NEWS AND COMMENTARY

Genomics

Anthrax and the art of war (against ascertainment bias)

M Worobey

Heredity (2005) 94, 459–460. doi:10.1038/sj.hdy.6800636

Published online 19 January 2005

Single-nucleotide polymorphisms (SNPs) are the most widespread type of sequence variation in genomes, and are quickly emerging as highly valuable genealogical and phylogenetic markers (Morin *et al.*, 2004). However, SNPs often come with a steep price attached. As they are usually identified using a limited subset of the sequences from a population, they represent a highly biased sample of genetic variation. A recent paper by Pearson *et al.* (2004) faces this problem of ‘discovery bias’ head on. In an embodiment of the maxim ‘Know thine enemy’, the authors thoroughly investigate, then neutralize, some of the problems associated with using SNPs. They go on to build atop this theoretical foundation the most thorough empirical study of *Bacillus anthracis* evolutionary history to date, one that could serve as a model for other whole-genome phylogenetic analyses.

So, what is discovery bias, and why is it a problem? Imagine you are presented with two aligned, completely sequenced anthrax genomes. You compare them, base for base, and discover, scattered throughout the aligned genomes, a thousand sites where they have different nucleotides. These sites are SNPs, and your two-genome comparison has yielded a panel that can now be used to survey new, unsequenced strains. Characterize just those 1000 SNPs in some other strains, and you have built yourself an attractive data set for phylogenetic analysis, one that exploits the complete-genome perspective without the trouble and expense of generating new complete genomes. All is well until your phylogenetics software package returns a strange-looking tree: at either end of a single trunk are your two original strains, while in between, stuck onto various parts of the trunk, but with no secondary branches of their own, are your newly typed strains (Figure 1, red and green trees). Instead of the hierarchically branching oak tree you were expecting, you get a piece of bamboo.

The curious phenomenon of ‘branch collapse’ was originally identified in an elegant study on SNPs of *Mycobacterium tuberculosis* (Alland *et al.*, 2003). Pearson *et al.* extend this work and explicitly point out, using simulations, that only those mutational events falling on the evolutionary path between the two strains used to generate the panel of SNPs will be discovered. The upshot is that other sequences will appear on the linear tree linking the reference strains, but only at the point they started to embark on their own evolutionary pathway. Their independent evolutionary

history is not captured by the SNPs defined by the reference sequences, so their branches collapse back to the most recent common ancestor shared with one of the reference strains. Any substitutions that occurred after this point are invisible using this biased set of SNPs.

Having familiarized themselves with the enemy (ascertainment bias, that is), Pearson *et al.* forged ahead with an empirical study of anthrax, using the following recipe. Firstly, use an existing phylogeny to choose four strains for full genome sequencing, then define SNPs based on pairwise comparisons between these newly sequenced genomes and the Ames strain, which was already fully sequenced. Next, reconstruct linear phylogenies using each SNP panel. Lastly, combine the linear phylogenies into a composite phylogeny that fleshes out several secondary branches, as in Figure 1. In addition to the five complete genomes, chosen to cover the main anthrax lineages, Pearson *et al.* added 21 diverse *B. anthracis* strains, plus three outgroup strains, typed for SNPs.

Figure 1 SNP discovery bias and what to do about it. The black tree at the top represents the phylogeny that would be obtained if all strains were completely sequenced and all SNPs sampled. In practice, a subset of reference strains is typically used to define a panel of SNPs. Using any single pair of complete genomes will lead to a highly biased selection of sites. Only substitutions lying on the evolutionary pathway linking that pair will be detectable, as shown by the red and green trees. All secondary branches collapse to zero length, resulting in unusual linear phylogenies. Linear phylogenies derived from SNPs generated from different pairs of reference sequences can, however, be combined. In this case, most of the tree can be reconstituted. Only taxon B is left with a zero branch length, and an unknown history. However, the point at which it started to diverge is accurately captured by the position of the node on the composite tree.

Their composite tree was able to reconstitute much of the fully branched phylogeny. While the secondary branches of the nonreference strains remained invisible, the overlapping portions of the four linear phylogenies showed almost perfect correspondence in node placements of the 29 strains. The 'visible' portions of the resulting tree were exquisitely accurate, and supported by a vast amount of extremely clean phylogenetic signal. Out of 990 SNPs, representing noncoding, synonymous, and nonsynonymous sites, there was only a single site with conflicting phylogenetic signal (homoplasy).

This extraordinary lack of homoplasy immediately revealed two things. First, the anthrax lineage must indeed be relatively young (or very slowly evolving) compared to other bacterial clades, since there is no evidence of conflicting character states arising via multiple substitutions at the same site. Older lineages, with their greater evolutionary distances among sequences, tend to exhibit this sort of homoplasy.

Second, the lack of homoplasy confirms that there has been little or no recombination in the evolutionary history of the genomes under consideration; that anthrax is, in effect, a clonal organism. The almost total lack of secondary branching observed here with a large data set from across the entire genome is the clearest possible test of clonality. It will be interesting to see how much secondary branching will be observed in such studies of organ-

isms in future, with greater divergence and/or considerable levels of recombination.

The high number of highly stable SNP loci also made it possible for Pearson *et al* to root the anthrax tree in a robust way for the first time, indicating that the rare 'C' lineage is basal to the more common 'A' and 'B' lineages. Moreover, the accurate branch lengths on the main branch of the tree indicated that different anthrax lineages evolve at very different rates. The 'C' lineage appears to evolve at about a third the pace of the 'A' and 'B' lineages, and, within the 'B' lineage, strains from Europe evolve at a fraction of the rate of those from Africa, which experience different epidemiological dynamics.

Such evolutionary rate variation, fully viewed for the first time in this study, is particularly fascinating in *B. anthracis*, which is essentially frozen in evolutionary time between transmissions, while it sits and waits as a soil spore for its next chance to proliferate in a new host. Is the underlying neutral mutation rate slower in some branches of the tree due to lack of transmission opportunity, or do different selective pressures account for the different rates? Since there is an abundance of both synonymous and nonsynonymous substitutional variation in the data, it should be possible to test whether both synonymous and nonsynonymous rates follow the same patterns and hence whether selection or mutation rate differences underlie lineage-specific rate differences.

The fact that we can now pose and tackle such questions for an organism whose limited genetic variation has previously been seen as an obstacle to such work speaks to the tremendous potential of SNPs for phylogenetic analysis, when discovery bias problems are controlled. Pearson *et al* go far beyond strain typing, crucial in its own right for such problems as the molecular epidemiology of bioterrorism (Read *et al*, 2002). Their work forms a bridge between genomic data and important biological questions regarding the natural history, ecology, and tempo and mode of evolution of an important pathogen. Until complete genomes can be routinely churned out in the sorts of numbers still reserved for subgenomic studies, their approach might be worth considering for anyone who wishes to extract a great deal of informative phylogenetic signal out of a population with still limited genomic data.

M Worobey is at the Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA.

e-mail: worobey@email.arizona.edu

Morin PA, The SNP Workshop Group (2004). *Trends Ecol Evol* **19**: 208–216.

Pearson T *et al* (2004). *Proc Natl Acad Sci USA* **101**: 13536–13541.

Alland D *et al* (2003). *J Bacteriol* **185**: 3392–3399.

Read TD *et al* (2002). *Science* **296**: 2028–2033.

Further Reading

Beaumont MA (2004). Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity* **92**: 365–379.