BST281: Genomic Data Manipulation, Spring 2019

Monday 12: Comparative Genomics

**Molecular evolution**

* Biological systems replicate imperfectly, resulting in genomic (genotype) variation
	+ Deleterious (harmful) variants tend to be purged (replicate less or not at all) = negative selection
	+ Neutral variants may become more common by random drift = neutral evolution
	+ Advantageous variants tend to replicate more = positive selection
* Non-neutral changes in genotype are observed as changes in phenotype
* Most changes to “important” DNA are deleterious and tend to be purged from the population
* Extent of conservation (lack of change) at a locus across species is an indicator of potential importance
* In CDSs, the ratio of non-synonymous to synonymous changes (dN/dS) estimates selective constraint
* Methods: *MEGA*, *PAML*, the *UCSC genome browser*

**Phylogenetic trees**

* Organize a collection of modern-day sequences according to their evolutionary history
* Distance-based methods operate on a table of pairwise distances between sequences
	+ UPGMA: ~agglomerative clustering; naïve method for tree-building; assumes a molecular clock (all branches changing at the same rate), which is generally not true
	+ Neighbor-Joining: Fast and reasonably accurate if distances are additive or nearly additive; iteratively pulls out pairs of nodes to minimize tree size; root by midpoint or by outgroup
* Multiple Sequence Alignment (MSA)-based methods
	+ Maximum Parsimony: Considering all possible tree topologies (computationally expensive!), pick the one that explains observed changes using the smallest number of point mutations
	+ Maximum Likelihood: Analog of Maximum Parsimony that attempts to identify the most likely tree rather than the cheapest one; these are the most common methods used today
* A single-gene tree does not always reflect/capture evolution of source species
	+ Requires an ortholog in each species; must be variable enough to resolve close relatives
	+ Lateral gene transfer (LGT) causes confusion (gene tree ≠ species tree)
	+ More data helps: Use more genes, either concatenating or “voting” over their individual trees
* Methods: *RAxML*, *PhyML*, *MrBayes*, *PHYLIP*, *FastTree*

**Multiple Sequence Alignment (MSA)**

* Align three or more sequences to identify homologous sites (columns of the MSA)
* Useful as an aid to some tree-building techniques, and for defining and studying protein families
* Computationally intractable to find the optimal MSA, so heuristic (progressive) methods used instead
* Methods: *MUSCLE* andthe *CLUSTAL* family of algorithms

**Other mechanisms of sequence/genome evolution**

* Genes, chromosomal regions, or entire genomes can be duplicated, deleted, shuffled, or inverted
* Genes can be acquired from other genomes (LGT)
* These processes result in variation in gene order (synteny) across organisms
* Synteny can change rapidly while gene content and gene sequences remain conserved
* Methods*: MUMMER* and *Mauve* are tools for studying syntenic variation in genomes

**Suggested reading**

* Lesk: Ch. 4, pp. 165-173 (phylogeny); Ch. 7, pp. 243-255 (comparative genomics)
* Pevsner: Ch. 6, pp. 205-222 (MSA), pp. 229-235 (sequence conservation); Ch. 7 pp. 245-296 (phylogenetics)
* Drummond, D. Allan, et al. "[Why highly expressed proteins evolve slowly](http://www.pnas.org/content/102/40/14338.short)." Proceedings of the National Academy of Sciences of the United States of America 102.40 (2005): 14338-1434.
* Lindblad-Toh, Kerstin, et al. "[A high-resolution map of human evolutionary constraint using 29 mammals](http://www.nature.com/nature/journal/v478/n7370/abs/nature10530.html)." Nature 478.7370 (2011): 476-482.