npg

NEWS AND COMMENTARY

Searching for signatures of selection ·····································

Who believes in whole-genome scans for selection?

J Hermisson

·····································

The days of the neutral theory are seemingly over. In time for the Darwin year celebrations, recent research has allowed a remarkable comeback for selection as the dominant force in shaping the diversity of genotypes and phenotypes. This change in perception results mainly from the emerging field of evolutionary genomics. On the basis of newly available genome-wide polymorphism and divergence data, and driven by Big Genomics endeavours like the human HapMap Project, selection is detected without a phenotype, directly from DNA sequence data. Two main results emerge: (1) selection affects non-coding regions throughout the genome as well as coding regions, leading almost to a shortage of sequence material that can be considered reliably neutral in some species (Wright and Andolfatto, 2008). (2) There is evidence for frequent positive selection in the recent history of several species, including fruitflies, mice and particularly humans. Thousands of candidate regions for recent positive selection have been identified in >20 genome-wide scans in humans (Akey, 2009). Using a haplotype test, Hawks *et al*. (2007) found traces of adaptations in 7% of all human genes in as few as 40 000 years (3000 generations). Even higher estimates have been reported by Foll and Gaggiotti (2008). They used an F_{ST} based test and data from 53 human populations to find evidence for positive or balancing selection in 131 out of 560 (>23%) randomly distributed STR marker loci. This is a staggering number, but how reliable are these estimates? It has long been known that demographic effects can confound the results, but how severe are these problems in real world applications? A new study by Excoffier *et al*. (2009) suggests that they can be very severe indeed. In particular, the findings demonstrate the importance of an accurate characterization of population structure for methods based on F_{ST}.

Genomic tests for selection can be distinguished with respect to the summary statistics they use. Several of these statistics are used to detect hitchhiking events, also known as selective sweeps (Schlötterer, 2003; Pavlidis *et al*., 2008). These methods build on the characteristic footprint of recent positive selection on linked neutral DNA. The main effect is a local reduction in polymorphism, but the signal can also be picked up in the frequency spectrum and the local linkage disequilibrium or haplotype pattern. The strengths and weaknesses of the hitchhiking approach are quite well understood (Teshima *et al*., 2006; Thornton *et al*., 2007). The problem that is considered most severe is that certain events in the demographic history of a population can mimic the polymorphism patterns produced by selection. Population bottlenecks of a critical strength are the most dreaded alternative scenario. The reason is easy to understand: bottlenecks readily lead to large variances in the genealogical (coalescent) history of samples from different loci along a chromosome. These histories can either be short if the entire sample coalesces to a common ancestor during the bottleneck, or very much longer if several lines of descent extend through the bottleneck into a large ancestral population. As a result, almost all summary statistics show large variances, turning a population bottleneck into a neutral null-model that is hard to reject. Similarly, if a simpler demography is (wrongly) assumed, tests will produce an excess of false positives.

An alternative method to detect selection from genomic data goes back to Lewontin and Krakauer (1973). It is based on genetic diversity between subpopulations (demes) as measured by F_{ST} and follows a simple intuition: regions under diversifying selection should exhibit larger divergence among demes than neutral loci (high F_{ST}). Similarly, regions under uniform balancing selection in all demes should be less differentiated (low F_{ST}). More recently, these ideas have been developed into sophisticated statistical frameworks to detect selection from genome scans (for example, Beaumont and Nichols, 1996; Beaumont and Balding, 2004; Foll and Gaggiotti, 2008). Compared with the hitchhiking approach, the F_{ST} method focuses on a different selection scenario: diversifying local selection instead of population-wide positive selection. Consequently, one expects to detect partly complementary sets of candidate loci. The method was criticised early on by Robertson (1975) concerning robustness with respect to demography; however, recent theoretical considerations and a limited number of simulations have led to speculations that the method might be less vulnerable (Beaumont, 2005).

With the new work by Excoffier *et al*. (2009), this issue can be considered as settled. The authors convincingly establish a neutral model with hierarchical population structure as the 'bottleneck scenario' of the F_{ST} based approach. The reason is analogous to the case of sweeps and bottlenecks: due to hierarchical structure (and similarly due to range expansion or sequential population splits and mergers) different demes draw from different migrant pools, leading to higher levels of variance in F_{ST} than expected under an island model. To avoid excessive false positives, knowledge about the population structure needs to be built into the null distribution of F_{ST} that is used. For a hierarchical model, Excoffier *et al*. (2009) show how this can be done. The results are drastic—and sobering. In their re-analysis of human STR data, introduction of hierarchical structure based on five previously established geographic regions reduces the frequency of selection candidates from 23% (Foll and Gaggiotti, 2008) to no more than expected by chance (that is, comparable with the 1% significance level applied).

What do these results imply? Certainly that many numbers in published studies are up for revision. But not necessarily that selection is rare. The problem is that our knowledge about false negatives is even more rudimentary than about false positives. For panmictic populations, the power of many tests to detect selection is known to be rather low. For a structured population, this information is basically missing. Realistic models of selection should further account for local adaptation, adaptation from standing genetic variation or interference among selected loci. It will be important to characterize the expected genomic footprints under realistic scenarios in much more detail and to construct adequate (combinations of) summary statistics to detect the

resulting footprints against the background of demographic noise. The good news is that method development for selection mapping has not yet been pushed to its limits. In that sense, the contribution by Excoffier *et al.* (2009) is a valuable step on a longer road.

Professor J Hermisson is at the Mathematics and BioSciences Group, University of Vienna, Vienna, Austria

e-mail: joachim.hermisson@univie.ac.at

Akey JM (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res* **19**: 711–722.

Beaumont MA (2005). Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol Evol* **20**: 435–440.

Beaumont MA, Balding DJ (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **17**: 3425–3427.

Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc London* **B263**: 1619–1626.

Excoffier L, Hofer T, Foll M (2009). Detecting loci under selection in a hierarchically structured population. *Heredity* **103**: 285–298.

Foll M, Gaggiotti OE (2008). A genome scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**: 977–993.

Hawks J, Wang ET, Cochran GM, Harpending HC, Moyzis RK (2007). Recent acceleration of human adaptive evolution. *Proc Natl Acad Sci USA* **104**: 20753–20758.

Lewontin RC, Krakauer J (1973). Distribution of gene frequency as a test of the theory of selective neutrality of polymorphisms. *Genetics* **74**: 175–195.

Pavlidis P, Hutter S, Stephan W (2008). A population genomic approach to map recent positive selection in model species. *Mol Ecol* **17**: 3585–3598.

Robertson A (1975). Gene frequency distributions as a test of selective neutrality. *Genetics* **81**: 775–785.

Schlötterer C (2003). Hitchhiking mapping—functional genomics from the population genetics perspective. *Trends Genet* **19**: 32–38.

Teshima KM, Coop G, Przeworski M (2006). How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**: 702–712.

Thornton KR, Jensen JD, Becquet C, Andolfatto P (2007). Progress and prospects in mapping recent selection in the genome. *Heredity* **98**: 340–348.

Wright SI, Andolfatto P (2008). The impact of natural selection on the genome: emerging patterns in Drosophila and Arabidopsis. *Annu Rev Ecol Evol Syst* **39**: 193–213.

Editor's suggested reading

Thornton KR, Jensen JD, Becquet C, Andolfatto P (2007). Progress and prospects in mapping recent selection in the genome. *Heredity* **98**: 340–348.

Stinchcombe JR, Hoekstra HE (2007). Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity* **100**: 158–170.

Barker JSF, Frydenberg J, González J, Davies HI, Ruiz A, Sørensen JG, Loeschcke V (2009). Bottlenecks, population differentiation and apparent selection at microsatellite loci in Australian Drosophila buzzatii. *Heredity* **102**: 389–401.

Fauvelot C, Lemaire C, Planes S, Bonhomme F (2007). Inferring gene flow in coral reef fishes from different molecular markers: which loci to trust? *Heredity* **99**: 331–339.

Campos R, Storz JF, Ferrand N (2008). Evidence for contrasting modes of selection at interacting globin genes in the European rabbit (Oryctolagus cuniculus). *Heredity* **100**: 602–609.