Genetica
DOI 10.1007/s10709-008-9321-3

Population genomics and speciation

Roger K. Butlin

Received: 25 March 2008 / Accepted: 20 August 2008

Abstract The process of speciation begins with genomically-localised barriers to gene exchange associated with loci for local adaptation, intrinsic incompatibility or assortative mating. The barrier then spreads until reproductive isolation influences the whole genome. The population genomics approach can be used to identify regions of reduced gene flow by detecting loci with greater differentiation than expected from the average across many loci. Recently, this approach has been used in several systems. I review these studies, concentrating on the robustness of the approach and the methods available to go beyond the simple identification of differentiated markers. Population genomics has already contributed significantly to understanding the balance between gene flow and selection during the evolution of reproductive isolation and has great future potential both in genome species and in non-model organisms.

Keywords F_{ST} · Genome scan · Reproductive isolation · Local adaptation

Introduction

In most cases, the origin of species is a slow process. Polyploid or hybrid speciation can occur in one or a few generations but other modes of speciation require the progressive build-up of reproductive isolation. The available data are limited but the duration of this process appears to be extremely variable, from <4,000 years in the explosive radiation of cichlid species flocks to >1 million years in allopatric *Drosophila* species pairs (reviewed by Coyne and Orr 2004) and 0.01 to >5 million years in birds alone (Price 2007). In the early stages of reproductive isolation, gene flow is likely to be reduced at just a few loci, scattered around the genome, which contribute directly to local adaptation, mate choice, sexual conflict or genetic incompatibility between diverging populations. The effect of these loci spreads to other parts of the genome by a variety of mechanisms such as restriction of recombination by inversions (e.g. Ortiz-Barrientos et al. 2002), a general reduction in gene flow due to reduced fitness of immigrants and their offspring (Nosil et al. 2005) or the evolution of assortative mating (Servedio and Noor 2003). This spreading stage remains a particularly poorly understood part of the speciation process (Wu 2001).

Since the speciation process takes many generations, it has to be studied by comparison of many snap-shots of divergent and partially-isolated population pairs. It is generally not possible to foresee whether these populations will eventually evolve into completely isolated species but, nevertheless, they can be used to determine the genetic architecture of reproductive isolation and to dissect the contributions of different traits to the overall reduction in gene flow. A powerful approach to this problem is to map loci responsible for key traits, such as hybrid inviability (Presgraves 2003) or sterility (Ting et al. 1998) in *Drosophila*, pheromone composition and response in *Ostrinia* (Roelofs et al. 1987) or components of host adaptation in *Rhagoletis* (Feder et al. 2003). However, this approach is only available in species that can be crossed and reared in controlled conditions and it focuses on previously-known isolating traits. 'Population genomics' offers a powerful alternative approach that is not subject to these restrictions.

R. K. Butlin (✉)
Department of Animal and Plant Sciences, The University of Sheffield, Western Bank, Sheffield S10 2TN, UK
e-mail: r.k.butlin@sheffield.ac.uk

Luikart et al. (2003) applied the term 'population genomics' to a method originally suggested by Lewontin and Krakauer (1973). The principle is straightforward and can be illustrated by considering the pattern of divergence between two populations that are connected by gene flow but under divergent selection pressures. The expectation for a measure of divergence, such as F_{ST}, is the same for all neutral loci and is determined by the balance between mutation, drift and gene flow. The stochastic effects of drift, and of the sampling required to estimate F_{ST}, together produce a distribution of values around this expectation. Loci under divergent selection are expected to have higher F_{ST}, with the actual value dependent on the strength of selection but higher than the neutral expectation. Neutral loci closely linked to loci under selection will also have higher levels of divergence. Therefore, if sufficient markers are available, it may be possible to detect genomic regions influenced by selection because loci in these regions will appear as outliers, relative to the neutral distribution of divergence for the majority of loci. As Luikart et al. (2003) emphasised, the separation of markers into two classes, neutral and influenced by selection, not only identifies candidate regions for reproductive isolation or local adaptation but also improves inferences about population size and gene flow that are based on an assumption of neutrality. See also the reviews by Storz (2005), which focuses on the genetics of local adaptation, and Stinchcombe and Hoekstra (2007), which concentrates on the interaction between population genomic and quantitative genetic approaches.

Molecular methods now make it possible to develop large numbers of markers for any organism. Suitable data were available first for human populations (Beaumont and Nichols 1996) but the Amplified Fragment Length Polymorphism (AFLP) approach (Vos et al. 1995) has made population genomics possible for essentially any species. AFLP have to be used with care: dominance creates significant problems, for example with bias in the estimation of allele frequencies (Zhivotovsky 1999), homoplasy is common when many bands are scored from a single primer combination (Vekemans et al. 2002) and can cause underestimation of differentiation among subpopulations (Caballero et al. 2008), and error rates have to be carefully checked (e.g. Pompanon et al. 2005; Bonin et al. 2006). Nevertheless, AFLP have been used successfully in a rapidly growing number of studies (e.g. Wilding et al. 2001; Emelianov et al. 2004; Bonin et al. 2006; Savolainen et al. 2006; Egan et al. 2008; Minder and Widmer 2008; Nosil et al. 2008). In species with genome sequences available, it is possible to analyse a sample of loci (SNPs or microsatellites) spread throughout the genome in known locations, rather than a random sample of loci. The study of divergence among human populations by Akey et al. (2002) was the first of this type but subsequent analyses have included mosquitoes (Turner et al. 2005) and mice (Harr 2006a). These systematic surveys are known as 'genome scans'. The term has also been applied to studies based on random samples of loci (e.g. Murray and Hare 2006) but it may be preferable to reserve it for high-density, genome-wide analyses using mapped markers.

The principle of population genomics, that neutral loci follow a consistent distribution and loci influenced by selection can be detected as outliers, applies to measures of variability within populations as well as differentiation between populations. When selection fixes a new advantageous allele, genetic diversity in the surrounding sequence is reduced (a 'selective sweep'). The selective sweep also has an effect on the distribution of allele frequencies, detectable by measures such as Tajima's D that are sensitive to the abundance of rare substitutions relative to overall sequence diversity, and on linkage disequilibrium (see Nielsen 2005 for a review). An example of this type of population genomic study is the survey of variability on a region of the X chromosome of *Drosophila melanogaster* using microsatellite polymorphism which was followed up by analysis of sequence variation in a smaller candidate region (Pool et al. 2006). This study was able to identify fixed, derived substitutions in transcription factor binding sites upstream of the gene *roughest* that appear to have been the target of a recent selective sweep in African populations. However, the source of selection is not currently known.

Robustness of the F_{ST}-based population genomics approach

Population genomics approaches suffer from two classical problems in evolutionary and quantitative genetics: separation of the effects of selection from those of population history and dealing with false positives that arise from making multiple comparisons. A serious concern is the sensitivity of the outlier detection method to features of population history that might increase the variance of the F_{ST} distribution for neutral loci. Outlier detection relies on assumptions about the underlying population structure and history. It is most commonly achieved using the Fdist package (now Fdist2 for co-dominant markers and Dfdist for dominant markers such as AFLP, available from www.rubic.rdg.ac.uk/~mab/) which implements a coalescent simulation of a uniform, finite-island model, as described by Beaumont and Nichols (1996). An alternative, hierarchical-Bayesian approach has been developed by Beaumont and Balding (2004) and may perform marginally better than the original method, at least in some circumstances such as when mean F_{ST} is high or sample size is low. The underlying assumption is that the population sampled approximates the infinite-island model with uniform local population size and migration rate. This method is implemented in the BayesFst program (www.rdg.ac.uk/statistics/genetics/software.html). There is some confusion between methods in the literature: for example, Nosil et al. (2008, p. 321) describe Dfdist as an implementation of the Beaumont and Balding approach. In either case, it is possible that factors such as uneven population sizes or migration rates, isolation-by-distance or historical effects such as bottlenecks or population expansion might increase the variance of the F_{ST} distribution relative to these simple models. However, Beaumont and Nichols (1996) and Beaumont and Balding (2004) simulated data sets under a variety of demographic and population structure scenarios and found the effects on the F_{ST} distribution, and on the probability of correctly detecting loci influenced by selection, to be small.

Other methods for outlier detection are based on contrasting models. The DetSel package (Vitalis et al. (2003), www.genetix.univ-montp2.fr/detsel.html) models divergence without gene flow between two populations of unequal size derived from a single population that may not have been at demographic equilibrium. Wilding et al. (2001) considered reciprocal gene exchange between two populations of equal size and the same approach has been used by Bernatchez and co-workers (Campbell and Bernatchez 2004; Rogers and Bernatchez 2007). Eveno et al. (2008) have developed a method specific to bi-allelic loci such as SNPs. Despite these varying assumptions, similar sets of outliers are identified by the different methods (e.g. Bonin et al. 2006; Eveno et al. 2008). This suggests that the approach is robust to uncertainty about the true population structure and history.

Murray and Hare (2006) examined an historical scenario that may be particularly relevant in speciation studies and especially likely to distort the F_{ST} distribution. Their study system was a zone of secondary contact in the oyster, *Crassostrea virginica*. Phylogeographic evidence suggests that the two populations diverged in allopatry and re-established contact after the last glaciation. The current level of divergence at neutral loci might reflect the decay, due to gene flow, of the differentiation accumulated in allopatry. Murray and Hare reasoned that the variance in F_{ST} might be greater during this decay than it is at equilibrium and this might prejudice outlier detection. Their simulations showed that both the mean and variance of F_{ST} decay rapidly (over less than 3,000 generations) following secondary contact, as expected. However, at any point during this decay, the distribution of F_{ST} values across neutral loci, given the observed mean F_{ST}, is not distinguishable from the equilibrium case. Thus, outlier detection should be as reliable with this extreme departure from the model assumptions (the prolonged period of allopatry) as it is when the study population conforms more closely to the modelled structure.

The robustness of F_{ST}-based population genomics contrasts with the impact of population history on measures of diversity. For example, cosmopolitan populations of *Drosophila melanogaster* are derived from an ancestral African population. Thornton and Andolfatto (2006) showed that the bottleneck and subsequent population growth experienced during this range expansion can account for most, if not all, of the observed patterns of diversity, Tajima's D and linkage disequilibrium among X-linked loci. Given the problems of precisely estimating the population history, it is not possible to exclude demographic effects completely and so it is difficult to obtain firm evidence for the influence of selection. Selection may be more easily detected in the relatively stable ancestral populations (Pool et al. 2006). Other authors have reached similar conclusions about the impact of population history on the population genomic approach (Kelley et al. 2006; Teshima et al. 2006). However, these conclusions do not apply to the F_{ST}-based approach, at least when it is applied to populations connected by current gene flow. This is because of the different timescales on which gene flow and mutation operate (Beaumont 2005): the distribution of alleles among demes (which is measured by F_{ST}) is dependent on recent events in the 'scattering phase' (Wakeley 1999) while the allele frequencies at the start of this phase depend on the genealogy in the 'collecting phase' which typically behaves as a single population. This 'separation of timescales' underlies the model used in the BayesFst program (Beaumont and Balding 2004) and is implicit in other F_{ST}-based methods. It may be violated in some circumstances, for example where mutation rates are high, but generally it appears to be robust (Beaumont 2005). It does not apply to approaches that rely in diversity or linkage disequilibrium because these patterns depend on mutation and recombination.

A related set of issues concerns the power of the population genomics approach and the false discovery rate (i.e. the probability of detecting a locus that is genuinely influenced by selection and the probability of falsely identifying a neutral locus as an outlier, respectively). Since many loci have to be considered in a population genomics study, it is essential to control the experiment-wide significance level when identifying outliers. One approach is to make multiple comparisons. Wilding et al. (2001) studied the rocky-shore gastropod, *Littorina saxatilis*, in three localities, in each case obtaining one high-shore morph (H) and one mid-shore morph (M) sample. They identified as outliers loci with F_{ST} above the 99th percentile of the simulated neutral distribution in all three between-morph comparisons. Given 300 AFLP loci in total, such a pattern was very unlikely by chance. However, the reliability of this approach is unclear since populations at the three localities are connected by gene flow and so are not truly independent replicates.

BayesFst uses a prior distribution to control the experiment-wide probability and so should be free from the multiple-test problem (Beaumont and Balding 2004). This is not true for Fdist2 and Dfdist which means that confusion of these methods risks inflating the Type 1 error rate (cf. Nosil et al. 2008, p. 323). Therefore, other authors have followed Wilding et al. (2001) in placing greater reliance on loci that are identified in more than one pairwise test (e.g. Bonin et al. 2006; Nosil et al. 2008). Clearly, the false discovery rate (FDR) depends on the critical probability value used to identify outliers. It also depends on the method used to match the simulated F_{ST} distribution to the empirical distribution. One approach is to use a trimmed mean of the empirical F_{ST} distribution as the target mean for the simulation. Nosil et al. (2008) experimented with various levels of trimming (10%, 20% or 30%) but this did not change their conclusions about 'different-host' outliers (see below). However, simulations by Caballero et al. (2008) show that the use of the trimmed mean can inflate the FDR in some circumstances because it tends to shift the simulated distribution towards lower values of F_{ST}. Using the median, conditional on heterozygosity, is a conservative alternative to the trimmed mean which is recommended in the latest version of the Dfdist Readme file (www.rubic.rdg.ac.uk/~mab/stuff/). An alternative, iterative approach was recommended by Beaumont and Nichols (1996) and used by Wilding et al. (2001), for example. Initially, the untrimmed empirical mean is used, outliers are then removed and the simulation repeated with the new mean. This process is repeated until no further outliers are generated.

Recent simulation studies (De Kovel 2006; Singleton 2008) have considered both the power and the FDR of the population genomic approach. Not surprisingly, the power to detect the influence of selection on a marker locus is greatest where selection on a target locus is strong, the marker is closely linked to the target of selection, the average neutral F_{ST} is low and both the population sizes and sample sizes are large. The FDR is lowest under similar conditions. Significantly, power as low as 50% and FDR as high as 50% is not uncommon under realistic conditions. This has serious implications for extrapolation from the proportion of outliers detected in a sample of loci to the genome as a whole (see below). The range of methods and cut-offs used also makes it very difficult to conduct meaningful comparisons, among studies, of the proportions of outliers.

The practical implications of this analysis are that the F_{ST}-based population genomics approach is robust to departures from the modelled population structure and history and therefore the choice of outlier detection method is not critical. However, it is critical to interpret outliers with caution. The safest approach may be to view outliers simply as markers of candidates for regions of the genome influenced by selection and to seek ways to take further steps to identify these regions and test for selection in other ways. Alternatively, or in addition, one can seek patterns in the outliers that might be robust to problems of power or false discovery. Finally, it may be preferable to move away from the categorisation of loci into outlier and non-outlier groups.

Population genomics and speciation

Several studies have applied the population genomics approach to address the genetic basis of local adaptation and the progression towards increasing reproductive isolation and speciation in the face of gene flow. Here, I will briefly review some of these studies to illustrate how they contribute to the understanding of speciation and, particularly, how one can go beyond the simple identification of outliers.

Two morphs of the intertidal gastropod, *Littorina saxatilis*, occupy different parts of rocky shores in northeast England. Like similar morph pairs in other parts of Europe, they are adapted to different combinations of environmental conditions, primarily crab predation and the risk of dislodgement by wave action (see Wood et al. 2008 and references therein). Wilding et al. (2001) found 15 outliers in a survey of 290 polymorphic AFLP loci and these were consistent across three partly independent comparisons, as mentioned above. The genomic distribution of these outliers is unknown. An important question arises, therefore: do the outliers mark 15 independent targets of selection or are they clustered, perhaps in a region of low recombination such as might be generated by a chromosomal rearrangement? A detailed survey of two shores (Grahame et al. 2006) showed steep, coincident clines for all 15 loci but only a very slight elevation in linkage disequilibrium among these loci in cline-centre populations. The power of this disequilibrium analysis was low because sample sizes around the cline centre were small but, nevertheless, it clearly indicates that the loci are not tightly linked. The fine-scale sampling also showed that differentiation at neutral loci is greater in between-morph than within-morph comparisons, after accounting for isolation by distance. Thus there appears to be a general barrier to gene exchange between morphs which may be due to local adaptation, assortative mating or genetic incompatibility. Similar patterns have been detected in subsequent studies and considered in more detail (see below).

AFLP outliers are more likely to be linked to loci under selection than they are to be the direct targets of selection. Therefore, it is interesting to determine the pattern of sequence differentiation around outlier loci. In particular, one might be able to detect coding regions within islands of differentiation that show independent evidence of selection, for example in the ratio of synonymous to non-synonymous polymorphism, and so represent candidate adaptive loci. This sort of follow-up study is difficult in non-model organisms. However, Wood et al. (2008) have made progress with two of the outliers from the Wilding et al. (2001) study, with surprising results. They probed a large-insert genomic library with sequences of outlier and control loci, sequenced four of the bacterial artificial chromosome (BAC) inserts so identified and then sequenced the regions containing the AFLP sites, and flanking regions, in new samples from natural populations. The two outlier AFLP loci correspond to indel polymorphisms. In both cases, the insertions have characteristics of transposable elements and appear to be recent in origin because sequences that contain insertions are significantly less diverse than those that do not. Interestingly, Minder and Widmer (2008) have also found retrotransposon sequences associated with outliers in a comparison between hybridizing species of *Silene*. In *Littorina*, insertion frequencies differ strongly between morphs. However, sequences in flanking regions approximately 5 kb either side of these insertions, including nearby coding regions, are not differentiated between morphs. Thus the insertions themselves are candidate targets of selection, perhaps through effects on expression of downstream loci.

Charlesworth et al. (1997) predict that excess differentiation will be observed over a recombination distance of the same order as the selection coefficient favouring locally adapted alleles. Wood et al. (2008) estimated the selection on the insertion polymorphisms to be in the range 0.03–0.12, on the basis of the observed difference in frequency and the gene flow inferred from neutral loci. Therefore, the lack of differentiation in flanking regions is surprising. Selection for local adaptation may have been over-estimated, there may be stabilizing selection on flanking regions or gene flow from neighbouring populations may help to homogenize variation around the selected locus.

Taken at face value, the *Littorina* story currently suggests that there are many independent targets of selection in the genome. If 5% of randomly chosen regions are influenced by selection (15 out of 290 AFLP loci), and each marks only a small genomic region (<10 kb), then the total number of selected loci must be very large (>1,000 given a genome size $\sim 10^9$ bp). This is, of course, a very rough extrapolation from a small sample. It also ignores both the limited power of population genomics to detect weakly selected regions (implying that the number of selected loci is under-estimated) and the potentially-high FDR (which suggests that the number is over-estimated). It will be interesting to see how these estimates are influenced by data from the characterization of additional outliers.

In their influential study of sympatric speciation in *Howea* palms, Savolainen et al. (2006) identified four outliers among 274 AFLP loci. Given the high mean F_{ST} in this case (0.31), it is particularly difficult to extrapolate from this observation to the probable number of loci under selection (because power declines and FDR increases with increasing mean F_{ST}, Singleton 2008). A simulation inspired by the *Howea* case, by Gavrilets and Vose (2007), found that speciation was most likely when the number of loci determining the key ecological and isolating traits was low, four being the lowest number simulated. This seductive coincidence of numbers is unfortunate. It is really too early to conclude, from the empirical data, even that the number of loci involved in either trait is ‘small’. Savolainen et al. also note that the ‘L-shaped’ distribution of F_{ST} values they observe is consistent with sympatric speciation, because gene flow keeps most values low while selection increases divergence at a minority of loci. This is true but should not be considered a test for sympatric divergence since the distribution of F_{ST} among neutral loci is expected to be ‘L-shaped’ in the early stages of allopatric divergence and the long tail may be accentuated by selection in allopatry as well as in sympatry.

A population genomic analysis in the lake whitefish, *Coregonus clupeaformis*, has been followed-up in different ways. Two morphs of whitefish, normal and dwarf, co-exist in North American postglacial lakes and differ in morphological, life-history and behavioural traits that are believed to be adaptive in their contrasting environments, benthic and limnetic respectively. An initial study by Campbell and Bernatchez (2004) used 440 AFLP loci in four sympatric morph pairs and found 2–4% outlier loci. Subsequently, Rogers and Bernatchez (2005, 2007) have asked to what extent selected loci identified by the population genomic approach correspond to quantitative trait loci (QTL) for the putatively adaptive traits. Among those AFLP loci that were included in both population genomic and QTL experiments, significantly more outliers were associated with QTL (i.e. within 1.5 LOD support limits of the inferred QTL position) than expected by chance. This was true for backcrosses to both parental types and primarily for growth QTL, although associations were also revealed for other putatively adaptive traits. Segregation distortion in genomic regions associated with one-third to one half of QTL, and with some outlier loci, indicated selection resulting from genetic incompatibility. This analysis suggests that at least a proportion of the outlier loci identified by the population genomics approach experience selection through their effects on quantitative traits. However, the limited power of both QTL and outlier approaches makes it difficult to quantify this association: outliers that are not associated with QTL might indicate other sources of selection but they might also mark QTL that were not detected in the backcross experiments. It might be preferable to find ways of combining the two methods of analysis, rather than performing them separately and then comparing results.

A new study (Via and West 2008) takes the combined outlier and QTL analysis much further by estimating the decline in the detection of outliers with genetic distance from QTL for key components of reproductive isolation in the pea aphid, *Acyrthosiphon pisum*. They find surprisingly large islands of differentiation extending $\sim$10 cM on either side of QTL. Since selection for host-plant association is strong ($\sim$0.1 per locus), this is actually consistent with the predictions of the Charlesworth et al. (1997) model. The implications for sympatric speciation are considerable because reduction in gene flow over these large genomic distances makes the accumulation of further differentiation between host races much easier. The large islands result from a form of hitchhiking, dubbed 'divergence hitchhiking' by Via and West. Hitchhiking extends further than it would in an undivided population because recombination requires movement between hosts, survival and interbreeding to bring haplotypes from the two races into the same individual. Since these events are rare, the effective rate of recombination is much reduced.

Few other studies have mapped outliers and considered their distribution across the genome. Comparing larch budworm, *Zeiraphera diniana*, from two alternative hosts, Emelianov et al. (2004) showed significant heterogeneity in divergence among chromosomes, which was consistent across replicate comparisons. This heterogeneity presumably derives from clustering of outlier markers around loci under selection. Genome-wide comparisons have been possible in two cases with partial reproductive isolation: the M and S forms of the mosquito, *Anopheles gambiae* (Turner et al. 2005) and the subspecies of house mouse, *Mus musculus musculus* and *M. m. domesticus* (Harr 2006a). The mosquito study used genomic DNA hybridization to microarrays and detected two well-supported and one weakly-supported region of the genome with greater differentiation than expected from the genome-wide average. These regions are small, containing only 60–70 genes, only some of which are directly implicated in reproductive isolation (Turner and Hahn 2007). Using publicly available SNP databases, Harr (2006a) found eight regions of elevated differentiation on the autosomes (7.5% of the autosomal genome) of mice as well as strong differentiation on the X chromosome, as expected from hybrid zone studies. There is a risk that this analysis was influenced by ascertainment bias (Boursot and Belkhir 2006; Harr 2006b). In both of these genome-wide analyses, sample sizes of individuals were, necessarily, small. Regions of differentiation were detected using a sliding window of markers and this means that small islands of differentiation may have been missed.

A further, novel approach has recently been applied to fish populations by the Bernatchez group, described as a 'transcriptome scan' (Roberge et al. 2007). The outlier principle was applied not to differences in allele frequency between populations but, instead, to differences in gene expression. The study system was a recently-separated pair of Atlantic salmon (*Salmo salar*) populations with significant genetic differentiation ($F_{ST} \approx 0.03$). Gene expression was measured using a 16,000 cDNA clone microarray for a set of half-sib families from each population. This made it possible to treat expression of each locus as a quantitative trait and to calculate the F_{ST} analogue, Q_{ST}. Outliers from the Q_{ST} distribution for all loci with significant heritability of expression levels were then detected using an arbitrary cut-off of the top 1.5% of values. This process identified 16 loci with expression levels that were unusually divergent between populations (average $Q_{ST} = 0.11$) out of 1,044 with significant heritability in expression (modal Q_{ST} close to zero). These 16 loci are candidates for adaptive divergence in expression. It will be interesting to see, in future, whether this approach identifies distinct sets of candidates compared with allele frequency based methods applied to the same populations. Are some loci experiencing adaptive substitutions as well as evolving expression differences or do genome scans identify transcription factors or their recognition sites that control expression of loci identified in transciptome scans?

Vasemagi et al. (2005) argued that gene-associated markers, in their case microsatellites discovered in expressed sequence tags (ESTs) from Atlantic salmon, should be more likely to reveal candidate loci for local adaptation because they are more likely to be closely linked to the direct targets of selection. Surprisingly, they did not find a difference in the proportion of outliers between gene-associated and random genomic markers in their sample of loci, which was strongly biased towards those from ESTs. However, a more recent study has found the expected effect. The sunflowers, *Helianthus annuus* and *H. petiolaris*, exchange genes across hybrid zones in North America. Using locally allopatric populations close to, but not in hybrid zones, Yatabe et al. (2007) considered several factors that might contribute to outlier behaviour across 108 microsatellite loci. Unlike Vasemagi et al., they found EST-associated microsatellites to be outliers more frequently than loci outside genes: in fact, all five outliers were EST-associated. They also found that loci close to chromosomal break points were more likely to be outliers but that being on a re-arranged chromosome, close to a

QTL for a phenotypic trait distinguishing the species or close to a QTL for hybrid male sterility did not influence outlier probability. Yatabe et al. argued that the effects of selected loci extend over relatively small areas of the genome, a result that is consistent with the *Littorina* observations discussed above, even though the *Littorina* outliers are not in coding regions.

Parallel studies of host-plant-associated insect populations (*Timema cristinae* stick-insects: Nosil et al. 2008, *Neochlamisus bebbianae* leaf beetles: Egan et al. 2008) have added a novel dimension to the interpretation of outliers. Partly following a strategy used by Bonin et al. (2006) to identify outliers specifically related to altitude in the frog, *Rana temporaria*, these two studies distinguish several classes of outlier after making all possible pairwise comparisons among populations. Some outliers occur only in comparisons between insect populations from the same host, some only in comparisons between hosts and some in both types of comparison. Some outliers are detected in one comparison only, some in multiple pair-wise comparisons. Those loci that are observed repeatedly in between-host comparisons are arguably the best candidate markers for genomic regions implicated in host-adaptation. Since repeated observation of the same locus is unlikely by chance, as in *Littorina*, it is also unlikely that these repeated outliers are false-positives. Each of the two studies finds a small number of these repeated, different-host outliers in a large sample of AFLP loci.

The separation between outlier and non-outlier loci involves an arbitrary cut-off and it is clear that, in reality, there is a gradual increase in the probability that a locus is influenced by selection (and/or an increase in the strength of selection) as its observed differentiation increases. Therefore, it might be preferable to find ways to combine information across comparisons, without first categorising loci. It may also be true that populations differ in their adaptation to a host or environment, for example because some populations are exposed to more gene flow that others from differently-adapted populations. Based on these principles, Nosil et al. (2008) introduce the concept of 'isolation by adaptation'. This is analogous to isolation by distance but with the spatial distance matrix replaced by a matrix, or matrices, of 'adaptive distance', such as the difference in host preference or the morphological distance between pairs of populations. They then use partial Mantel tests to determine whether genetic distance is correlated with adaptive distance, with or without a correction for spatial distance. If genetic distance is calculated locus-by-locus, it is possible to seek loci with unusually strong isolation-by-adaptation patterns, compared with the overall pattern. Using this approach, Nosil et al. found that 9% of AFLP loci showed nominally significant isolation-by-adaptation. This is nearly twice the proportion expected given the 5% critical probability across many comparisons. Only about one in six of these loci was detected in the outlier analysis suggesting that the approach may be more powerful. However, as currently applied, the test does not make allowance for the effects of drift. One might expect that variance among populations due to drift would be accounted for by controlling the effect of spatial distance but it is common for historical effects to cause departures from isolation-by-distance and these might inflate the variance of the isolation-by-adaptation correlations. An adjustment of the critical *P*-value may be needed, presumably based on simulation.

Future directions

The population genomics approach clearly has the potential to document the progression, from a handful of loci responding to divergent selection pressures through to genomically-widespread barriers to gene exchange that is envisaged in the genic view of speciation. However, progress to date illustrates not just the power but also the problems:

1. Very few studies have truly scanned the genome and those that have done so have used small samples of individuals and so have had limited resolution.
2. For studies using a random sample of markers, the genomic distribution of selected loci and the sizes of the regions marked by individual outliers are unknown in most cases and difficult to infer in non-model organisms. The currently available evidence is equivocal. Two studies suggest small regions of differentiation (Yatabe et al. 2007; Wood et al. 2008) which, in turn, suggests that a very high density of markers is needed if all selected regions are to be detected. However, the study by Via and West (2008) seems to indicate very large genomic islands of differentiation around a few strongly selected loci. These studies differ in several ways, including the stage in divergence of the populations compared and the methods for inferring the sizes of differentiated regions (see Smadja et al. 2008 for further discussion).
3. Classification of loci into two classes—neutral versus influenced by selection—is not only an over-simplification but it is also unreliable. Typically, it is clear that there are some loci influenced by selection but it is difficult to say how many because of uncertainty about the power of the test and about the false discovery rate.
4. There are currently conflicting results concerning the association of selected regions with genes (Vasemagi et al. 2005; Yatabe et al. 2007; Wood et al. 2008).

There are two areas where progress is being made and where further developments can be expected: testing of more explicit hypotheses and identification of outliers. More specific predictions always lead to more powerful tests. There are a number of ways in which this principle can be applied in population genomics. One is illustrated by the classification of outliers into different classes (Nosil et al. 2008; Egan et al. 2008) and the extension to isolation-by-adaptation analysis (Nosil et al. 2008). In future, the first approach will hopefully be freed from the need for an initial classification of markers and more powerful analyses of isolation-by-adaptation will be developed. Another strategy to make tests more specific is to define classes of marker a priori and then to test predictions about the influence of selection on these classes. The studies by Vasemagi et al. (2005) and Yatabe et al. (2007) move in this direction, as does the use of candidate genes by Eveno et al. (2008) and of QTL-linked markers by Mäkinen et al. (2008) (although the latter two papers were not specifically in the context of speciation).

In terms of characterising outliers or, more generally, the loci most likely to be influenced by selection, two broad approaches are illustrated by the studies reviewed here. The first is to test for association with QTLs and so infer the source of the selection pressures on particular outliers or the proportion of outliers explained by selection on particular traits. This has already produced some valuable insights but it suffers from the statistical problems associated with both approaches and is only practical in a limited range of species where QTL analysis is possible. The other approach is a library-based search for the genomic regions marked by outliers, followed by molecular characterisation, as initiated by Wood et al. (2008) for *Littorina*. This strategy can be applied to any organism. However, library production and sequencing of large inserts such as BACs is costly, although less so given recently developed parallel sequencing technology (Poinar et al. 2006), and it may be necessary to screen large genomic regions to resolve the sizes of islands of differentiation and detect all regions under selection (Via and West 2008; Smadja et al. 2008). Moving from characterisation of sequences around outliers to functional understanding is also going to be very difficult in many cases. Nevertheless, the population genomic approach provides a route to the characterisation of the genetic basis of speciation which is likely to continue to provide important insights.

Acknowledgements I am very grateful to Jacob Hoglund and Gernot Segelbacher for organising, and inviting me to attend, the very stimulating workshop from which these proceedings are derived. The workshop was funded by the European Science Foundation. Discussions with Juan Galindo about population genomics approaches have been important in the development of this paper. My work on *Littorina* is in collaboration with John Grahame and has been funded by the BBSRC.

References

Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. Genome Res 12:1805–1814. doi:10.1101/gr.631202

Beaumont MA (2005) Adaptation and speciation: what can F_{ST} tell us? Trends Ecol Evol 20:435–440. doi:10.1016/j.tree.2005.05.017

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13:969–980. doi:10.1111/j.1365-294X.2004.02125.x

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B Biol Sci 263:1619–1626. doi:10.1098/rspb.1996.0237

Bonin A, Taberlet P, Miaud C, Pompanon F (2006) Explorative genome scan to detect candidate loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). Mol Biol Evol 23:773–783. doi:10.1093/molbev/msj087

Boursot P, Belkhir K (2006) Mouse SNPs for evolutionary biology: beware of ascertainment biases. Genome Res 16:1191–1192. doi:10.1101/gr.5541806

Caballero A, Quesada H, Rolán-Alvarez E (2008) Impact of AFLP fragment size homoplasy on the estimation of population genetic diversity and the detection of selective loci. Genetics 179(1): 539–554

Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. Mol Biol Evol 21:945–956. doi:10.1093/molbev/msh101

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet Res 70:155–174. doi:10.1017/S0016672397002954

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates, Sunderland

De Kovel CGF (2006) The power of allele frequency comparisons to detect the footprint of selection in natural and experimental situations. Genet Sel Evol 38:3–23. doi:10.1051/gse:2005024

Egan SR, Nosil P, Funk DJ (2008) Selection and genomic differentiation during ecological speciation: isolating the contributions of host-association via a comparative genome scan of *Neochlamisus bebbianae* leaf beetles. Evolution (online early)

Emelianov I, Marec F, Mallet J (2004) Genomic evidence for divergence with gene flow in host races of the larch budmoth. Proc R Soc Lond B Biol Sci 271:97–105. doi:10.1098/rspb.2003.2574

Eveno E, Collada C, Guevara MA, Leger V, Soto A, Diaz L et al (2008) Contrasting patterns of selection at *Pinus pinaster* Ait. drought stress candidate genes as revealed by genetic differentiation analyses. Mol Biol Evol 25:417–437. doi:10.1093/molbev/msm272

Feder JL, Roethele FB, Filchak K, Niedbalski J, Romero-Severson J (2003) Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. Genetics 163:939–953

Gavrilets S, Vose A (2007) Case studies and mathematical models of ecological speciation. 2. Palms on an oceanic island. Mol Ecol 16:2910–2921. doi:10.1111/j.1365-294X.2007.03304.x

Grahame JW, Wilding CS, Butlin RK (2006) Adaptation to a steep environmental gradient and an associated barrier to gene exchange in *Littorina saxatilis*. Evolution 60:268–278

Harr B (2006a) Genomic islands of differentiation between house mouse subspecies. Genome Res 16:730–737. doi:10.1101/gr.5045006

Harr B (2006b) Regions of high differentiation—worth a check. Genome Res 16:1193–1194. doi:10.1101/gr.5787706

Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 16:980–989. doi:10.1101/gr.5157306

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

Luikart G, England PR, Tallmon D, Jordan S, Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. Nat Rev Genet 4:981–994. doi:10.1038/nrg1226

Mäkinen HS, Cano JM, Merilä J (2008) Identifying footprints of directional and balancing selection in marine and freshwater three-spined stickleback (*Gasterosteus aculeatus*) populations. Mol Ecol 17:3565–3582. doi:10.1111/j.1365-294X.2008.03714.x

Minder AM, Widmer A (2008) A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species. Mol Ecol (online early)

Murray MC, Hare MP (2006) A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. Mol Ecol 15:4229–4242. doi:10.1111/j.1365-294X.2006.03060.x

Nielsen R (2005) Molecular signatures of natural selection. Annu Rev Genet 39:197–218. doi:10.1146/annurev.genet.39.073003.112420

Nosil P, Vines TH, Funk DJ (2005) Perspective: reproductive isolation caused by natural selection against immigrants from divergent habitats. Evolution 59:705–719

Nosil P, Egan SR, Funk DJ (2008) Heterogeneous genomic differentiation between walking-stick ecotypes: “isolation by adaptation” and multiple roles for divergent selection. Evolution 62:316–336. doi:10.1111/j.1558-5646.2007.00299.x

Ortiz-Barrientos D, Reiland J, Hey J, Noor MAF (2002) Recombination and the divergence of hybridizing species. Genetica 116:167–178. doi:10.1023/A:1021296829109

Poinar HN, Schwarz C, Qi J, Shapiro B, MacPhee RDE, Buigues B et al (2006) Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311:392–394. doi:10.1126/science.1123360

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. Nat Rev Genet 6:847–859. doi:10.1038/nrg1707

Pool JE, DuMont VB, Mueller JL, Aquadro CF (2006) A scan of molecular variation leads to the narrow localization of a selective sweep affecting both afrotropical and cosmopolitan populations of *Drosophila melanogaster*. Genetics 172:1093–1105. doi:10.1534/genetics.105.049973

Presgraves DC (2003) A fine-scale genetic analysis of hybrid incompatibilities in *Drosophila*. Genetics 163:955–972

Price T (2007) Speciation in birds. Roberts and Company, Greenwood Village

Roberge C, Guderley H, Bernatchez L (2007) Genomewide identification of genes under directional selection: gene transcription Q_{ST} scan in diverging Atlantic salmon subpopulations. Genetics 177:1011–1022. doi:10.1534/genetics.107.073759

Roelofs W, Glover T, Tang XH, Sreng I, Robbins P, Eckenrode C et al (1987) Sex pheromone production and perception in European corn-borer moths is determined by both autosomal and sex-linked genes. Proc Natl Acad Sci USA 84:7585–7589. doi:10.1073/pnas.84.21.7585

Rogers SM, Bernatchez L (2005) Integrating QTL mapping and genome scans towards the characterization of candidate loci under parallel selection in the lake whitefish (*Coregonus clupeaformis*). Mol Ecol 14:351–361. doi:10.1111/j.1365-294X.2004.02396.x

Rogers SM, Bernatchez L (2007) The genetic architecture of ecological speciation and the association with signatures of selection in natural lake whitefish (*Coregonas* sp., Salmonidae) species pairs. Mol Biol Evol 24:1423–1438. doi:10.1093/molbev/msm066

Savolainen V, Anstett MC, Lexer C, Hutton I, Clarkson JJ, Norup MV et al (2006) Sympatric speciation in palms on an oceanic island. Nature 441:210–213. doi:10.1038/nature04566

Servedio MR, Noor MAF (2003) The role of reinforcement in speciation: theory and data. Annu Rev Ecol Evol Syst 34:339–364. doi:10.1146/annurev.ecolsys.34.011802.132412

Singleton DR (2008) Evaluating the reliability of F_{ST}-based methods at detecting the signature of selection in microsatellite markers linked to a selection-targeted locus. Ph.D. thesis, University of Reading

Smadja C, Galindo J, Butlin RK (2008) Hitching a lift on the road to speciation. Mol Ecol (in press)

Stinchcombe JR, Hoekstra HE (2007) Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. Heredity 100:158–170. doi:10.1038/sj.hdy.6800937

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. Mol Ecol 14:671–688. doi:10.1111/j.1365-294X.2005.02437.x

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16:702–712. doi:10.1101/gr.5105206

Thornton K, Andolfatto P (2006) Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics 172:1607–1619. doi:10.1534/genetics.105.048223

Ting CT, Tsaur SC, Wu ML, Wu CI (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282:1501–1504. doi:10.1126/science.282.5393.1501

Turner TL, Hahn MW (2007) Locus- and population-specific selection and differentiation between incipient species of *Anopheles gambiae*. Mol Biol Evol 24:2132–2138. doi:10.1093/molbev/msm143

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. PLoS Biol 3:1572–1578. doi:10.1371/journal.pbio.0030285

Vasemagi A, Nilsson J, Primmer CR (2005) Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). Mol Biol Evol 22:1067–1076. doi:10.1093/molbev/msi093

Vekemans X, Beauwens T, Lemaire M, Roldan-Ruiz I (2002) Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol Ecol 11:139–151. doi:10.1046/j.0962-1083.2001.01415.x

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Mol Ecol (in press)

Vitalis R, Dawson K, Boursot P, Belkhir K (2003) DetSel 1.0: a computer program to detect markers responding to selection. J Hered 94:429–431. doi:10.1093/jhered/esg083

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M et al (1995) AFLP—a new technique for DNA-fingerprinting. Nucleic Acids Res 23:4407–4414. doi:10.1093/nar/23.21.4407

Wakeley J (1999) Nonequilibrium migration in human history. Genetics 153:1863–1871

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. J Evol Biol 14:611–619. doi:10.1046/j.1420-9101.2001.00304.x

Wood HM, Grahame JW, Humphray S, Rogers J, Butlin RK (2008) Sequence differentiation in regions identified by a genome scan for local adaptation. Mol Ecol 17(13):3123–3135

Wu CI (2001) The genic view of the process of speciation. J Evol Biol 14:851–865. doi:10.1046/j.1420-9101.2001.00335.x

Yatabe Y, Kane NC, Scotti-Saintagne C, Rieseberg LH (2007) Rampant gene exchange across a strong reproductive barrier between the annual sunflowers, *Helianthus annuus* and *H. petiolaris*. Genetics 175:1883–1893. doi:10.1534/genetics.106.064469

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. Mol Ecol 8:907–913. doi:10.1046/j.1365-294x.1999.00620.x