THE POWER AND PROMISE OF POPULATION GENOMICS: FROM GENOTYPING TO GENOME TYPING

Gordon Luikart, Phillip R. England, David Tallmon, Steve Jordan and Pierre Taberlet

Population genomics has the potential to improve studies of evolutionary genetics, molecular ecology and conservation biology, by facilitating the identification of adaptive molecular variation and by improving the estimation of important parameters such as population size, migration rates and phylogenetic relationships. There has been much excitement in the recent literature about the identification of adaptive molecular variation using the population-genomic approach. However, the most useful contribution of the genomics model to population genetics will be improving inferences about population demography and evolutionary history.

RANDOM GENETIC DRIFT
Random fluctuations in allele frequencies between generations owing to sampling effects. It increases as the effective population size decreases.

GENE FLOW
The movement of genes among populations. Often expressed as the proportion of gene copies (or breeding individuals) that are immigrants from a different population.

Laboratoire d'Ecologie Alpine, Génomique des Populations et Biodiversité, CNRS UMR 5553, Université Joseph Fourier, B.P. 53, F-38041 Grenoble, Cedex 9, France. Correspondence to G.L. e-mail: gordon.luikart@ujf-grenoble.fr
doi:10.1038/nrg1226

Population genomics — an emerging discipline and a new paradigm in population genetics[1] — combines genomic concepts and technologies with the population-genetics objective of understanding evolution. The term was apparently first used in a publication about human disease genetics by Gulcher and Stefansson[2], and subsequently has become increasingly popular (for example, see REFS 3–5).

Population genomics can be broadly defined as the simultaneous study of numerous loci or genome regions to better understand the roles of evolutionary processes (such as mutation, RANDOM GENETIC DRIFT, GENE FLOW and natural selection) that influence variation across genomes and populations. This broad definition includes issues ranging from understanding the pattern and degree of genome-wide heterogeneity (for example, chromosomal/positional differences in sequence diversity and recombination rates) to the origins, relationships and demographic history (interpopulation movement rates, relationships and relative divergence dates) of populations using genome-wide sampling (for example, see REFS 6,7).

According to a more narrow definition, as proposed by Black *et al.*[1], population genomics is the use of genome-wide sampling to identify and to separate locus-specific effects (such as selection, mutation, assortive mating and recombination) from genome-wide effects (such as drift or BOTTLENECKS, gene flow and inbreeding) to improve our understanding of MICROEVOLUTION. This is crucial because only genome-wide effects inform us reliably about population demography and phylogenetic history, whereas locus-specific effects help identify genes that are important for fitness and adaptation. An example of a locus-specific effect is directional selection whereby one allele is selected for in population X but another is selected for in population Y. Such selection would generate a large allele-frequency difference (high F_{ST}) at this locus relative to the F_{st} at distant non-linked NEUTRAL LOCI across the genome.

The two main principles of population genomics are that neutral loci across the genome will be similarly affected by demography and the evolutionary history of populations, and that loci under selection will often behave differently and therefore reveal 'outlier' patterns of variation. Consequently, it is extremely important to identify OUTLIER LOCI both to reliably infer population-demographic history (in which case outliers often should be excluded) and to detect selected (adaptive) loci. Selection will also influence linked markers along a chromosome, such that a SELECTION SIGNATURE (outlier effects) can often be detected by genotyping markers

that are scattered across chromosomes (even if the marker is not in the gene that is affected by selection). The selection signature will decay with time owing to recombination, and therefore ancient/historical selection might not be detectable.

Here, we outline four steps that constitute a basic population-genomic approach, which is based on genotyping numerous molecular markers and testing for outlier loci in population data sets. We illustrate the concept of 'outlier loci', discuss recent statistical and molecular genomic approaches that detect outliers (including available computer software), and quantify the magnitude of bias caused by outliers when estimating POPULATION PARAMETERS (for example, migration rates). Complementary population-genomic approaches, such as quantitative trait loci (QTL) mapping in controlled populations, population-based mapping of genes through linkage disequilibrium (LD) and association studies, have been reviewed elsewhere[8–10] and are not discussed here. We conclude with a brief discussion of some important uses of outlier markers for biodiversity conservation; other uses (such as detecting SELECTIVE SWEEPS) have been reviewed elsewhere[11–14].

We focus mainly on non-human and non-model organisms because the increasing availability of genome-scale data sets in this more diverse set of taxa will yield evolutionary insights that are more broadly applicable. In addition, further genome-wide thinking is needed in studies of non-model taxa to improve evolutionary inferences. Readers who are interested in human and medical genomics (pharmacogenomics) should consult REFS 5,11. We also concentrate on molecular marker data — microsatellites, single nucleotide polymorphisms (SNPs) and AMPLIFIED FRAGMENT-LENGTH POLYMORPHISMS (AFLPs) — as they will continue to be the most widely useful markers for non-model organisms in the near future.

The population-genomic approach: step 1

The first step involves sampling dozens or hundreds of individuals from one or more populations (FIG. 1). Large samples are often required to avoid biased estimates of population parameters[9,15]. For many purposes, the sampling should be geographically broad and representative with no '*a priori*' assumptions about population locations or boundaries. This can help avoid bias that is associated with the subjective sampling of perceived (but not true) population units. For example, researchers might sample individuals from a certain geographic location or a certain phenotype that is assumed to, but does not, represent a breeding group[16]. Also, individual-based analyses should be adopted when possible because they treat the individual (not the population) as the operational unit. For example, genetic distance can be calculated between individuals[17] to cluster them into populations (using no *a priori* assumptions) or to test for correlations between the genetic distance and geographic distance between individuals.

With the growing availability of numerous DNA markers and improved statistical methods, individual-based analyses are becoming increasingly feasible[18]. However, both the sampling strategy adopted and the sample size needed are dictated by the question that is being addressed. For example, 30–50 individuals from a single location could be enough to estimate the EFFECTIVE POPULATION SIZE of an isolated population (for example, see REF. 19), but hundreds of individuals might be

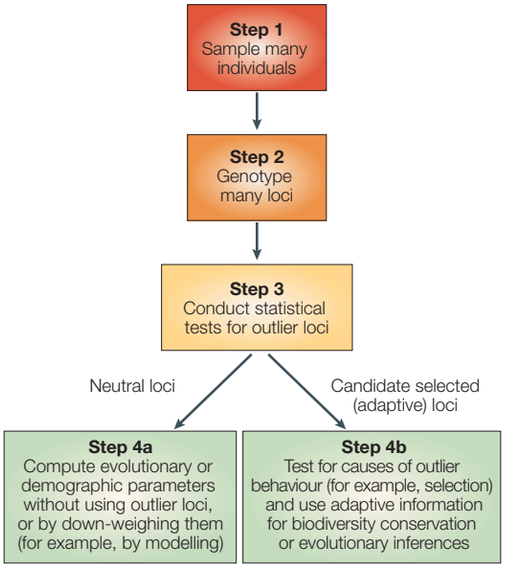


Figure 1 | **Flow chart of the four main steps in the population-genomic approach.** The approach summarized here can be used to identify loci that are under selection (adaptive genes) and to better estimate population history and demography. Step 1, if searching for adaptive variation, is to sample groups of individuals with divergent phenotypes or to sample across a 'selection gradient' (for example, in disease exposure, environmental conditions or phenotype). Large populations are sampled because selection signatures will be detectable only if they are not obscured by drift (small effective population size, N_e). The selection coefficient (s) must be large relative to the N_e for selection to be detectable (for example, $(N_e \times s) > 1$; see REF. 33). Step 2 is to conduct genome-wide genotyping, preferably with mapped loci. If inferring demographic status, independent neutral loci (for example, pseudogenes, random markers and non-coding sequences) are used. If searching for adaptive variation, markers that are in or near genes (ideally of known function and associated with phenotype or environment) are used, as well as many neutral markers. In step 3, outlier loci are those that behave unlike most other loci in the sample; for example, those with an extremely high F_{st} (genetic divergence). Such loci are potentially under selection and could mark adaptive variation; they could also bias estimates of parameters such as gene flow, population size and structure, and therefore should not be used (or be accounted for by modelling). The key to improving many applications of molecular markers in population genetics is the development and validation of improved statistical tests to identify and deal with outlier loci. Step 4a is to estimate N_e, N_m (dispersal rate), F_{st}, structure and phylogenies, to test for bottlenecks, expansion, Hardy–Weinberg and genotypic disequilibrium, and so on. Step 4b is to validate selection as the cause of outlier behaviour — for example, by correlating patterns at outlier loci with selection gradients of environmental variables[6] (disease presence, temperature gradients, predation and so on). Selected markers should be used in studies to better understand adaptation or to plan conservation-management strategies.

BOTTLENECK
A marked reduction in population size that often results in the loss of genetic variation and more frequent matings among closely related individuals.

MICROEVOLUTION
Evolutionary processes or changes over relatively short time periods — such as change in allele frequencies, genotypic composition or gene expression — within or between populations.

F_{ST}
The most widely used index of genetic divergence between populations. A standardized measure of the distribution of genetic variation between populations on a scale between 0 (identical allele frequencies in populations) and 1 (populations fixed for different alleles).

NEUTRAL LOCI
Loci that are not evolving directly in response to selection, the dynamics of which are controlled mainly by genetic drift and migration. These loci can, however, be influenced by selection on nearby (linked) loci.

OUTLIER LOCI
Genome locations (or markers or base pairs) that show behaviour or patterns of variation that are extremely divergent from the rest of the genome (locus-specific effects), as revealed by simulations or statistical tests.

SELECTION SIGNATURE
The molecular footprint of a selection event from the recent past (for example, an excess of rare alleles at a locus relative to the abundance of rare alleles at loci across the rest of the genome).

POPULATION PARAMETERS
Parameters that characterize populations such as gene flow, migration rates, effective size, change in size, relatedness and phylogeny.

SELECTIVE SWEEP
The increase in frequency of an allele (and closely linked chromosomal segments) that is caused by selection for the allele. Sweeps initially reduce variation and subsequently lead to a local excess of rare alleles (homozygosity excess) as new unique mutations accumulate.

needed from each of several independent sets of populations that span a geographic selection gradient (such as a latitudinal gradient of increasing selection for a certain trait) to estimate adaptive genetic differentiation[6] (FIG. 1, step 1).

The population-genomic approach: step 2

The second step involves genotyping tens to hundreds of marker loci, including many putative neutral loci from across the genome. Recent advances in molecular technologies for GENOME TYPING (BOX 1) have made genome-wide sampling in populations feasible. Genotyping many neutral loci is not only a prerequisite for accurately inferring demographic history, but also for providing a baseline (the neutral baseline distribution) to test for statistical outlier loci (step 3) that are potentially under selection. To find neutral marker loci, researchers can choose markers that are located far away from known genes, if genome maps or sequences are available; for example, 50 independent non-coding sequences (far from genes) were used to estimate ancestral population sizes of humans and related primates[20].

Finding neutral marker loci in non-model organisms generally requires the genotyping of many arbitrary (unmapped) loci, followed by statistical tests (see below) to confirm that they are neutral. However, when searching for selection signatures — for example, to study adaptive loci —researchers should also genotype many strong candidate genes (preferably

AMPLIFIED FRAGMENT LENGTH POLYMORPHISM (AFLP). A DNA fragment-length polymorphism that is revealed by a PCR-based DNA fingerprinting technique that generates dozens of polymorphic marker bands (presence or absence of a restriction enzyme site) in a single gel lane. The marker bands are usually dominant in that we generally cannot see the difference between a heterozygote and homozygote.

EFFECTIVE POPULATION SIZE (N_e). Roughly the number of breeding individuals that produce offspring that live to reproductive age. It influences the rate of loss of genetic variation, the efficiency of natural selection, and the accumulation of beneficial and deleterious mutations. It is frequently much smaller than the number of individuals in a population.

GENOME TYPING
The simultaneous genotyping of hundreds of loci from across the genome, which ideally includes mapped loci and different classes of loci such as allozymes, microsatellites and AFLPs, or synonymous (non-coding) and non-synonymous nucleotide polymorphisms.

SEMI-MODEL SPECIES
Species that are not as extensively studied as classical model systems such as mice, *Arabidopsis* and *Drosophila*, but for which large data sets and effective genomic tools are beginning to be developed.

EXPRESSED SEQUENCE TAGS (ESTs). Short DNA sequences (several hundred base pairs) that are produced by reverse transcription of mRNA into DNA. ESTs are cDNAs that consist of exons and the sequences that flank exons. The sequencing of ESTs allows rapid identification ('tagging') of genes and can expedite DNA marker (SNP) development in coding genes.

COMPARATIVE ANCHOR-TAG SEQUENCES (CATS). Exon sequences that are conserved across taxa allowing the design of primers that amplify in divergent species (for example, across mammal orders). CATS-like primers speed the discovery of SNPs (in exons or introns) and comparative genome mapping across taxa.

Box 1 | **Molecular markers for population genomics and genome typing**

The ideal molecular approach for population genomics should uncover hundreds of polymorphic markers (microsatellites and amplified fragment-length polymorphisms (AFLPs), or synonymous (non-coding) and non-synonymous nucleotide polymorphisms) that cover the entire genome in a single, simple and reliable experiment. Unfortunately, at present there is no such approach, although AFLPs[64] and diversity array technology (DArT[65]) partially fulfil these requirements.

The classical AFLP protocol

Selective PCR is used to produce hundreds of polymorphic markers that cover the entire genome, although AFLPs sometimes cluster around centromeres[66,67]. AFLPs are increasingly used to identify markers that are associated with traits that are under selection in non-model plant[68] and vertebrate[69] species (L. Bernatchez, manuscript in preparation).

Variants of the classical AFLP protocol

These methods use either three restriction enzymes (TE-AFLP)[70], one primer that contains a conserved sequence of a gene family (gene-targeted AFLP) or primers in widely-dispersed repeated sequences such as small inserted nuclear elements (SINEs; for example, Alu repeats)[68,71,72]. Unlike the classical AFLP protocol, the SINE-based approach requires only a single PCR. Gene-targeted AFLP can facilitate the detection of selection signatures and adaptive genes. Gene targeting (or avoidance) can also be facilitated by using GC-rich (or GC-poor) restriction enzymes, which tend to cut genomic DNA in gene-rich (or gene-poor) regions.

The DArT approach

This genotyping approach uses a microarray, in which each 'spot' contains a DNA fragment that has been amplified from a library of polymorphic markers that were identified during an initial screening phase[65]. DArT is attractive because a single PCR can amplify hundreds of polymorphic markers and because automation is easier using images rather than gel electrophoresis. Both AFLP and DArT need to be adjusted according to the genome complexity (for example, by digesting the genomic DNA with further restriction enzymes and/or by using selective nucleotides on the 3′ end of the primers). Unfortunately, AFLP-based methods, and perhaps DArT, mainly yield dominant markers, which are less informative and for which there are fewer software programs compared with co-dominant microsatellites and single nucleotide polymorphisms (SNPs).

Microsatellites or SNPs

The development of hundreds of microsatellite and SNP markers is time-consuming and expensive, and the genotyping of microsatellites would require too many DNA amplifications to be competitive with methods that allow a 'massively parallel' analysis (for example, AFLP and DArT). The rapid development of numerous SNPs (including non-synonymous and functional SNPs[73]) is becoming feasible in some non-model species and SEMI-MODEL SPECIES[74,75] (such as mammals, salmonids, agricultural plants and some insects), owing to the rapid growth of EXPRESSED SEQUENCE TAG (EST) databases, data-mining software[76] and primer-design strategies such as COMPARATIVE ANCHOR-TAG SEQUENCES (CATS) and EXON-PRIMED INTRON-CROSSING PCR (EPIC-PCR). Recent improvements in SNP genotyping technology[77] make SNPs attractive for population genomics (REF. 78 and P. A. Morin, G.L. and R. K. Wayne, manuscript in preparation; for example, see Illumina in online links box). A drawback of SNPs is that they are prone to severe ascertainment bias — bias in estimating population parameters — which arises when choosing markers on the basis of their polymorphism level, identifying SNPs using few individuals or transferring markers between populations[79–84].

Sequence data

In 5–10 years, the generation of sequence data might be affordable enough to be used to study numerous loci in hundreds of individuals from non-model species. Sequences are desirable because ascertainment bias is avoided, haplotypes can be identified (or inferred), and coalescent times and allele relatedness (genealogies) can be estimated. Difficulties with sequencing include the analysis of heterozygous sites and insertion/deletion polymorphisms[85].

mapped, with known function and including sites such as receptors and regulatory sequences). In non-model organisms, for which candidate genes might be unknown, gene-rich regions can be preferentially screened (BOX 1).

Even if only 10–20 loci are genotyped, researchers should test for outliers to avoid biased estimates of population parameters. This has seldom been done in published works, which is unfortunate because a single outlier can have an effect on evolutionary inference, even when using relatively few loci (see examples below). LD patterns across the genome can influence the number of markers that must be screened to achieve reasonable power for detecting selection signatures. In some species — for example, *Drosophila melanogaster* — LD typically decays within a few hundred nucleotides after many generations of recombination, so only recent selective sweeps (within the past tens of generations) might be detectable, even if thousands of random or mapped markers are screened.

The recent discovery of HAPLOTYPE BLOCKS[9] indicates that genome typing might be reasonably successful, because if large blocks are shared across populations, then relatively few loci need to be screened to achieve genome-wide coverage. For example, 'haplotype tagging' (genotyping of only a few markers per large block) could be an efficient strategy for genome-wide scans to detect selection. However, shared haplotype blocks have not been investigated in non-model taxa[9,21] and haplotype tagging requires mapping studies that are rare in non-model taxa (for more discussion of LD, see REFS 8,9).

The population-genomic approach: step 3

The third step — testing for outlier loci (FIG. 1, step 3) — is perhaps the most important, because most applications of molecular markers in population genetics require the use of neutral loci (and loci that are inherited according to Mendelian laws and are in HARDY–WEINBERG proportions). Aberrant behaviour of a locus can range from having exceptionally high or low F_{st} between populations (FIG. 2), to having an excess or deficit of low frequency alleles in a population — a HOMOZYGOSITY EXCESS or homozygosity deficit, respectively (FIG. 3). Another locus-specific behaviour is an excess or deficit of heterozygous genotypes (F_{is}-outliers, where F_{is} is an index of deviation from Hardy–Weinberg proportions at a locus; FIG. 2), which is typically tested for at individual loci using standard statistical tests for Hardy–Weinberg proportions (and without consideration of the average genome-wide F_{is}). Other practical applications of tests for outlier loci are briefly discussed in BOX 2.

Testing for outlier loci is important because outliers are likely to be fairly common across data sets, even if they are rare within data sets. There are also other reasons. First, natural selection is often strong in wild populations[22,23] and can be strong in genome regions — with SELECTION COEFFICIENTS of up to 0.69 (REF. 24) — but there are few estimates of selection coefficients for individual loci[25]. Second, reports of molecular selection signatures are increasingly common[26,27], although some might represent false positives owing to violations of assumptions and oversimplified models (for example, with no population structure) that test for selection. Third, selective sweeps can cause LD across large chromosomal regions, thereby increasing the likelihood that many (linked) markers will behave as non-neutral outliers; this is true especially if background LD is high — as occurs in structured, ADMIXED or bottlenecked populations, and in species with close inbreeding. Fourth, data sets are becoming larger (including tens to hundreds of loci), thereby increasing the likelihood that some marker loci will fall in or near selected genes. Fifth, the risks of genotyping errors (which can generate outlier effects) can be high, especially if numerous markers are used or DNA quality is poor (for example, see REF. 28). Sixth, many phenomena, as well as genotyping errors, can cause outlier behaviour, including null alleles (for example, non-amplified alleles), aberrant mutation rates or

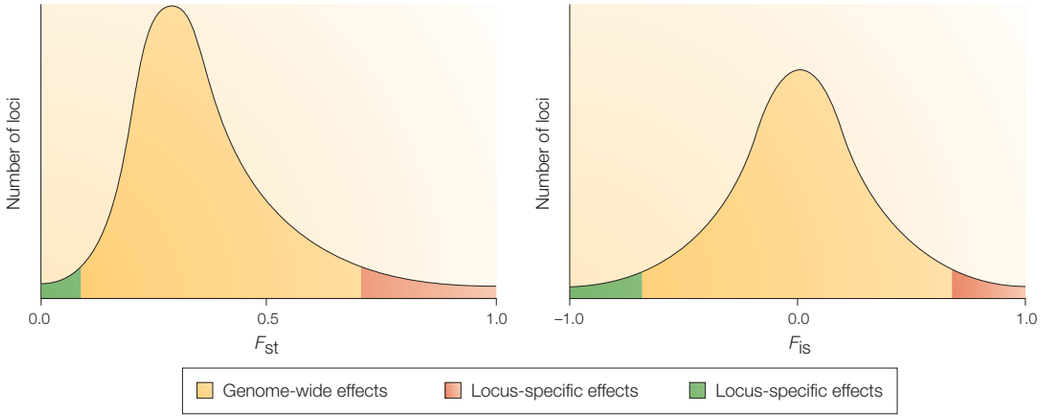


Figure 2 | **Identifying outlier behaviour.** A hypothetical distribution of F_{st} (genetic divergence) and F_{is} (deviation from Hardy–Weinberg proportions) among neutral loci that are sampled from across the genome. Locus-specific effects lead to a few outlier loci with a highly divergent F_{st} or F_{is} value relative to most other loci across the genome. Modified with permission from REF. 1 © (2001) Annual Reviews.

EXON-PRIMED INTRON-CROSSING PCR (EPIC-PCR). EPIC primers are designed in conserved exons and amplify intron sequences that are generally more polymorphic than exons, which are therefore useful for the development of SNP or RFLP markers.

HAPLOTYPE BLOCKS
Long stretches (tens of megabases) along a chromosome that have low recombination rates (and relatively few haplotypes). Adjacent blocks are separated by recombination hot spots (short regions with high recombination rates).

HARDY–WEINBERG
A law or model in which allele and genotype frequencies will reach equilibrium in one generation and remain constant from generation to generation in large random-mating populations with no mutation, migration or selection.

HOMOZYGOSITY EXCESS
A higher Hardy–Weinberg equilibrium homozygosity than that which is expected in a population at mutation–drift equilibrium with the same observed number of alleles. This is not an excess of homozygotes (deviation from Hardy–Weinberg proportions).

SELECTION COEFFICIENTS
The average proportional reduction in fitness of one genotype relative to another owing to selection (designated by '*s*').

ADMIXED
(Hybridized). An admixed population contains hybrids or offspring of individuals originating from genetically divergent parental populations.

CLINE
A gradient of variation across space. It usually refers to increased differences among populations in the frequency of an allele or trait with increased geographic distance.

EMPIRICAL DISTRIBUTION
The distribution of a test statistic (for example, F_{st} or F_{is}) that is computed from observed data obtained from hundreds of loci sampled genome-wide.

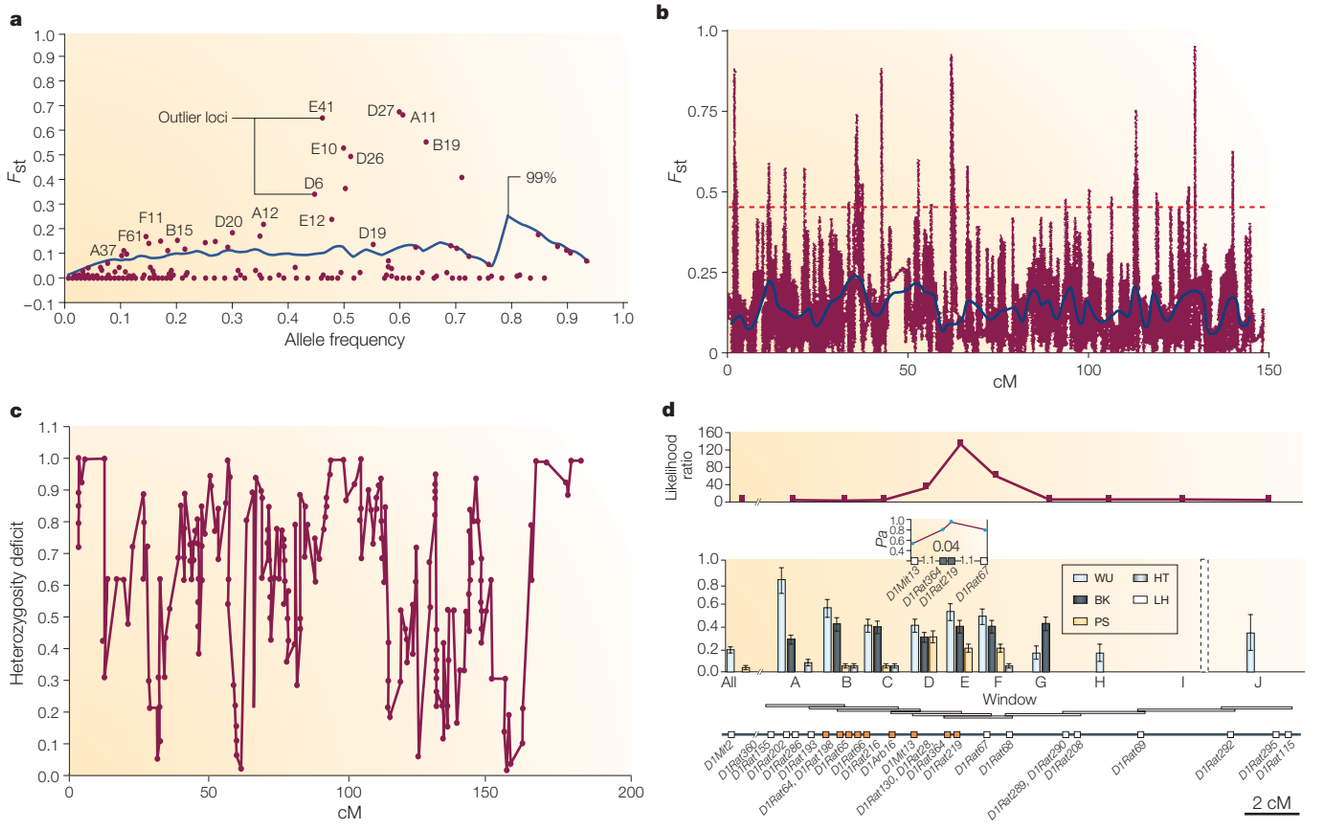


Figure 3 | **Examples of outlier behaviour. a** | F_{st}-outlier amplified fragment-length polymorphisms (AFLPs). Wilding *et al.*[6] genotyped 306 AFLP loci in the intertidal snail *Littorina saxatilis*, the populations of which show a CLINE in shell shape across rocky ocean shorelines. Fifteen loci (dots above the solid line) had extremely high F_{st} (genetic divergence) values (>0.20–0.30) compared with the mean observed F_{st} (<0.04) and with the null distribution of 'neutral' F_{st} values (~0.0–0.2) estimated using computer simulations[6]. The solid line is the upper 99th percentile of the null distribution of simulated neutral loci. The two *L. saxatilis* morphotypes are thin shell and wide aperture (morphotype H), and thick shell and narrow aperture (morphotype M). Reproduced with permission from REF. 6 © (2001) Blackwell Science. **b** | F_{st}-outlier single nucleotide polymorphisms (SNPs) across human chromosome 8 (REF. 35). The horizontal dashed line is a threshold for identifying exceptionally high F_{st} values (>0.45), which represent the upper 2.5% of the EMPIRICAL DISTRIBUTION of F_{st} values; the lower 2.5% threshold is approximately $F_{st} = 0$ (for at least two tightly-linked SNPs). Reproduced with permission from REF. 35 © (2002) Cold Spring Harbor Laboratory Press. **c** | Heterozygosity deficiency (homozygosity excess) and outlier microsatellites. A sliding-window plot of P-values from tests for a locus-specific excess of low-frequency alleles versus genomic position across human chromosome two[36]. It should be noted that heterozygosity deficiency (homozygosity excess) is a typical genome-wide signature of population expansion, but also a locus-specific signature of selective sweeps and directional selection. Therefore, a sliding-window approach that detects a heterozygosity deficit (relative to the whole genome) can be used to identify regions that are potentially under positive directional selection. Reproduced with permission from REF. 36 © (2002) Oxford University Press. **d** | Linkage disequilibrium (LD) outlier microsatellites across chromosome 1 in rat populations that are resistant to warfarin poison. For each window (A–J, horizontal overlapping bars), the fraction of locus pairs in LD (within one standard deviation; middle graph, in which *Pa* is the fraction of correctly classified rats) is shown by the histogram bar height. In the three most resistant populations — see bars colour-labelled grey (WU), black (BK) and yellow (PS) — there is high LD near the warfarin-resistance gene in windows D, E and F, but lower LD away from the gene (for example, windows H and I). This is consistent with a selective sweep of the poison-resistance allele to high frequency, and with recombination reducing LD far from the gene[62]. The top graph shows that the likelihood of correctly assigning an individual to its population of origin (through assignment tests) is highest when using loci from near the resistance gene (windows D, E and F), because those loci have a higher F_{st}[11,62]. cM, centimorgans. Reproduced with permission from REF. 62 © (2000) National Academy of Sciences.

NULL DISTRIBUTION
(Neutral distribution). The distribution (or range) of values across which we expect to observe the value of the test statistic if the null hypothesis is true (for example, neutrality). When conducting a standard *t*-test, *t* is the test statistic and the null distribution is the normal (Gaussian) distribution with *t* degrees of freedom.

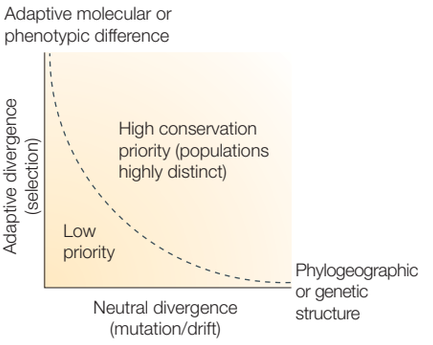
SUMMARY STATISTIC
A parameter estimate (such as F_{st} or F_{is}) that quantifies attributes of the data sampled from a population of interest.

dynamics[29], and aberrant recombination dynamics. Seventh, and finally, marker loci that are in or near important functional genes occasionally experience positive selection and such gene-markers are being used more often (for example, SNPs are being identified in coding genes[30]). Nevertheless, there are still relatively few well-established cases of strong positive selection, even in coding genes[25].

Theoretical and empirical test approaches. There are two general statistical approaches to test for outlier loci: one uses theoretical (simulated) and the other empirical (observed) NULL DISTRIBUTIONS of a SUMMARY STATISTIC such as F_{st} or homozygosity. The empirical approach is used less often because it requires the genotyping of tens to hundreds of loci from across the genome to build a robust null distribution.

Box 2 | Other applications for outlier-loci tests

Tests for outlier loci have other practical applications as well as identifying loci for studying molecular adaptation.



Prioritizing wild populations for conservation
Adaptive markers can be treated differently from neutral markers to identify populations for conservation (see figure; modified with permission from REF. 52 **© (2002) Taylor and Francis). For example, outlier (putative adaptive) loci could be removed to allow the accurate computation of genome-wide (neutral) genetic distinctiveness of populations; outliers that were shown to be genuinely adaptive could then be used as indicators of adaptive differentiation. Subsequently, both neutral and adaptive measures could be integrated to rank populations on the basis of genome-wide neutral and adaptive diversity. Ranking populations for conservation priority is difficult and risky (see main text), but adaptive markers could help to identify the most appropriate source population (the least adaptively differentiated population) from which to translocate individuals into small declining populations that require supplementation.**

Identifying immigrants
Assignment tests can be used to identify 'foreign' (non-resident) multi-locus genotypes[86] (see online links box) and to directly estimate dispersal rates (N_m). Genome screening and the subsequent identification of loci with a high F_{st} (genetic divergence) will increase the efficiency of assignment tests that require high F_{st} to achieve high power[87,88].

Detecting illegal trafficking
Assignment tests and F_{st}-outlier loci can also be used to identify the provenance of trafficked material[89] such as endangered species, animal parts, plants and plant-derived drugs (for example, marijuana).

Detecting fraudulent food products
High-F_{st}-outlier loci can be used to detect fraudulent foods[88,90]. Maudet *et al.*[90] sequenced thousands of nucleotides from many coding genes to find high-F_{st} single nucleotide polymorphisms (SNPs) to differentiate between the common black Holstein and the rare red French alpine cattle breeds. Several SNPs in the *MC1R* (colour receptor) gene had an F_{st} of almost 1.0, whereas most DNA polymorphisms between European cattle breeds have an F_{st} ~0.10–0.20. The high-F_{st} SNPs are now used in France to detect the fraudulent use of cheap Holstein milk in place of expensive mountain-breed milk to make speciality cheeses.

One theoretical approach, which is well illustrated by the classical analytical equations of the Ewens–Watterson test for neutrality at a locus[31,32], combines analytical models and mathematical formulae. Ewens[31] derived an equation that gives the number of different alleles (n) expected in a sample of size $2N$ individuals, given the observed gene diversity (expected heterozygosity) and assuming mutation–drift equilibrium (stable population size). Watterson[32] extended this work and developed a formal statistical test using a theoretical null distribution of homozygosity (the skew of allele-frequency distribution) for an equilibrium population[33].

Another theoretical approach, best illustrated by the F_{st}-outlier test[34] of Beaumont and Nichols (Fdist), uses computer simulations to model neutral loci. The advantage of this approach is that many different population structures and histories can be simulated to assess the influence of different demographic and non-equilibrium scenarios. Software programs that are based on simulations are available to identify F_{st} outliers if either co-dominant loci (allozymes, SNPs and microsatellites) or dominant loci (AFLPs) are used (TABLE 1). This approach was used to identify 15 outlier AFLP loci in a population of intertidal snails[6] (FIG. 3a).

The empirical approach is perhaps best illustrated by the impressive study of Akey *et al.*[35], who calculated the F_{st} for >26,000 SNPs sampled genome-wide in humans. This approach has the advantage of controlling for demographic effects, which can cause outlier behaviour that is similar to selection (for example, an excess of low frequency alleles; FIG. 3c). The disadvantage of this approach is that it requires numerous loci — more than have been available in most non-model organisms.

Single-population tests. Software programs are available to conduct the Ewens–Watterson homozygosity-excess test at individual loci (TABLE 1). This test has been modified for genome-wide sampling and a population-genomics approach by Payseur *et al.*[36], who used 5,257 microsatellites mapped across the human genome and found a genome-wide pattern of homozygosity excess that was consistent with the known recent human population expansion. A genome-wide homozygosity excess is expected in expanding populations because such populations accumulate rare alleles (and so a homozygosity excess) because new (unique) mutations are seldom lost through drift, which is weak in expanding populations. Payseur *et al.*[36] also identified outlier loci (relative to the genome-wide pattern) that might be targets of selection (FIG. 3c). Another approach for detecting homozygosity excess was recently developed by Storz and Beaumont[37]. This BAYESIAN multi-locus test

BAYESIAN
A framework of statistical inference in which previous beliefs (or data) and likelihoods are combined to estimate a parameter of interest given the observed data.

can detect aberrant loci while simultaneously testing for population expansion.

Multiple-population tests. The most widely used 'outlier tests' are multiple-population tests, because of the availability of software programs (TABLE 1) and perhaps because these tests can be used to identify loci that are implicated in adaptive divergence or speciation[6,35,38]. We therefore focus mainly on these tests. An interpopulation F_{st}-outlier test was first suggested by Lewontin and Krakauer[38]. Recent advances in statistics and in computer technology have made these tests feasible[13,34,39,40] (N. Raufaste and F. Bonhomme, manuscript in preparation) and more widely acceptable as a viable method for genome-wide analysis[11,35]. The software programs DetSel[39] and Neutrallelix[41] (N. Raufaste and F. Bonhomme, manuscript in preparation) are available for two F_{st}-based outlier tests similar to that of Beaumont and Nichols[34] (TABLE 1).

Schlotterer developed another interpopulation outlier test that is known as the lnRv test[13]. The test is based on the idea that the variance in allele length will be reduced for microsatellite loci that are linked to targets of directional selection relative to unlinked (neutral) loci. Importantly, the test does not depend on knowledge of mutation rates at marker loci or demographic history, but it can only be used on microsatellite loci (although similar tests based on gene diversity could be developed for use with any type of loci). The performance analysis by Schlotterer[13] reported that the lnRv test is robust under various demographic and sampling conditions, and can provide reasonably high power to identify selected (outlier) loci when 100 loci are genotyped.

When testing for F_{st}-outlier loci, two or more of the available methods and software programs should be used, because if they identify the same outlier there is more confidence in the results and also because different tests have different and complementary characteristics. For example, it is not surprising that the Fdist and DetSel programs identify different outlier loci from the same data set (see, for example, REF. 39), because they use different summary statistics: Fdist uses F_{st} (the standardized variance of allele frequencies) and DetSel uses F (an estimate of identity by descent — the probability that

Table 1 | **Statistical methods and software for population genomics and identifying outlier loci**

Model/test/program characteristics	Software	References
Multiple-population samples		
Simulate F_{st}-null distribution of neutral loci (multiple population island model*, SMM‡ or IAM§ mutation)	Fdist	34
Simulate F_{st}-null distribution for neutral loci (pairs of populations and IAM mutation only)	DetSel	39
Simulate F_{st}-null distribution for neutral loci (island model‡ of migration, IAM mutation)	Neutrallelix‖	41
Simulate F_{st}-null distribution of neutral loci (like Fdist, but for dominant markers)	Wink150	6
Gaussian¶ distribution as null distribution for lnRv# test (microsatellites only)	NA	13
Single-population sample		
Analytical null distribution for allele-frequency distribution under the standard neutral/equilibrium model (for example, Ewens–Watterson test on single loci)	Popgen, Arlequin	92,93
Likelihood-based evaluation of allele-frequency distribution — that is, homozygosity-excess-like test (coalescent, hierarchical Bayesian and MCMC-based**); locus-specific characteristics estimated simultaneously with population expansion/decline	MsVar	37
Detects F_{is} outliers (nucleotide sites with excessive deviation between the observed and expected number of heterozygotes)	PDFis (of GENOMETEST)	1
Detects LD‡‡ at pairs of nucleotides in a gene	LGENOME	1
General framework software for large-scale multilocus genotype data analysis		
Environment/platform using the R statistical package for graphical analyses of large data sets; computes/tests Hardy–Weinberg balancing or directional selection, haplotype frequencies (and distributions) and LD; useful across populations for one locus or across loci within one population	PyPop (python for population genetics; α-version)	94
Environment/platform of linked programs for automating analysis of sequence data, finding SNPs and computing basic summary statistics (diversity and divergence) and tests for selection (Tadjima's D)	PySNP (α-version)	85

*The island model consists of many subpopulations with the same probability of migrants between all subpopulations. ‡SMM is the stepwise mutation model, in which each mutation will either increase or decrease (with a 50:50 probability) the allele length by a single step (that is, one repeat unit at a microsatellite locus), so back mutation is possible. §IAM is the infinite allele model of mutation, in which all mutations give rise to a new (non-existing) allelic state, so back mutation or homoplasy is not possible. ‖N. Raufaste and F. Bonhomme, manuscript in preparation (see also online links box). ¶A Gaussian distribution is a normal bell-shaped distribution, such as is used in conventional *t*-tests. #lnRv is the natural log of the ratio of variance in allele lengths at a locus between two populations (the ratio is computed between two populations).**MCMC (Markov chain Monte Carlo) algorithms are computer-intensive stochastic-simulation methods for solving the mathematical integration that is necessary to calculate the likelihood distribution (for example, posterior distributions) for a parameter of interest (for example, F_{st} or N_e). ‡‡Linkage disequilibrium (LD) is the non-random association of alleles from different loci. F_{is}, deviation from Hardy–Weinberg; F_{st}, genetic divergence; NA, none available; N_e, effective population size; SNP, single-nucleotide polymorphism.

LIKELIHOOD-BASED TEST Statistical tests that consider how likely the data are given an assumed model.

MARKOV CHAIN MONTE CARLO (MCMC). A simulation-based computational technique for the numerical calculation of likelihoods.

two alleles sampled in a population are identical owing to shared ancestry). A new LIKELIHOOD-BASED TEST that combines characteristics of Fdist[34] and DetSel[39] is being developed (M. Beaumont, manuscript in preparation). This test could be more reliable and powerful because it will use more information from the raw data, unlike approaches that use a single summary statistic. Also, it will be Bayesian and therefore could potentially incorporate other information (for example, on population size or mutation rates) that can further increase power.

Future work. Other summary statistics should be recruited (for example, F_{is} and LD) to improve the detection of selection and the estimation of population parameters. Several of the pitfalls of genome typing will have to be overcome — violation of the assumption of independence among loci (which will be breached if large numbers of loci are used) and the problem that many likelihood-based methods are computationally demanding, often taking days to yield a single estimate. Also, some likelihood-based methods might not be reliable when many loci are used (for example, MARKOV CHAIN MONTE CARLO (MCMC)-based methods might never converge). Therefore, although the use and validation of likelihood-based methods is preferable, it is often not feasible with large genomic data sets.

A promising alternative to full likelihood-based approaches are the emerging so called 'summary statistics' approaches. These approaches use a Bayesian model-based framework and calculate several summary statistics, which can extract nearly all of the information from the data[42]. These computationally efficient methods allow

Table 2 | **Some recent studies showing F_{st}-outlier loci and the bias they cause when estimating N_{em}**

Species	Number of populations	Number of individuals per population*	Number and type of loci	Number of outlier loci (%)	Mean F_{st} with outlier loci (*n*)	Mean F_{st} with non-outlier loci (*n*)	F_{st} bias (%)	N_{em}‡ with all loci	N_{em}‡ without outlier loci	Refs
Mice										
Peromyscus californicus	13	24	17 alloz	2 (12)	0.367 (17)	0.234 (15)	36	0.43	0.82	26
Peromyscus gossypinus	50	50	37 alloz	2 (5)	0.178 (37)	0.089 (35)	50	1.15	2.56	26
Peromyscus maniculatus	7	60	15 alloz	1 (7)	0.050 (15)	0.019 (14)	62	4.75	12.91	26
Peromyscus leucopus	12	28	33 alloz	3 (9)	0.140 (33)	0.115 (30)	18	1.54	1.92	26
Peromyscus polionotus	28	30	15 alloz	3 (20)	0.382 (15)	0.283 (12)	26	0.40	0.63	26
Sockeye salmon										
Oncorhyncus nerka	4	50	26 alloz, RAPD and msat§	1 (4)	0.202 (26)	0.091 (25)	55	0.99	2.50	44
Atlantic cod										
Gadus mohua	6	~100	11 nucl RFLP	1 (9)	0.069 (11)	0.034 (10)	51	3.4	6.6	34,45
Drosophila										
Drosophila melanogaster	15	NA	61 alloz	8 (13)	0.23 (61)	0.17 (53)	26	0.84	1.22	34
Intertidal snails										
Littorina saxatilis	8	50	306 AFLP	15 (5)	0.039**	0.0259**	44**	1.9-3.9‖¶	5.5–308‖#	6
Humans										
Homo sapiens	3	NA	216 msat	2 (1)	D = 1.34 (216)	D = 0.74 (214)	45	Divergence >70,000 yBP	NA	95
Homo sapiens	3	~40‡‡	8,862 SNP (in genes)	253 (of 8,862) (2.8)§§	0.120 = autosomes; 0.195 = X chromsome¶¶	NA	NA	NA	NA	35

*Approximate median number of individuals sampled per population. ‡The number of migrants per generation, assuming the island model of migration. §13 allozymes, 8 microsatellites and 5 RAPDs (the outlier was an allozyme locus with F_{st} = 0.713).‖Range of F_{st} between population pairs each with a different morphotype (within a shoreline site). ¶No phylogeographic pattern (P = 0.37; Mantel test).#Significant phylogeographic structure (P<0.002; Mantel test). **J. W. Grahame, unpublished data. ‡‡12–53 individuals per population, genotyped in 6 different laboratories; the actual median number was difficult to estimate.§§156 genes had exceptionally high F_{st}, 18 had exceptionally low F_{st} and contained 253 SNPs and some genes had many outlier SNPs (for example, a gene with exceptionally low F_{st} had at least 2 SNPs with F_{st} = 0.0 and 1 SNP with F_{st} <0.005); it was difficult to estimate the percentage of outlier SNPs in non-gene-associated regions, as the number of F_{st}-outlier SNPs was not given. ¶¶Mean F_{st} is higher for the X chromosome, which is consistent with its smaller effective size; note that F_{st} was also significantly different between coding, intronic and noncoding SNPs (0.107, 0.118 and 0.123, respectively). AFLP, amplified fragment-length polymorphism; alloz, allozyme; F_{st}, index of genetic divergence; msat, microsatellite; NA, none available; N_{em}, absolute number of migrants (a measure of gene flow); nucl, nuclear; RAPD, random amplified polymorphic DNA; RFLP, restriction fragment-length polymorphism; SNP, single-nucleotide polymorphism; yBP, years before present.

the analysis of large data sets and extensive performance testing (for example, by simulating numerous different population histories and known selection regimes).

The population-genomic approach: step 4

The fourth step involves using validated neutral loci to infer demography or history (FIG. 1, step 4a), or the use of putative selected (outlier) loci for further studies, for example, to understand MOLECULAR ADAPTATION (FIG. 1, step 4a).

The population-genomics approach can improve molecular-based estimates of parameters (FIG. 1, step 4a), such as the effective population size, population growth/decline rates, phylogeny, population structure and dispersal rates, in two general ways. First, it allows the removal or down-weighting (modelling to achieve less influence) of outlier loci, which improves accuracy. Second, it increases the number and genomic coverage, and therefore the precision and accuracy, of the marker-sets used.

Assessing outlier bias. The inclusion of only a few outlier loci among many neutral loci can greatly bias estimates of population evolutionary history and PHYLOGEOGRAPHY. A notable example involves the 306 AFLP loci that were genotyped in the intertidal snail[6]. Phylogenies were constructed both with and without the 15 outlier AFLP loci that were inferred to be under selection using F_{st}-outlier simulation-based tests (TABLE 2). With the 15 outliers, the phylogeny showed that populations with a similar shell shape (and habitat type) grouped together, even though they originated from distant geographic locations. By contrast, without the 15 outliers, the phylogeny grouped the geographically proximate populations together (but not the morphologically similar ones; FIG. 4). A crucial finding was that the same outlier loci were identified in replicate distant geographic locations, which supported the hypothesis that selection caused the outlier behaviour. Only selection (not random drift or sampling error) is likely to cause the same loci to have the same outlier pattern across distant replicate locations.

The intertidal snail example[6] indicates that only non-outlier loci should be used to infer demographic history. It also illustrates the value of replicate sampling across habitats and MORPHOTYPES, and of combining genome-wide sampling with morphological and ecological data. One ambiguity in this study is whether the few F_{st} outliers result from recent selection across an environmental cline or from secondary contact followed by introgression in all but a few genomic regions. This question could be resolved using DNA sequence data (and COALESCENT or genealogical analyses) more reliably than with AFLPs or other categorical markers. Unfortunately, at present, the sequencing of numerous randomly chosen or candidate regions is not economically feasible for most non-model organisms.

The first thorough evaluation of the extent to which outlier loci can influence phylogeny is that of Landry *et al.*[43]. These authors found that a single microsatellite locus (out of 17 loci) with extremely high interpopulation variance (in allele lengths) can greatly bias phylogeny branching patterns and cause overestimation of phylogenetic-tree stability from BOOTSTRAP estimates. They introduced a statistical test to identify loci with excessively large interpopulation variance, and showed that removing such loci eliminates the false inflation of estimates of support for a phylogenetic tree. This and

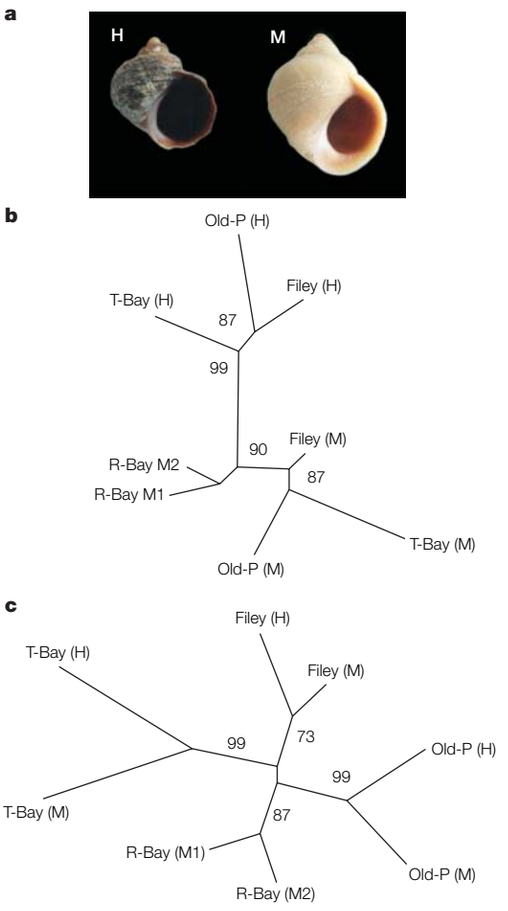


Figure 4 | **The effect of outlier loci on phylogenetic inference.** **a** | The two intertidal snail (*Littorina saxatilis*) morphotypes are thin shell and wide aperture (morphotype H), and thick shell and narrow aperture (morphotype M). The M and H morphotypes are from two habitat types (high and low, respectively, on the littoral zone) that span a strong selection gradient owing to differences in wave action (high) and predatory crabs (low) across the zone. **b** | When using all 306 amplified fragment-length polymorphism (AFLP) loci, including the outlier loci, the phylogenetic tree topology reflects morphotype and habitat type (the M samples from high on the littoral zone cluster together). **c** | When excluding the 15 highly differentiated loci (F_{st}-outlier loci) the phylogenetic tree shape reflects geography, whereby the populations from the same geographic location (Filey, T-Bay, Old-P and RH-Bay) cluster together. The neighbour-joining trees were constructed using AFLP frequency data and PHYLIP[6]. Note the high bootstrap support (numbers on branches, from 100 replicates)[6]. Panel **a** is provided courtesy of J. W. Grahame, Centre for Biodiversity and Conservation, University of Leeds, UK. Panels **b** and **c** are reproduced with permission from REF. 6 © (2001) Blackwell Science.

MOLECULAR ADAPTATION
Genetic change (for example, allele frequency shift or amino-acid substitution) in response to natural selection.

PHYLOGEOGRAPHY
The study of the geographic distribution of phylogenetic lineages, usually within species and to reconstruct the origins and diffusion of lineages.

MORPHOTYPES
Distinctive phenotypes. Organisms that are classified together on the basis of similar physical features without knowledge of their genetic relationships.

COALESCENT
Relating to the mathematical and statistical properties of genealogies. A modelling framework in which two DNA sequence lineages converge in a common ancestral sequence, going backwards in time.

BOOTSTRAP
A statistical approach that is often used to generate confidence intervals (measures of variation) around parameter estimates in which the data are re-sampled repeatedly (with replacement) using computer Monte Carlo simulations.

other studies[44,45] show the importance of applying outlier tests, even when using only 15–20 marker loci.

The effect of outlier loci on estimates of F_{st} in natural populations was recently assessed using seven allozyme data sets from *Peromyscus* mice[26] (TABLE 2). *Peromyscus* is a good genus for such an investigation because several studies have been conducted across the geographic distribution of *Peromyscus* species using many populations (7–50) and allozyme loci (10–37), the segregation and linkage relationships of which have been established. For small terrestrial mammals, *Peromyscus* mice have relatively high gene flow and apparently adaptive geographic population differences in physiology, morphology and behaviour. The Fdist test of Beaumont and Nichols[34] showed that five out of seven data sets had statistically significant F_{st}-outlier loci, with between one and three outlier loci detected per data set. Removal of outlier loci reduced F_{st} estimates substantially by 18–62% (TABLE 2).

Allendorf and Seeb[44] briefly reviewed published studies comparing F_{st} values from different classes of loci such as allozymes, microsatellites and nuclear restriction-fragment-length polymorphisms (RFLPs). They showed that one or a few outlier loci per data set were not uncommon, which was consistent with the findings of Stortz and Nachman[26]. However, in 12 out of 16 studies there was no difference in the mean divergence (F_{st}) estimated from different marker types (proteins versus nuclear DNA). The differences in F_{st} can generally be explained by one or two outlier loci. So, although sampling different marker types (and parts of the genome) is important, it is equally important to test for outlier loci and to sample many loci.

In perhaps the most impressive population-genomics study so far, Akey *et al.*[35] showed the usefulness of genome-wide sampling and of using many mapped loci. The authors quantified the heterogeneity in F_{st} among genome regions and estimated the percentage of SNPs with extreme outlier values. For each of 25,549 autosomal SNPs from three human populations (East Asians, African Americans and European Americans), the average F_{st} was 0.123 (12.3% of the total variation in allele frequencies is the result of interpopulation differences), which is low relative to intercontinental F_{st} values in most other organisms. The authors identified 156 outlier genes that contain SNPs with extremely high F_{st}, and 18 outlier genes with unusually low F_{st}, by studying a subset of 8,862 SNPs located in or near genes.

Interestingly, Akey *et al.*[35] found significant differences in the F_{st} for SNPs from exons versus SNPs outside of coding regions: 0.107 versus 0.123, respectively, which represents ~15% increase in F_{st}. The lower F_{st} in coding regions can be explained by purifying selection that

Box 3 | **Confirming outlier behaviour and selection**

There are several ways to help confirm that a locus is a genuine outlier, rather than a false positive, and to identify the cause of outlier behaviour.

Consider genome position
Markers in or near strong candidate genes — those with a known function that is related to the phenotype or environment being studied, and/or for which selection has been detected previously (for example, in other taxa)[91] — are more likely to be under selection than arbitrary markers or markers that are far from genes. A standard candidate-gene approach might not be useful if there are tens or hundreds of candidate genes, but a strong candidate-gene approach could be fruitful. Markers in or near exons that code for a site of known important function (such as receptors and antigen-binding sites) are likely to be under selection, especially if the function relates to selection pressures (for example, disease or stress) that vary across the study populations.

Conduct complementary population-genomic approaches
For example, quantitative trait loci (QTL) mapping in controlled environments with artificial selection. Bernatchez and colleagues have combined QTL mapping with the F_{st}-outlier approach to confirm that differentiation between sympatric whitefish ecotypes (the *Coregonus clupeaformis* complex) at adaptive QTL-linked loci is maintained by divergent selection (L. Bernatchez, personal communication).

Test for genotyping errors
Genotyping errors (such as false alleles and allelic dropout) can be tested for by re-genotyping samples or by testing for Mendelian segregation in family material.

Genotype across replicate independent populations (or taxa) spanning identical selection gradients
Repeated independent evidence for an outlier locus being correlated with a selection gradient is correlative evidence for selection (for example, see REF. 6).

Confirm support for selection
Genotype extra markers from the same genome region as the outlier/candidate locus.

Conduct significance tests
Both empirical null distributions and simulated distributions, across a wide range of demographic histories, can be used to better understand the potential role of demographic history versus selection in causing outlier behaviour.

Conduct multiple complementary statistical tests for different outlier behaviours
For example, F_{is}, F_{st}, homozygosity excess, gametic and Hardy–Weinberg disequilibrium.

maintains similar allele frequencies for slightly deleterious alleles (which are kept at low frequency) across populations. This study reported ~3% of SNPs with extreme F_{st} values, which was similar to studies on a smaller scale using other markers (TABLE 2). The study also illustrates the importance of knowing the map position of a marker, because most loci on the X chromosome have a high F_{st} relative to the rest of the genome (owing to the smaller effective population size (N_e) and higher drift for the X chromosome). So, chromosome-specific tests for outliers are needed for loci on the X chromosome. Numerous other genome-wide SNP studies for humans and model species have recently been published (for example, see REFS 9,21,46,47), which are not discussed further here.

Outlier bias in published data. To assess the magnitude of bias that can arise from outlier loci, we quantified the F_{st} bias from several real data sets (TABLE 2). Among 11 data sets with F_{st} outliers, one or a few outlier loci were generally detected (~1–10% of all markers per study). Across studies, the magnitude of change in multi-locus F_{st} estimates with and without outliers ranged from 18 to 62% (from the *Peromyscus* studies[26]). Although this bias is large, the effect on estimates of migration rates (N_{em}), and subsequent biological interpretations seem less severe. For example, the largest change in F_{st} (62%) leads to a change in N_{em} from 5 to only 13 (assuming an island model of migration and $F_{st} = 1/4\ N_{em} + 1$). This might not greatly influence biological inferences or management actions (especially if the N_{em} estimates are interpreted with necessary caution[48]). Nonetheless, the AFLP study by Wilding *et al.*[6] indicates that N_{em} estimates with and without outliers could change as much as from 3.9 to 308. This amount of difference would probably influence biological inferences and management strategies.

Identifying causes of outlier loci. Before inferring that a locus is an outlier and discarding it or concluding that it is under selection, it is important to confirm that it is a true outlier — erroneous identification is a substantial risk. This error (a type I error) is likely to arise, for example, when conducting several statistical tests for the study of many loci. When 100 tests with a type-I-error risk of 0.05 are conducted, an average of 5 loci will be erroneously identified as outliers. Also, if an over-simplified or unrealistic simulation model is used (for example, assuming migration–drift equilibrium) to generate the null distribution of neutral F_{st} values, the interlocus variation in F_{st} can be underestimated. This could cause neutral Mendelian loci to be identified as outliers.

Genome-wide heterogeneity in variation can be large even for neutral loci and thereby cause spurious outlier effects. For example, the amount of recombination, diversity and linkage disequilibrium can vary by an order of magnitude across human chromosomes (X chromosome versus autosomes, or centromere versus other regions)[49]. Furthermore, the degree of genome-wide heterogeneity can depend on demographic history; for example, the interlocus variance in F_{st} increases as the rate of gene flow decreases[38].

To avoid making erroneous conclusions from using outlier loci (or from wrongly excluding neutral loci), population parameters should be estimated with and without outliers (especially if outliers are not extremely aberrant). For example, if excluding the outlier makes little difference, the outlier probably can be ignored. Excluding or including weak outliers often makes little difference to biological interpretations (TABLE 2). This is fortunate because marginal or weak outliers might occur often and will be relatively difficult to confirm as being truly aberrant (as opposed to being only random sampling errors).

Confirming and using adaptive molecular variation. Perhaps the most exciting application of the population-genomic approach is identifying adaptive loci to better understand the genetic basis of adaptation and speciation (FIG. 1, step 4b). The study of positive (adaptive) Darwinian selection in natural populations has been notoriously difficult[50]. So, it is exciting that emerging molecular and statistical methods make it increasingly possible to detect and study adaptive molecular change. Population-genomic studies of adaptive molecular variation will improve our understanding of the genetic mechanisms of speciation[1,7,12,51] and will speed up the discovery of genes that are important for health and human medicine[2,4,5]. These topics have been reviewed elsewhere[1,2,4,5,12,51], so here we concentrate on the applications of population genomics in biodiversity conservation, which, although under-appreciated, are becoming increasingly urgent in light of the accelerating extinction crisis.

When searching for adaptive/outlier markers, it is important to state hypotheses and models of evolution *a priori* to avoid subsequent weak (and incorrect) inferences about the cause of selection. 'Fishing expeditions' without *a priori* hypothesis or strong candidate genes are potentially useful, but are susceptible to detecting false positives and drawing erroneous conclusions, because factors other than selection can cause aberrant outlier behaviour (see below).

Conservation biologists have been increasingly interested in identifying adaptive variation to help prioritize populations for conservation[52,53]. The population-genomic approach could help identify adaptive population differences (steps 3 and 4b) that might help a species to survive future environmental changes. Outlier loci, if confirmed as adaptive (BOX 3), could be used to prioritize the selection of populations for conservation, if, for example, they contain a high proportion of adaptive and unique alleles. This approach has several risks. First, it can be extremely difficult to verify whether genes that behave as outliers are genuinely adaptive. Establishing the adaptive importance of a gene can require repeated experiments in replicate taxa or field sites (for example, see REFS 6,26) and in controlled environments such as in the laboratory or in captivity. This is not possible for many threatened species. Association or correlation (not causation) between a marker and a trait or an environmental variable will often be the best evidence for adaptive significance. It will be even more difficult, or impossible, to

obtain a large representative genome-wide sample of adaptive genes, such as is probably required to reliably prioritize populations. The second pitfall of using the population-genomic approach to identify variation that is of concern in conservation is that the adaptive genes detected in samples today might not represent the genes that will be adaptive in future environments: it is difficult to predict which genes will be adaptive in the future. Third, prioritizing certain populations on the basis of high F_{st} values or diversity in a sample of adaptive genes could actually reduce diversity across the rest of the gene pool of a species. This could jeopardize the adaptive potential of a species to future environmental changes[54,55].

One way that candidate adaptive markers might be generally useful in conservation is in choosing the source populations for the translocation of individuals to supplement and rescue a declining population. For example, if there are two candidate source populations, but one has many F_{st}-outlier loci compared with the declining population[6,26], then the source population with many outliers might not be favoured — especially if the outlier loci have a function (for example, in disease resistance).

Another important but under-exploited application of genomic technology in biodiversity conservation is in rapid biodiversity screening and molecular taxonomy[56]. Quick and inexpensive genome typing could greatly speed up the inventory and identification of taxa[57–59], and the delineation of geographic areas[60] for conservation and reserve design. Further development and application of genomic technology for conservation-management purposes is urgently needed to help curb the accelerating extinction crisis.

Another promising application of genomics and 'outlier tests' involves the identification of sets of populations and interpopulation linkages in which the 'process' of adaptive evolution is occurring (for example, local adaptation in the face of high gene flow). Conservationists have called for more efforts to preserve the process of evolution, as well as patterns such as historical population structures. However, such efforts have been hindered, until now, by a lack of molecular and statistical tools for detecting adaptive molecular change and the interactions of gene flow, selection and genetic drift.

An excellent illustration of the usefulness of population genomics for detecting adaptive variation in natural populations is presented by Kohn *et al.*[61,62] (FIG. 3d). They detected outlier behaviour (both high LD and F_{st} outliers) in wild rat populations at microsatellite loci positioned near a poison-resistance gene, but not at microsatellites positioned far from the gene. Populations exposed to rat poison (warfarin) had outlier-selection signatures whereas control populations did not. These studies and others cited above[6,26] illustrate the potential power of population-genomic approaches for detecting selection signatures and adaptive variation, and for studying positive selection in the wild.

Perspectives

The increasing availability of molecular markers will promote the development of genome-wide tests for molecular adaptation and the identification of outlier loci. However, the existing statistical tools need to become more sophisticated and powerful before they can fully exploit the explosion of data that are becoming available for many species. For example, tests that simultaneously incorporate information from physically linked and unlinked markers in gametic disequilibrium (for example, see REF. 63) are needed to fully make use of marker sets that almost saturate the genome. We also urgently need the development and performance testing of more multi-locus tests (other than tests for F_{st} outliers) for detecting outlier behaviour and selection (for example, F_{is} outliers, locus-specific homozygosity excess and so on). Less computer-intensive likelihood-based methods and perhaps methods that are based on several summary statistics[42] will facilitate analyses of large data sets. Little is known about the statistical power of existing tests or their robustness to breached assumptions, variations in population history, marker mutation rates, non-random interlocus associations and sample sizes. Until these are known, the strength of our inferences about genomic patterns and population history will be limited not by data, but by analysis.

The vastly increased availability of molecular markers represents an enormous boon for population genetics, but it can also tempt users to folly. Now that hundreds of markers can be genotyped there is much greater potential to detect evidence of selection in the genome and results will often become statistically significant if enough markers are used. Because TEST STATISTICS (such as F_{st}) from real data will not always follow assumed idealized null distributions, and because statistical evidence is not necessarily biological evidence for selection, false positives are inevitable. Therefore, conservative interpretations of data will be required. Clearly, statistical outliers in large data sets will include both neutral and non-neutral markers, and tests for outliers will miss markers that are evolving in direct response to selection. It will be tempting to infer intricate (and incorrect) biological mechanisms to explain patterns that emerge following data analyses.

To avoid drawing spurious conclusions from large data sets it is important to develop *a priori* hypotheses and models of evolution before carrying out any analysis, because without this approach data analysis is merely a data-mining exercise. Population genomics has the potential to revolutionize the inference of population-demographic history and the detection of adaptation (even pointing to causal processes). It could pave the way for important studies aimed at providing a reliable detailed understanding of the role of selection in the evolution of genomes and populations.

Conclusions

Does population genomics warrant recognition as a new discipline and paradigm? On the one hand, population genomics is nothing new. Geneticists have long realized that analysing only a few loci, or only one class of loci (for example, allozymes), can provide an incomplete or biased view of the genome and of population history or relationships. On the other

TEST STATISTIC
The summary value (often a summary statistic) of a data set that is compared with a statistical distribution to determine whether the data set differs from that expected under a null hypothesis.

hand, only now is it becoming feasible to genotype vast numbers of marker loci (genome typing) in many individuals and populations of non-model organisms. Many statistical methods and computer programs have only recently become available to test for outlier loci and to resolve locus-specific effects versus genome-wide patterns in populations (for example, see REFS 34,37,39).

It is evident from the numerous publications that fail to test for outlier loci before estimating population parameters — the interpretations of which rest heavily on assumptions of neutrality — that the power and promise of population genomics is not fully appreciated among population biologists and geneticists. It can be argued that a conceptual shift that emphasizes a genome-wide perspective is still needed. Embracing a genomic perspective would improve population-genetic studies, including study design (for example, strategic sampling across genomes, populations, phenotypes and environments) and data analysis (testing for outlier loci). Recognition of population genomics as a model could help promote genome-wide thinking, which would improve evolutionary studies.

Molecular technologies are bridging the gap between genotyping and genome typing, which promises to help unlock the secrets of adaptive evolution and to refine inferences about population history. Population genomics will advance our understanding of the genetic basis of fitness, adaptation and speciation, in ways that were impossible only a few years ago. For example, we will have genome-wide studies that estimate the number, map position and relative contribution of the genes that are involved in inbreeding depression, adaptation to extreme climates and the onset of reproductive isolation. The population-genomic approach will speed the discovery, conservation and use of economically important molecular variation in agricultural species by identifying the genes that are important for drought and disease resistance and for milk, meat and grain yield — but also by improving estimates of population size and evolutionary relationships. By providing candidate SNPs (as in Akey *et al.*[35]) population genomics will contribute to the identification of disease-related genes in humans through association studies.

The understanding of adaptive evolution is exciting and important, but improved inference of population parameters and reconstruction of the evolutionary history of populations will probably be the widest influence of population genomics on population genetics. The population-genomic approach will vastly improve the power and sensitivity of many molecular investigations in conservation, ecology and population genetics by ensuring that the assumption of selective neutrality is met by as many markers as possible. Recent advances in molecular and statistical methodology have bolstered the population-genomic approach; nonetheless, statistical methods must mature before they can adequately and reliably deal with the molecular genetic data explosion.

1. Black, W. C., Baer, C. F., Antolin, M. F. & DuTeau, N. M. Population genomics: genome-wide sampling of insect populations. *Annu. Rev. Entomol.* **46**, 441–469 (2001).
This article defines and lays the foundation for population genomics in terms of separating locus-specific effects versus genome-wide effects. It illustrates population-genomic concepts and principles through hypothetical examples and illustrations.
2. Gulcher, J. & Stefansson, K. Population genomics: laying the groundwork for genetic disease modelling and targeting. *Clin. Chem. Lab. Med.* **36**, 523–527 (1998).
3. Goldstein, D. B. & Weale, M. E. Population genomics: linkage disequilibrium holds the key. *Curr. Biol.* **11**, 576–579 (2001).
4. Jorde, L. B., Watkins, W. S. & Bamshad, M. J. Population genomics: a bridge from evolutionary history to genetic medicine. *Hum. Mol. Genet.* **10**, 2199–2207 (2001).
5. Gibson, G. & Mackay, T. F. C. Enabling population and quantitative genomics. *Genet. Res.* **80**, 1–6 (2002).
6. Wilding, C. S., Butlin, R. K. & Grahame, J. Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *J. Evol. Biol.* **14**, 611–619 (2001).
This article indicates that the F_{st}-outlier-detection approach can work surprisingly well if applied to populations that span known selection gradients. It was the first to use AFLP markers, which is encouraging as these are the most readily available markers for genome-wide studies in non-model organisms. One particular strength of this study is the genotyping of replicate sets of populations that span the same kind of selection gradient in different distant geographic locations.
7. Albertson, R. C., Markert, J. A., Danley, P. D. & Kocher, T. D. Phylogeny of a rapidly evolving clade: the cichlid fishes of Lake Malawi, East Africa. *Proc. Natl Acad. Sci. USA* **96**, 5107–5110 (1999).
8. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* **4**, 701–709 (2003).
9. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
10. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet.* **2**, 91–99 (2001).
11. Bamshad, M. & Wooding, S. P. Signatures of natural selection in the human genome. *Nature Rev. Genet.* **4**, 99–111 (2003).
12. Nielsen, R. Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**, 641–647 (2001).
13. Schlötterer, C. A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* **160**, 753–763 (2002).
14. Schlötterer, C. Hitchhiking mapping — functional genomics from the population genetics perspective. *Trends Genet.* **19**, 32–38 (2003).
15. Long, A. D. & Langley, C.H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
16. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
17. Hardy, O. J. & Vekemans, X. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol. Ecol. Notes* **2**, 618–620 (2002).
18. Manel, S., Schwartz, M., Luikart, G. & Taberlet, P. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.* **18**, 189–197 (2003).
This article summarizes the statistical approaches that are available for relating spatial variation in population-genetic patterns to spatial variation in environmental patterns, This article and the population-genomic concepts discussed here show the feasibility of a 'landscape genomic' approach using association studies between the genome and environments.
19. Waples, R. S. Genetic methods for estimating the effective size of cetacean populations. *Report of the International Whaling Commission (Special Issue)* **13**, 279–300 (1991).
20. Yang, Z. Likelihood and Bayes estimation of ancestral population size in hominoids using data from multiple loci. *Genetics* **162**, 1811–1823 (2002).
21. Wiltshire, T. *et al.* Genome-wide single-nucleotide polymorphism analysis defining haplotype patterns in mouse. *Proc. Natl Acad. Sci. USA* **100**, 3380–3385 (2003).
22. Endler, J. A. *Natural Selection in the Wild* (Princeton Univ. Press, Princeton, New Jersey, 1986).
23. Conner, J. K. How strong in natural selection? *Trends Ecol. Evol.* **5**, 215–217 (2001).
24. Ungerer, M. C., Linder, C. R. & Rieseberg, L. H. Effects of genetic background on response to selection in experimental populations of *Arabidopsis thaliana*. *Genetics* **163**, 277–286 (2003).
25. Olson, S. Seeking the signs of selection. *Science* **298**, 1324–1325 (2002).
26. Storz, J. F. & Nachman, M. W. Natural selection on protein polymorphism in the rodent genus *Peromyscus*: evidence from interlocus contrasts. *Evolution* (in the press).
This paper quantifies the potential effects of outlier loci on parameter estimation. The authors suggest that outlier loci are rare within data sets but are fairly common across data sets. They also show that the same loci are often outliers across independent data sets (support for selection as the cause of outlier behaviour).
27. Fay, J. C., Wyckoff, G. J. & Wu, C-I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**, 1024–1026 (2002).
28. Taberlet, P., Waits, L. P. & Luikart, G. Noninvasive genetic sampling: look before you leap. *Trends Ecol. Evol.* **14**, 321–325 (1999).
29. Flint, J. *et al.* Minisatellite mutational processes reduce F-st estimates. *Hum. Genet.* **105**, 567–576 (1999).
30. Sunnucks, P. Efficient genetic markers for population biology. *Trends Ecol. Evol.* **15**, 199–203 (2000).
31. Ewens, W. J. The sampling theory of selectively neutral alleles. *Theoret. Popul. Genet.* **3**, 87–112 (1972).
32. Watterson, G. A. The homozygosity test of neutrality. *Genetics* **88**, 405–417 (1978).
33. Hedrick, P. W. in *Genetics, Demography, and Viability of Fragmented Populations* (eds Young, A. & Clarke, G.) 113–125 (Cambridge Univ. Press, Cambridge, UK, 2000).
34. Beaumont, M. A. & Nichols, R. A. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B* **263**, 1619–1626 (1996).

This paper improves and largely revives the Lewontin and Krakauer F_{st}-outlier approach (reference 38) as a viable method for detecting loci that are candidate selected/adaptive. Real and simulated data (from non-equilibrium populations and various migration patterns) indicate that outliers can be reliably detected. A software program is made freely available to conduct the F_{st}-outlier tests.

35. Akey, J. M., Zhang, G., Zhang, K., Jin, L. & Shriver, M. D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
This study describes the most extensive genome-wide sampling that has been done so far, which provides empirical distributions of F_{st} for different genome regions (X chromosome, exons, introns and non-coding regions).
36. Payseur, B. A., Cutter, A. D. & Nachman, M. W. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol. Biol. Evol.* **7**, 1143–1153 (2002).
37. Storz, J. F. & Beaumont, M. A. Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**, 154–166 (2002).
The first extension of the single-locus homozygosity-excess test (by Ewens–Watterson, references 31 and 32) for use in a genome-wide approach.
38. Lewontin, R. C. & Krakauer, J. K. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* **74**, 175–195 (1973).
39. Vitalis, R., Dawson, K. & Boursot, P. Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811–1823 (2001).
40. Baer, C. F. Among-locus variation in Fst: fish, allozymes and the Lewontin–Krakauer test revisited. *Genetics* **152**, 653–659 (1999).
41. Arnaud-Haond, S., Bonhomme, F. & Blanc, F. Large discrepancies in differentiation of allozymes, nuclear and mitochondrial DNA loci in recently founded Pacific populations of the pearl oyster *Pinctada margeritifera*. *J. Evol. Biol.* **16**, 388–398 (2003).
42. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
43. Landry, P. A., Koskinen, M. T. & Primmer, C. R. Deriving evolutionary relationships among populations using microsatellites and $(\delta\text{-}\mu)^2$: all loci are equal, but some are more equal than others. *Genetics* **161**, 1339–1347 (2002).
44. Allendorf, F. W. & Seeb, L. W. Concordance of genetic divergence among sockeye salmon populations at allozyme, nuclear DNA, and mitochondrial DNA markers. *Evolution* **54**, 640–651 (2000).
This article indicates that outlier loci, although rare within data sets, might be common across large data sets, and that outliers occur with any type of molecular marker. It emphasizes that it is more important to genotype many markers (and test for outliers) than to use a certain marker type when computing population-genetic parameters.
45. Pogson, G. H., Mesa, K. A. & Boutilier, R. G. Genetic population structure and gene flow in the Atlantic cod: a comparison of allozyme and nuclear RFLP loci. *Genetics* **139**, 375–385 (1995).
46. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
47. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analysis in whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
48. Whitlock, M. C. & McCauley, D. E. Indirect measures of gene flow and migration: FST ≠ (4Nm + 1). *Heredity* **82**, 117–125 (1999).
49. Nachman, M. W. Single nucleotide polymorphism and recombination rate in humans. *Trends Genet.* **17**, 481–485 (2001).
50. Hughes, A. L. *Adaptive Evolution of Genes and Genomes* (Oxford Univ. Press, New York and Oxford, 1999).
51. Wu, C-I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
52. Moritz, C. Strategies to protect biological diversity and the evolutionary processes that sustain it. *Syst. Biol.* **51**, 238–254 (2002).
53. Crandall, K. A., Bininda-Emonds, O. R. P., Mace, G. M. & Wayne, R. K. Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.* **15**, 290–295 (2000).
54. Vrijenhoek, R. C. & Leberg, L. P. Let's not throw the baby out with the bathwater: a comment on management for MHC diversity in captive populations. *Cons. Biol.* **5**, 252–253 (1991).
55. Lacy, R. C. Should we select genetic alleles in our conservation breeding programs? *Zoo Biol.* **19**, 279–282 (2000).
56. Wilson, E. O. The encyclopaedia of life. *Trends Ecol. Evol.* **18**, 77–80 (2003).
57. Ronquist, F. & Gardenfors, U. Taxonomy and biodiversity inventories: time to deliver. *Trends Ecol. Evol.* **18**, 269–270 (2003).
58. Baker, S. C., Dalebout, M. L., Lavery, S. & Ross, H. A. DNA-surveillance: applied molecular taxonomy for species conservation and discovery. *Trends Ecol. Evol.* **18**, 271–272 (2003).
59. Blaxter, M. & Floyd, R. Molecular taxonomics for biodiversity surveys: already a reality. *Trends Ecol. Evol.* **18**, 268–269 (2003).
60. DeLong, E. F. Microbial population genomics and ecology. *Curr. Opin. Microbiol.* **5**, 520–524 (2002).
61. Kohn, M. H. *et al.* Locus-specific genetic differentiation among warfarin resistant rat populations. *Genetics* **164**, 1055–1070 (2003).
62. Kohn, M. H., Pelz, H-J. & Wayne, R. K. Natural selection mapping of the warfarin-resistance gene. *Proc. Natl Acad. Sci. USA* **97**, 7911–7915 (2000).
63. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure II. Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
64. Vos, P. *et al.* AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* **23**, 4407–4414 (1995).
65. Jaccoud, D., Peng, K., Feinstein, D. & Kilian, A. Diversity arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* **29**, 25 (2001).
This paper described the DArT approach, which promises to increase the number of RFLP-like markers that can be genotyped in a single PCR by an order of magnitude. The technique uses microarray hybridization, which increases speed and reduces cost.
66. Young, W. P., Schupp, J. M. & Keim, P. DNA methylation and AFLP marker distribution in soybean. *Theor. Appl. Genet.* **99**, 785–792 (1999).
67. Lindner, K. R. *et al.* Gene-centromere mapping of 312 loci in pink salmon by half-tetrad analysis. *Genome* **43**, 538–549 (2000).
68. Skot, L., Sackville, H., Mizen, S., Chorlton, K. H. & Thomas, I. D. Molecular genecology of temperature response in *Lolium perenne*. 2. association of AFLP markers with ecogeography. *Mol. Ecol.* **11**, 1865–1875 (2002).
69. Wang, Z., Baker, A. J., Hill, G. & Edwards, S. V. Reconciling actual and inferred population histories in the house finch (*Carpodacus mexicanus*) by AFLP analysis. *Evolution* (in the press).
70. van der Wurff, A., Chan, Y., van Straalen, N. & Schouten, J. TE–AFLP: combining rapidity and robustness in DNA fingerprinting. *Nucleic Acids Res.* **28**, 105 (2000).
71. van Tienderen, P., de Haan, A., van der Linden, C. & Vosman, B. Biodiversity assessment using markers for ecologically important traits. *Trends Ecol. Evol.* **17**, 577–582 (2002).
Gene-targeted AFLP and other methods for identifying adaptive genes (mainly in agricultural species) are described in this paper.
72. Waugh, R. *et al.* Genetic distribution of *Bare-1*-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). *Mol. Gen. Genet.* **253**, 687–694 (1997).
73. Ramensky, V., Bork, P. & Sunyaev, S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3894–3900 (2002).
74. Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J. & Edwards, D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tags. *Plant Physiol.* **132**, 84–91 (2003).
75. Davey, G. C., Caplice, N. C., Martin, S. A. & Powell, R. A survey of genes expressed in the Atlantic salmon as identified by expressed sequence tags. *Gene* **363**, 121–130 (2001).
76. Everitt, R. *et al.* RED: the analysis, management of and dissemination of expressed sequence tags. *Bioinformatics* **18**, 1692–1693 (2002).
77. Chen, J. W. *et al.* A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res.* **10**, 549–557 (2000).
78. Kennedy, G. *et al.* Large-scale genotyping of complex DNA. *Nature Biotechnol.* **2**, 1233–1237 (2003).
79. Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**, 439–447 (2000).
80. Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. The discovery of single-nucleotide polymorphisms — and inferences about human demographic history. *Am. J. Hum. Genet.* **69**, 1332–1347 (2001).
81. Brumfield, R. T., Beerli, P., Nickerson, D. A. & Edwards, S. V. The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.* **18**, 249–256 (2003).
82. Akey *et al.* The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* **20**, 232–242 (2003).
83. Nielsen, R. & Signorovitch, J. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. *Theor. Popul. Biol.* **63**, 245–255 (2003).
84. Clark, A. *et al.* Linkage disequilibrium and inference of ancestral recombination in 538 single-nucleotide polymorphism clusters across the human genome. *Am. J. Hum. Genet.* **73**, 285–300 (2003).
85. Schmid, K. *et al.* Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res.* **13**, 1250–1257 (2003).
86. Paetkau, D., Slade, R., Burden, M. & Estoup, A. Direct, real-time estimation of migration rates using assignment methods: a simulation-based exploration of accuracy and power. *Mol. Ecol.* (in the press).
87. Banks, M. A., Eichert, W. & Olsen, J. B. Which genetic loci have greater population assignment power? *Bioinformatics* **19**, 1436–1438 (2003).
88. Cornuet, J. M., Piry, S., Luikart, G., Estoup, A. & Solignac, M. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989–2000 (1999).
89. Manel, S., Berthier, P. & Luikart, G. Detecting wildlife poaching: identifying the origin of individuals using Bayesian assignment tests and multi-locus genotypes. *Cons. Biol.* **16**, 650–657 (2002).
90. Maudet, C. & Taberlet, P. Holstein's milk detection in cheeses inferred from melanocortin receptor 1 (*MC1R*) gene polymorphism. *J. Dairy Sci.* **85**, 707–715 (2002).
91. Pletcher, S. D. & Stumpf P. H. Population genomics: ageing by association. *Curr. Biol.* **12**, 328–330 (2002).
This study is an example of how genes cause similar fitness effects in different taxa (humans and mice). This indicates that genes with known adaptive/fitness effects from one species can be used in another species as 'strong candidate genes' in population-genomics association studies.
92. Yeh, F. C., Yang, R-C., Boyle, T. B. J., Ye, Z-H. & Mao, J. X. POPGENE, the user-friendly shareware for population genetic analysis. Molecular Biology and Biotechnology Centre, University of Alberta, Canada. [online], (cited 20 October 2003), <http://www.ualberta.ca/~fyeh/faq.htm> (1997).
93. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes. *Genetics* **131**, 479–491 (1992).
94. Lancaster, A., Nelson, M. P., Single, R. M., Meyer, D. & Thomson, G. in *Pac. Symp. Biocomput. 2003* (eds Altman, R. B. *et al.*) 514–525 (World Scientific, Singapore, 2002).
95. Cooper, G. *et al.* An empirical estimate of the δ-μ genetic distance for 213 human microsatellite markers. *Am. J. Hum. Genet.* **6**, 1125–1133 (1999).

Acknowledgements

We thank F. Allendorf, M. Beaumont, T. Mitchell-Olds, K. Schmidt, P. Sunnucks and three anonymous reviewers for providing references, discussions and helpful comments. W. Amos and J. W. Grahame provided unpublished data and correspondence. S.J. and D.T. were funded by the United States National Science Foundation. G.L, P.R.E and P.T. were supported in part by the European Union ('Econogene' project).

Competing interests statement

The authors declare that they have no competing financial interests.

Online links

DATABASES

The following term in this article is linked online to:
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
MC1R

FURTHER INFORMATION

DetSel software (Renaud Vitalis's web site): http://www.univ-montp2.fr/~genetix/detsel/detsel.html
Fdist software (Mark Beaumont's web site): http://www.rubic.rdg.ac.uk/~mab/software.html
GeneClass2: http://www.montpellier.inra.fr/URLB/GeneClass2/Aide
Illumina: http://www.illumina.com
LECA web site: http://www2.ujf-grenoble.fr/leca
Neutrallelix: http://www.univ-montp2.fr/~genetix/neutrality.htm
PyPop software: http://allele5.biol.berkeley.edu/pypop
Zhenshan Wang's web site: http://depts.washington.edu/scotte/research/postdocs/wang.html
Access to this interactive links box is free online.