Gauging the utility of compound microsatellites for shallow scale phylogeny estimation.

Jay Hogan, Alan Lemmon
Florida State University, Department of Scientific Computing

Background

Microsatellites are common genetic motifs in which short nucleotide sequences are repeated many times (e.g. ATATATAT…). They are generally unconstrained by selection, mutate rapidly and show a high degree of polymorphism. This polymorphism makes them attractive candidates as tools for shallow scale phylogenetic analysis, but the rapidity with which they mutate results in a homoplasy problem that tends to obscure their ancestry even at shallow depths. That is to say that one often cannot infer whether two individuals with the same number of repeats at a locus owe this similarity to common ancestry or to convergence. This difficulty may be alleviated by focusing on compound microsatellites, sites at which multiple microsatellite alleles are located close enough together on a chromosome such that recombination is minimal. At these sites multiple characters are independently evolving at what may be approached as one locus, which may allow for sufficient accuracy for phylogeny reconstructions.

Methods

Microsatellites can be simulated as continuous characters allowed a rate of Brownian motion along the branches of a tree (see Figure 1). For this experiment, coalescent trees were generated using Mesquite for a range of tree sizes (from 4 to 64 taxa) and population sizes (from 1000 to 1000000). Alleles were simulated along each using a Brownian mutation rate of 1.0. For each treatment, maximum likelihood trees were estimated for the simulated data using the contml program in the PHYLIP package. These trees were constructed using a range of character counts (from 1 to 128). These likelihood trees were then compared back to their true trees with PHYLIP's treedist program, producing Robinson-Foulds distances, which reflect the number of branch rearrangements separating the two trees. This analysis was conducted for 80 replicates. Average Robinson-Foulds distances across replicates were used to compute the branches recovered as a proportion of the total branches.

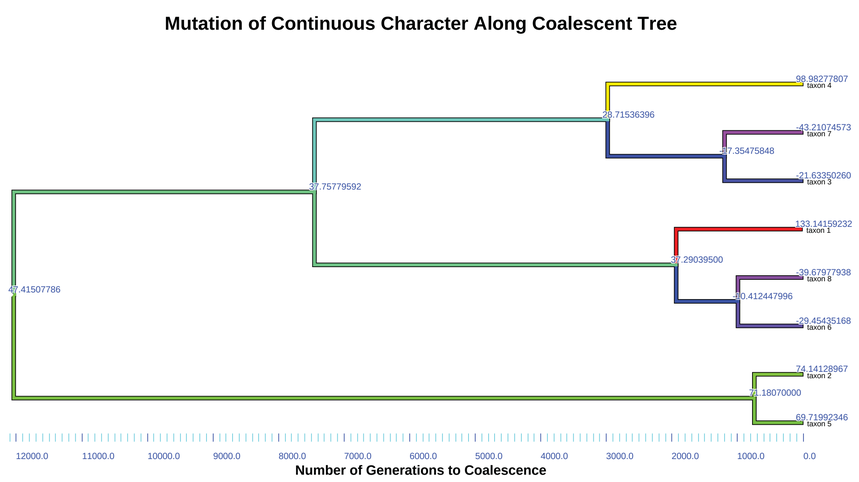

Figure 1. An example of simulated mutation of a continuous character across an eight taxon coalescent tree. Note the extent of homoplasy, particularly over long branch lengths.

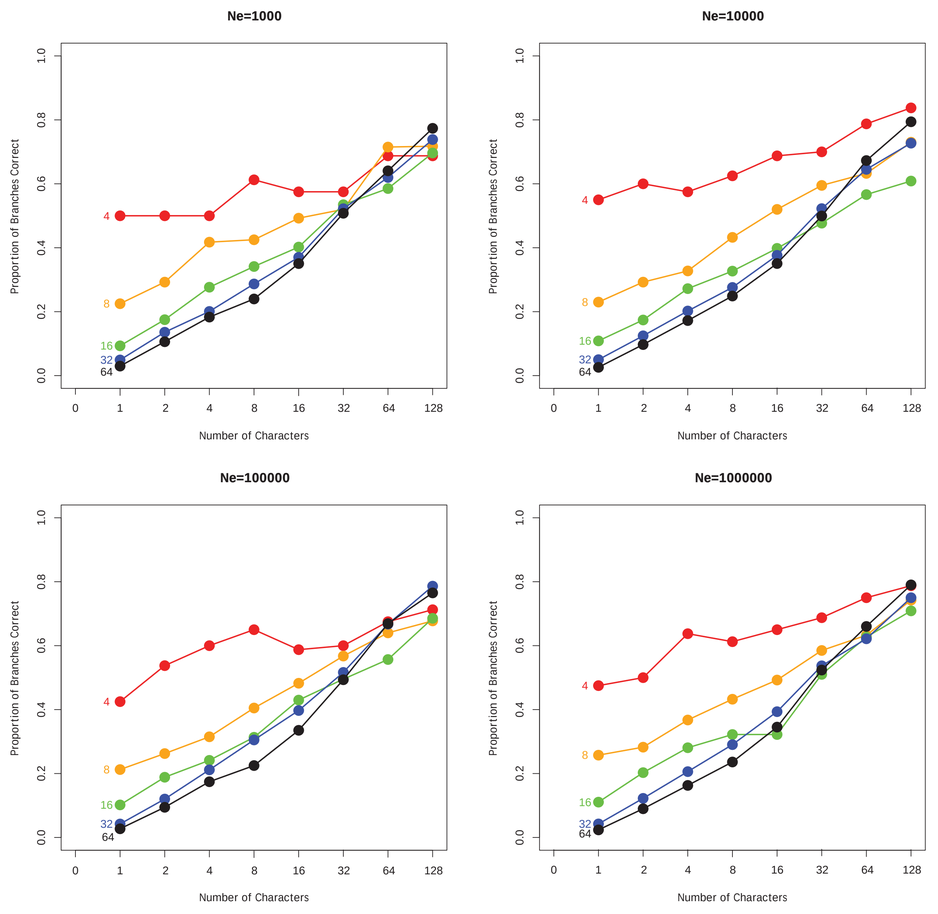

Figure 2. Proportion of branches of the true tree accurately recovered in maximum likelihood tree by number of characters analyzed. Shows results for five tree sizes (4-64) and four effective population

	Microsatellite counts for a given number of repeats							
	2	**3**	**4**	**5**	**6**	**7**	**8**	**>9**
Homo sapiens	51997	6096	1198	335	106	41	7	12
Maccaca mulatta	52796	6565	1389	433	155	49	10	10
Mus musculus	137237	26551	6561	2080	652	241	99	114
Rattus norvegicus	113077	16505	2632	607	170	78	19	32
Ornithorhynchus anatinus	1791	105	13	4	0	0	0	0
Gallus gallus	7782	610	115	17	6	2	0	0
Danio rerio	71280	15703	4163	1641	592	336	143	301
Drosophila melanogaster	685	29	0	0	0	0	0	0

Figure 3. Compound microsatellite sizes and counts for eight model species, as reported in Kofler et al, 2008.

Results

Results are displayed in Figure 2, which shows the trend in concordance between true and estimated trees with increasing character count. For a given tree size, as more characters are included in the analysis, a higher percentage of branches are successfully recovered. At low character counts, larger trees are more difficult to reconstruct than smaller trees. However, beyond a certain threshold of character count, the tree sizes converge towards a common accuracy. Results in the figure are for continuous microsatellite sizes, but real microsatellite counts are discrete. The use of continuous precision thresholds obscures the trend that would otherwise be observed across population sizes, with lower accuracies reported for smaller populations.

Discussion

The character counts were varied during the likelihood analysis in order to determine the number of characters necessary to provide reliable inferences about tree structure. When few characters are used, the ancestral history is more ambiguous and the likelihood tree is a poor estimate of the true tree. As more characters are included, this ambiguity is reduced and likelihood estimates are stronger.

As compound microsatellites in real genomes are limited in size and frequency, the threshold at which the character count becomes sufficient for tree estimation has obvious significance to the usefulness of these motifs as phylogenetic tools. Reported data suggest that the reliability of tree estimation improves with increased character counts, and that this improvement begins within the range of compound microsatellite size consistent with biological reality. Trees may be estimated at an acceptable accuracy threshold of >20% given eight microsatellite characters. Figure 3 reports data from Kofler et al (2008) profiling compound microsatellite size and count from a sample of model species, suggesting that compound microsatellites of eight characters or more may be found in real genomes for many species. This may suggest a novel approach to the resolution of shallow scale phylogenies, especially in light of approaching technological advances that are expected to make whole genomes practical to obtain for phylogenetic research.

References

Bull, L.N., Pabon-Pena, C.R. and N.B. Freimer. 1999. Compound microsatellite repeats: practical and theoretical features. *Genome Res*. 9:830–8.

Estoup, A., Jarne, P. and J-M Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology.* 11:1591–1604.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle.

Kofler, R., Schlötterer, C., Luschutzky, E. and T. Lelley. 2008. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. *BMC Genomics*. 9:612.

Maddison, W.P., and D.R. Maddison. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org

Wilson, I.J. and D.J. Balding. 1998. Genealogical inference from microsatellite data. *Genetics*.150:499–510.