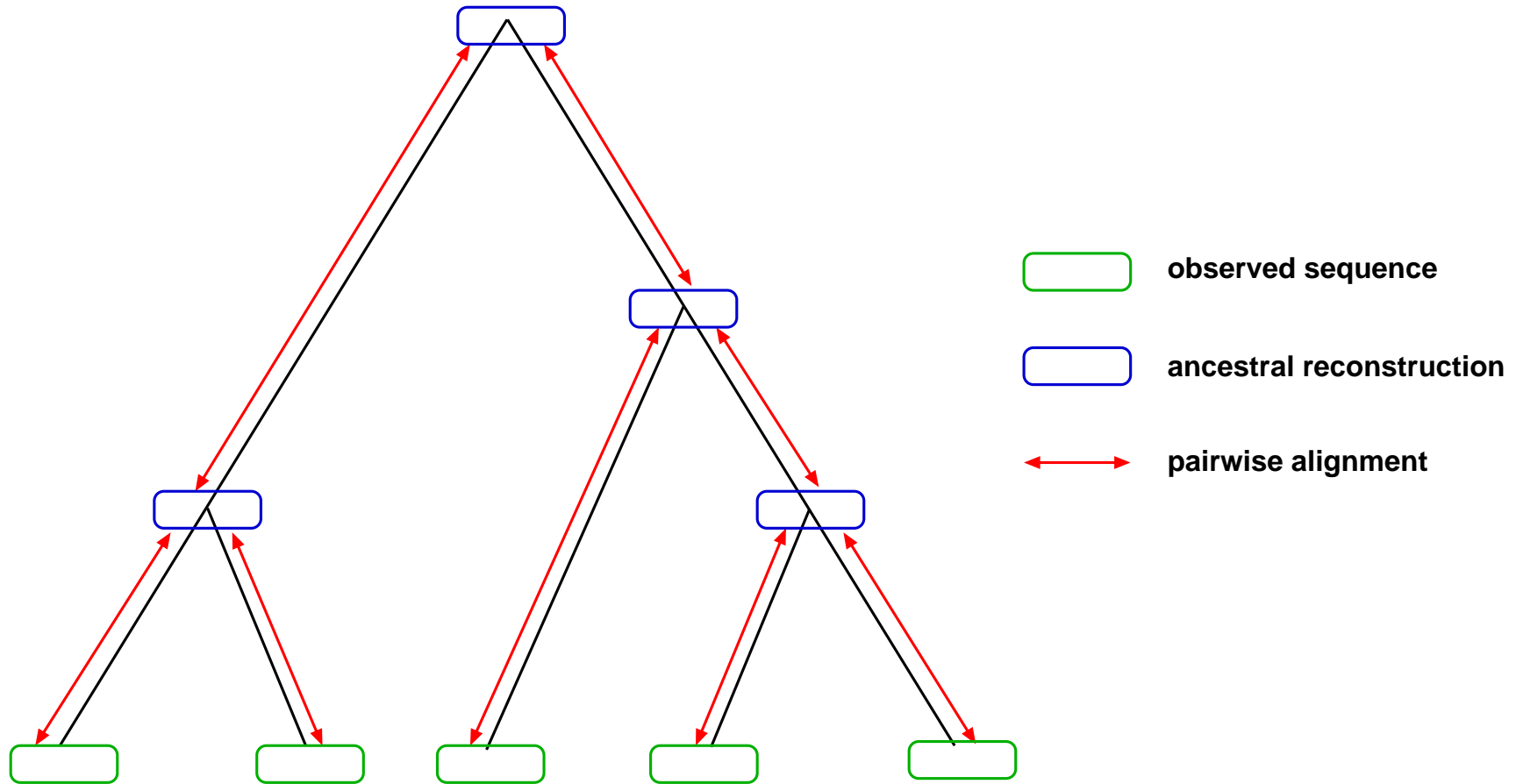
# Multiple sequence alignment

It remains the “single point of failure” of comparative genomics (not to mention of phylogenetic reconstruction).

- *all methods attempt to reduce multiple alignment to a series of pairwise alignments*
- *every popular tool is based on progressive alignment, using some assumed (or heuristically built) phylogenetic tree*
- *even the best alignment packages (MAFFT, ProbCons, Muscle) handle only point mutations and indels*
- *results tend to be poor for sequences with significant divergence*

In comparative genomics, the phylogeny is often known, yet even then progressive alignment may be poor.

# Sankoff's problem



*All pairwise alignments involve at least one ancestral sequence.*

# Sankoff's problem

*This formulation requires reconstruction of the full history of the given sequences from their last common ancestor—a very hard task.*

*No tool exists for this problem at present, except for small-scale work on gene-order data (e.g., Lancia's GESTALT, 1999).*

*Warnow's group has developed SATé (to appear in Science), the first usable tool for the general problem.*

# The median problem

The “other” big computational problem in comparative genomics is deceptively simple:

- *given  $k$  genomes (usually 3), find a new genome that minimizes the sum of the pairwise genomic distances from itself to the given  $k$  genomes*

Finding a median is

- *a key step in phylogenetic reconstruction and for Sankoff’s problem*
- *the most common approach to ancestral reconstruction*

Median optimization is intractable under most measures of genomic distance.

# Taming median computations

*We cannot avoid the problem entirely (except with progressive alignment), but we can find ways of estimating or approximating medians quickly:*

*minimum spanning tree: easily computed, then altered heuristically to produce a phylogenetic tree*

*tight bounding on edge scores: based on mixed integer-linear programming, one set for each tree*

*greedy methods: in a median of three, repeatedly move from one end towards the other two simultaneously; fork when no such move exists*

*path-bundling methods: commuting and non-interfering operations are grouped as single operators*

*decomposition methods: successfully used for DCJ medians*

# A note on ancestral genomes

Medians are used to reconstruct ancestral genomes, but **the two are quite distinct**.

Medians:

- *store intermediate algorithmic results*
- *answer a very simple optimization criterion*

Ancestral genomes:

- *biological constraints are presumably numerous, but we know very little about them, and so*
  - *good probabilistic models are lacking*
  - *reconstruction is severely underconstrained*
  - *“optimal” solutions (i.e., medians) abound*

# Positive results

- *Tracking gene clusters or operons across lineages.*
- *Ancestral genomes claimed for mammalian genomes at coarse resolution.*
- *Assembly approach instead of median:  
ancestral genome in a star tree built from many known syntenic blocks by selecting some and assembling them into a (possibly incomplete) genome.*
- *Signature approach by identifying rearrangements common to all shortest evolutionary paths.*
- *Automated phylogenetic reconstruction (incl. ancestral genomes) for yeast (with doubling) rivals year-long manual work in K. Wolfe's lab.*

# Conclusion

*The theory and practice of genomic evolution at whole-genome scale (rearrangements, duplications and losses, HGTs, recombinations) has come of age:*

- *In phylogenetic analysis, it has gone beyond reproducing phylogenies built by “conventional” (sequence-based) means.*
- *In the study of evolution, it now contributes significant insights into mechanisms and proffers serious candidates for evolutionary scenarios.*
- *In comparative genomics, it can handle complete eukaryotic genomes down to the sequence level if need be—providing alignments, orthology assignments, indication of the type of selection pressure, etc.*
- *Within AToL, it can be used directly for prokaryotes, thanks to next-gen sequencing, and for eukaryotic families related to model organisms (Saccharomyces, Drosophila, etc.).*