STUDY DESIGNS

Making sense of genomic islands of differentiation in light of speciation

Jochen B. W. Wolf[1,2] *and Hans Ellegren*[1]

Abstract | As populations diverge, genetic differences accumulate across the genome. Spurred by rapid developments in sequencing technology, genome-wide population surveys of natural populations promise insights into the evolutionary processes and the genetic basis underlying speciation. Although genomic regions of elevated differentiation are the focus of searches for 'speciation genes', there is an increasing realization that such genomic signatures can also arise by alternative processes that are not related to population divergence, such as linked selection. In this Review, we explore methodological trends in speciation genomic studies, highlight the difficulty in separating processes related to speciation from those emerging from genome-wide properties that are not related to reproductive isolation, and provide a set of suggestions for future work in this area.

Postzygotic intrinsic isolation
Lowered hybrid fitness in the form of sterility or reduced viability of zygotes that are produced in a cross between two groups of individuals, often two species.

Gene flow
Movement of chromosomes (or chromosomal regions) across genetically structured populations, resulting in a change of allele frequencies.

[1]*Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden.*
[2]*Section of Evolutionary Biology, Department of Biology II, Ludwig Maximilian University of Munich, Grosshaderner Strasse 2, Planegg-Martinsried 82152, Germany.*
jochen.wolf@ebc.uu.se; hans.ellegren@ebc.uu.se

doi:10.1038/nrg.2016.133
Published online 14 Nov 2016

The question of how new species originate and adapt to novel environments has been foundational to the field of evolutionary biology[1]. The merging of Darwinian principles with Mendelian genetics in the 'Modern Synthesis' during the first half of the twentieth century laid the conceptual foundation for the field of speciation genetics[2,3], which has the objective of elucidating the molecular underpinnings of species formation. This quest has traditionally been dominated by studies on postzygotic intrinsic isolation that make use of genetic crosses of diverged species in laboratory model systems such as *Drosophila, Arabidopsis* or *Saccharomyces* species[4,5]. At the same time, it has long been recognized that genetic studies are most insightful when they further our understanding of the evolutionary forces that shape natural genetic variation[2,6]. The rapid development of high-throughput sequencing technologies during the past decade has opened this avenue, and speciation genetic research has expanded into wild populations[7,8]. Genome-wide investigation in natural populations is now possible in essentially any organism of choice.

By studying patterns of segregating genetic variation across the genomes of multiple individuals sampled in their natural environment, evolutionary processes governing the accumulation of genetic differences between incipient evolutionary lineages can be 'caught in the act'. In conjunction with ecological and behavioural investigations, genome-wide population surveys hold great promise to provide essential information on the biogeographic history of species[9,10] and address long-standing questions in speciation research. For example, how common is speciation with gene flow (BOX 1)? What is the genetic architecture of reproductive barriers (BOX 2)? What is the timeline of speciation? What is the role of sex chromosomes? What role do chromosomal rearrangements have? In addition, population genomic data more generally contribute to our knowledge of genomic processes that need not be causally related to adaptation and speciation, but that leave 'footprints' that mimic or interfere with signals from adaptation or reproductive isolation[11]. Examples include evidence for recombination rate variation across the genome of individuals and species[12,13], biased gene conversion[14], the link between life history traits and genome evolution[15], and the relative importance of genetic drift, background selection and genetic draft in shaping overall levels of genome-wide genetic diversity[16]. Although interesting in their own right, the genomic footprints generated by these processes complicate the interpretation of genome scans in the context of speciation[17–19].

Originally confined to studies of human evolution, genome scans lie at the core of speciation and adaptation genomic approaches in natural populations[20] and have turned into an industry. It is tempting to draw parallels with the dramatically increased use of mitochondrial DNA (mtDNA) sequencing and microsatellite genotyping in molecular ecology several decades ago. Population genomic studies using genome scans often build on the idea of a 'genic model of speciation' under conditions of gene flow (BOX 1). This model predicts that divergent selection against gene flow is initially confined to few genic elements[21–23]. Allelic variation at loci that are under selection and confer reproductive isolation will be less likely to cross population boundaries than

unlinked, selectively neutral loci[24]. As a result, targets of divergent selection and loci in close linkage with those targets are relatively protected from the homogenizing process of gene flow. Consequently, genetic differentiation between diverging populations is expected to vary along their genomes (FIG. 1).

Although often framed in a strictly ecological setting of divergent selection (or, as a special case, disruptive selection) across an ecological contrast, any large-effect 'barrier locus' contributing to pre-mating, post-mating, prezygotic or postzygotic isolation has the potential to create genomic areas of elevated differentiation relative to the genomic background. Importantly, this elevated differentiation includes genic or non-genic elements that promote meiotic drive or centromeric drive, which have been implicated in speciation[25,26]. Such genomic 'outlier' regions have been referred to as 'differentiation islands' (REF. 27) or more speculatively as 'speciation islands' (REFS 28,29). Their amplitude and width should, in principle, be given by a function of the amount of gene flow, the strength and timing of selection, the recombination rate and the underlying genetic architecture of the trait under selection[22,30] (BOX 2). The complex interplay of these factors is instrumental in deciding whether substantial local genomic differentiation can arise in the first place, and whether it can spread to the whole genome as populations continue diverging and eventually lead to full reproductive isolation sealing off gene flow altogether[31]. Even without adding confounding factors that are not related to adaptation or forms of reproductive isolation, the interaction of these often-unknown parameters makes quantitative predictions of genomic differentiation across the genome anything but straightforward.

Consistent with the genic model of speciation, a number of high-profile studies published recently have revealed strong heterogeneity in genomic differentiation upon population divergence in a variety of taxa[27,28,32–36] (Supplementary information S1 (table)) — a pattern that requires explanation. To describe this pattern, metaphors such as '(heterogeneous) genomic landscape of differentiation' or 'islands or continents of speciation' are commonly used. Although the metaphors are, in principle, agnostic about the underlying process, they often imply

Reproductive isolation
Any mechanism or process that reduces the probability of mating, survival or reproduction between members of different groups and their offspring.

Genetic drift
Random change in allele frequencies between generations as a consequence of stochastic sampling in a finite population.

Background selection
Change in allele frequencies and reduction in diversity at neutral loci as a result of selection against deleterious alleles at linked loci.

Genetic draft
Also known as genetic hitch-hiking. Pervasive reduction of genetic diversity owing to recurrent selective sweeps.

Divergent selection
Natural selection for different trait values in diverging lineages.

Disruptive selection
A special case of divergent selection in which natural selection favours two extreme values of a phenotypic distribution.

Meiotic drive
A mechanism of segregation distortion during female meiosis in which one allele at a locus is transmitted to the offspring (gametes) more often than are other alleles, even in the absence of a selective advantage of that allele.

Centromeric drive
A special case of meiotic drive involving centromeres that are in evolutionary conflict to increase their odds of transmission during asymmetric (female) meiosis.

Introgression
The transfer of genetic information (gene flow) between divergent populations or species as a result of hybridization and repeated backcrossing.

Good species
Well-separated lineages that clearly form distinct species and no longer interbreed.

Box 1 | Speciation with or without gene flow

The question of whether speciation occurs with or without gene flow has long been a central topic in speciation genetics, as it has implications for our understanding of the underlying processes. Depending on the degree of inter-population gene flow, expectations differ about the relative importance of selection versus drift, the role of prezygotic isolation and the genetic architecture of barrier loci[4,31] (also see BOX 2).

By historical convention, the intensity and modality of gene flow are often conceptualized in geographical terms (for example, allopatry, peripatry, parapatry and sympatry), and views on the prevailing geographical mode have changed through time. Charles Darwin saw potential for speciation in large interconnected populations and was comfortable with the idea of many coexisting varieties with intermediate forms[124], which was interpreted as an argument for sympatric speciation[125] or at least an opening for continuity of forms allowing for introgression[126]. A century later, Ernst Mayr popularized the biological species concept and with it the view that speciation with gene flow was bound to be rare, if it occurred at all[127]. Capitalizing on the idea that the evolution of reproductive isolation is an essential component of speciation, Mayr proposed that gene flow has an antagonistic role by counteracting reproductive isolation and by homogenizing gene pools. As a consequence, it seems most straightforward to achieve speciation in isolation. Indeed, genome-wide linkage disequilibrium is readily generated between isolated demes that are free to differentiate in many genomic regions via selection or genetic drift, with speciation resulting as a simple by-product that has an increasing probability of occurring as time progresses (the predicted snowball effect[128]). Central speciation models such as the Bateson–Dobzhansky–Muller (BDM) model[2,129,130] or Oka's model[131], which explain the evolution of hybrid sterility or inviability by negative epistasis between populations, work best in the absence of gene flow.

Whereas Mayr's view dominated the second half of the twentieth century and continues to be the main null hypothesis today, the past two decades have seen a shift in the perception of gene flow as the main antagonist to speciation. Both modelling and empirical work acknowledge the challenge, but also emphasize the plausibility of speciation under conditions of gene flow[22,132], even in sympatry[133–135]. Fortunately, a much-heated debate on the plausibility of speciation for certain geographical modes has given way to a more balanced discussion on how the interplay between central population genetic parameters (such as selection, recombination, genetic drift and migration) may affect the outcome of incipient population divergence[31].

With an unprecedented amount of data characterizing genetic variation within and between natural populations, speciation research is in a unique historical position to gain a more comprehensive view of the role of gene flow in speciation. Methods leveraging genome-wide information to infer the presence and/or amount of gene flow from multi-locus data are becoming increasingly available[67]. Empirical surveys to characterize the extent to which population divergence is accompanied by gene flow are now accessible for a broad range of taxa[8,34]. As studies are accumulating, comparative analyses have clear potential to quantify the frequency of speciation with gene flow[136]. However, there is a limit to what can be achieved with empirical data alone. By paradigm, speciation genetic research is limited to investigating the genetic basis of reproductive isolation in 'good species' or in current-day populations that are in the process of diversification. For good species, lineage sorting has probably been completed, and inference on gene flow during incipient stages is missing. For current-day populations, gene flow during divergence can be estimated, although it remains unclear whether differentiation will result in speciation or whether a certain level of differentiation will perpetuate at an arrested stage[65]. It is therefore important to accompany empirical work with theoretical studies that define the parameter space where speciation can occur even under conditions of gene flow[137–139].

Magic traits
Traits that are subjected to divergent selection and contribute to non-random mating; they facilitate speciation with gene flow.

Supergenes
Clusters of tightly linked loci where two or more haplotypes give distinctly different phenotypes.

F_{ST}
A common statistical measure of genetic differentiation between populations that compares the variance in allele frequencies between populations to the variance within populations. It is sensitive to genetic drift, demographic change, mutation, migration and genetic variation of each population.

Linked selection
Selection that changes allele frequencies, often leading to reduced diversity, at loci genetically linked to the focal locus.

Box 2 | **Speciation with gene flow — the importance of genome architecture**

The genome is not a mere collection of independent genes. Organization into chromosomes (physical linkage) and population processes such as admixture (statistical linkage) introduce non-independence among genes, the extent of which is determined by the recombination rate. In the context of speciation with gene flow it is therefore crucial to consider barrier loci that convey reproductive isolation in a genomic context. The vicinity around loci that experience divergent selection or negative epistatic interactions among populations (in the form of Bateson–Dobzhansky–Muller incompatibilities) may effectively be protected against gene flow (a phenomenon known as divergence hitch-hiking). In this scenario, the rest of the genome is free to introgress if it is sufficiently disassociated by recombination, which generates heterogeneity in the levels of genetic diversity within and among populations[21,38]. Speciation will be facilitated if barrier loci coding for different components of reproductive isolation become coupled. Whether coupling occurs crucially depends on the genetic architecture, including the number of segregating barrier loci, the strength of selection on each locus and the degree of antagonistic recombination[99,132,140]. The combination of these factors will shape the patterns of genomic differentiation and determine whether single peaks of differentiation are expected (divergence hitch-hiking) or whether differentiation can generalize to the entire genome (genome hitch-hiking)[22,138,141].

What do we expect? A broad variety of genetic architectures seem plausible. Magic traits constitute one extreme in which a single trait — in the most extreme case this is encoded by a single gene — can facilitate speciation with gene flow[142,143]. When several genes contribute to reproductive isolation, genomic features such as centromeres or inversions that reduce recombination between populations can promote coupling by creating and maintaining linkage disequilibrium[39,88]. The presence of barrier loci in inverted regions or even their organization in supergenes has been suggested in a variety of taxa[83,144] and may be theoretically expected to be promoted under conditions of gene flow[30]. The upper-end scenario suggested by numerous studies on ecological speciation involves many genes, each with small effect, spread across the genome[36,58].

Ultimately, characterization of the genetic architecture of barrier loci requires substantial empirical effort in a number of systems across several time-points of population divergence. Genome scans have the potential to extract candidate barrier loci, provided that patterns of genetic diversity are interpreted with due caution against alternative explanations. Importantly, however, genome scans have an inherent ascertainment bias, as they more easily detect traits under putative selection that have a simple genetic architecture. This is similarly true for functional validation of traits with complex genetic architectures or gene–environment (G × E) interactions, which will constrain the power of functional genetics approaches in natural populations. It will be important to more fully recognize this limitation and develop alternative approaches to study such complex traits. One pragmatic way forward is to develop comprehensive research programmes that use as much independent information as possible, including genome scans on larger sample sizes than are common today, functional tests, transgenic validation and admixture analyses, in order to elucidate current-day selection acting in zones of contact[56,118].

the genic model of speciation and interpret genomic regions of elevated differentiation as primary hosts for 'speciation genes', or as 'crystallization points' (REF. 37) of reproductive isolation[22,38].

However, local peaks of genomic differentiation need not necessarily arise as a result of divergent selection involving allelic variation of genetic elements that advance reproductive isolation. Alternative explanations are possible in which heterogenic differentiation is not determined by differential gene flow across the genome[17,39] (FIG. 1). As nicely illustrated by the meta-analysis of Cruickshank and Hahn[19], there is increasing evidence that genomic regions with elevated differentiation, which are typically inferred by relative measures of genetic differentiation such as F_{ST}, can emerge by processes that are not related to speciation *per se*. Even in the absence of gene flow, linked selection — in the form of either genetic draft (on loci that may or may not be relevant for speciation)[40,41] or background selection[42,43] — can significantly contribute to heterogeneity in differentiation.

In this Review, we discuss central aspects of studies that gather genome-wide data from natural populations to investigate the genetic basis of reproductive isolation. We highlight trends, illustrate the difficulty in separating processes related to speciation from those emerging from genome-wide properties that are not related to reproductive isolation, and provide a set of suggestions for future work in this area. TABLE 1 provides an overview of the suggested workflow, and indicates pitfalls of and best-practice procedures for this research.

Study designs

We scanned the literature and identified 67 studies that have used genome-wide approaches to investigate genomic differentiation between diverging populations or species. Throughout this Review, these studies form the 'data set' that we used to describe current trends and practices in speciation genomic research. Central aspects of these studies are summarized in Supplementary information S1 (table) and presented in FIG. 2.

Taxonomic distribution

Taxon sampling (at the level of taxonomic class) has been very uneven and strongly dominated by ray-finned fish and insects, followed by birds, mammals, angiosperms and gastropods. This bias reflects the publicly available genomic resources, and probably a general skew in research focus on vertebrates and insects in evolutionary genetics. Moreover, the dominance of ray-finned fish and insects is, to a large extent, explained by a highly prolific literature on sticklebacks, *Drosophila* spp. fruitflies and *Heliconius* spp. butterflies. Research in these speciation models is certainly important, as it allows the consistency of inferred evolutionary processes and patterns to be tested through replication

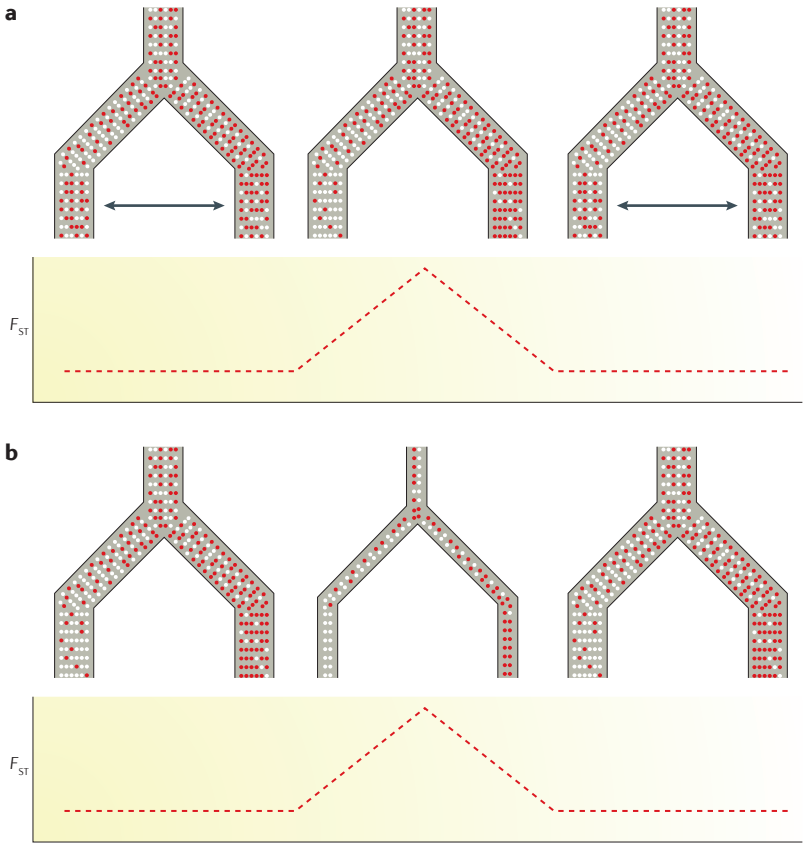

Figure 1 | **A schematic of alternative processes that generate regional genomic islands of elevated differentiation.** Red and white circles represent different alleles in a population at the depicted genomic region. The branching schematic indicates the segregation of these alleles between populations that are diverging. **a** | In regions of gene flow (indicated by the double-headed arrows), differentiation becomes reduced relative to loci where there is selection against gene flow because of reproductive incompatibility, for example. **b** | In regions where the effective population size (N_e) is reduced by processes that are independent from gene flow (middle panel) the rate of lineage sorting is enhanced relative to background levels, leading to elevated differentiation.

across independent populations, pairs of species or subspecies, or ecological contrasts. However, to gain a more general understanding of genome-wide processes that occur during population divergence there is a need for broader taxon sampling, in particular with respect to organisms that span a wide range of effective population sizes (N_e) (and hence a wide range in the efficacy of selection), recombination rates (for example, degree of selfing) and dispersal (and hence expected gene flow). There is potential to build on common interest in research communities that embrace under-represented taxa[44] if impediments resulting from deviant paradigms and terminology can be overcome.

Population sampling regime

Temporal sampling at different levels of population divergence. Time to the most recent common ancestor is a central parameter when studying population divergence. Signatures of selective sweeps that are relevant to population divergence will — depending on the recombination rate, its mutational origin and the selection coefficient — only persist for a short time after the split. By contrast, effects of background selection on local N_e will only be exposed at intermediate divergence times, and measures of absolute divergence such as d_{xy} (BOX 3) are only informative beyond a minimum level of divergence[19]. Moreover, to distinguish cause and consequence of reproductive isolation on patterns of genome-wide variation, temporal resolution is needed to reconstruct the sequence in which reproductive barriers emerge. However, with the exception of experimental microbial systems[45], the temporal progression of speciation cannot be studied in real time.

A promising way forward is to sample several populations at different stages of population divergence, which are sometimes referred to as a 'speciation continuum'. This approach has been used in an increasing number of studies and has yielded valuable insights into the chronology and relative importance of the underlying evolutionary processes[8,46]. For instance, by sampling within and across species, studies focusing on advanced stages of speciation have shown that linked selection has a major role in shaping common differentiation landscapes[33,47]. Other multi-population studies have shown that the very onset of population divergence can be characterized by extreme differentiation peaks confined to a small proportion of the genome[35,48]; these extreme differentiation peaks emerge against heterogeneous background levels of differentiation shaped by linked selection that is common to all populations[9,49–51].

Temporal sampling at different levels of population divergence is particularly crucial to test the 'divergence hitch-hiking model' of speciation with gene flow[52,53]. According to this model, primary regions of elevated differentiation that are protected from gene flow by divergent selection may promote the secondary accumulation of genetic elements under weaker selection, which launches a cascade of local genomic divergence. As a consequence, regions of elevated differentiation would grow in width as time progresses, a prediction that to date has gained limited empirical support[35,50,54,55].

Replication. Population replication is powerful for unveiling parallel selection pressures (or other forces that repeatedly occur) in independent population pairs that have a similar divergence age. In a replicate design across parallel hybrid zones, Nadeau *et al.*[56] exposed commonalities in the genetic architecture of wing patterning within and between *Heliconius* species that confers reproductive isolation. Similarly, Vijay *et al.*[49] studied patterns of genomic differentiation across independent phenotypic contact zones in crows and suggested that the patterns reflected context-dependent selection on a multigenic trait architecture superimposed on a common background of linked selection. Multiple replication in worldwide stickleback populations across a variety of ecological contrasts revealed repeated effects of the same major-effect genes but also peculiarities for each genotypic background[50,57–59].

More generally, population replication is central to our understanding of the importance of context dependence for adaptation and speciation. Theory and experimental evidence in model organisms clearly predict

Effective population sizes (N_e). A somewhat abstract population genetic measure of the size of an idealized population in which the strength of genetic drift is the same as that in the population of interest.

Time to the most recent common ancestor
In genetic genealogy, the time, in years or generations, to the most recent individual from whom all individuals in the sample under consideration descended.

Hybrid zones
Narrow geographical regions where two species or divergent populations are found in close proximity and hybridize.

Epistatic interactions
The interaction between two or more genes that causes a phenotype to be dependent on the particular combination of alleles at these loci.

Coalescent times
The most recent time-point in the past at which two gene copies share a common ancestor.

Structural genetic variation
Polymorphisms involving differences in the length, orientation, order, copy number or chromosomal organization of DNA sequences.

Quantitative trait loci
(QTLs). Genomic regions that are statistically associated with non-discrete variation in a phenotypic trait.

Isolation-by-ecology
Genome-wide differentiation between groups of individuals according to environmental or phenotypic contrasts between populations (rather than, for example, according to geographical distance).

Spatial autocorrelation
Genetic similarities that are attributable to geographical proximity between populations.

that phenotypic variation, and hence selection on this variation, is strongly dependent on the environment (gene–environment (G × E) interaction)[60]. Furthermore, in addition to single-gene effects, epistatic interactions are expected to contribute substantially to reproductive isolation[5,16,45]. Epistatic interactions are context-specific, and will not only depend on the underlying genotypic background (gene–gene (G × G) interaction) but also on the environment (gene–gene–environment (G × G × E) interaction) and on higher-order components[61]. Although clear-cut answers on the genetic architecture of traits under selection will be difficult to obtain from studies in natural populations, it is important to consider this dimension in the design and interpretation of results.

As a final note on the 'speciation continuum', time to the most recent common ancestor cannot be approximated simply by mean levels of genome-wide F_{ST}. Although *F*-statistics reflect mean coalescent times within and between populations[62], they cannot distinguish between common ancestry that is due to a recent population split versus that due to substantial migration between populations. Thus, population comparisons with similar F_{ST} can differ widely in their histories, with clear implications for the evolutionary processes shaping heterogeneity in differentiation across the genome.

A priori contrasts. The genomics of population divergence can be studied by simply quantifying intra-population and inter-population genetic variation between any populations. This 'agnostic' approach has been successfully applied in numerous systems and has exposed the impact of linked selection[47] and structural genetic variation[63,64] on heterogeneity in genomic differentiation. However, in the majority of cases, population comparisons are complemented by a priori contrasts, which generally refer to morphometric parameters (that is, phenotypes suspected to be under selection, often colour patterns) or ecological parameters (for example, habitat contrast and temperature gradients), which can be easily monitored in the field. The advantage of leveraging a priori contrasts lies in the opportunity to experimentally validate loci emerging as candidates from the genomic analyses. Examples include forms of validation via gene expression network analyses[35], through overlap with quantitative trait loci (QTLs) (several studies in stickleback and *Heliconius* spp.), via selection experiments in the wild[36] or by reference to published information on the genetic architecture of the phenotype in question. Although valuable in principle, a priori contrasts need to be applied with some caution. They can misguide interpretation and devalue regions of interest that reflect relevant contrasts that are not considered, and in cases of suboptimal population sampling isolation-by-ecology can be confounded by spatial autocorrelation[65]. More generally, a priori contrasts bear the inherent risk of biasing our perspective to a small subset of biometric or colouration traits that can easily be measured in the field[66]. Importantly, comparisons between populations that do not differ in the contrast of interest always need to be incorporated as a control. As the field matures, this basic principle of experimental design needs to be embraced more thoroughly.

Demographic history

The interpretation of the observed landscape of genomic differentiation directly depends on the timing and amount of gene flow between populations[19]. It is therefore somewhat surprising that a large proportion of speciation genomic studies in our data set (70%) did not explicitly estimate gene flow between the populations in question or they loosely refer to published information that is not necessarily derived from the target populations under study. Considering the increasing

Table 1 | **Speciation genomics workflows and considerations**

Overall workflow	Pitfalls	Best practice
Sampling of DNA from two or more natural populations	Focusing on a geographical sub-sample misrepresents divergence history and does not allow for generalization of the inferred processes	• Sample populations across the 'speciation continuum' • Replicate sampling in independent species pairs
Genome re-sequencing	• Failure to identify crucial signals, or the landscape of signals, when reduced-representation sequencing is done • Failure to associate regions of increased divergence with chromosomal features such as centromeres in the absence of a genome assembly	• Variant calling based on whole-genome, individual-based re-sequencing and read mapping to a genome assembly with scaffolds anchored to chromosomes • Application of long-read technology or other means for inferring structural variation
Population genomic analysis	• Genomic regions with high F_{ST} can be incorrectly interpreted as having high absolute divergence (without acknowledging that low diversity will give similar signals)	• Apply a suite of summary statistics • Include recombination rate data • Inference of demographic processes including gene flow
Inference of processes underlying genomic differentiation	*Ad hoc* ideas about how the genes located within outlier regions have biologically plausible links to speciation can lead to over-interpretation of the functional importance of these regions	Just as for any inference of selection using molecular data, a null model of what pattern is expected under a neutral scenario has to be formulated. In this case, neutral null models have to include, for example, variation in genomic differentiation due to linked selection

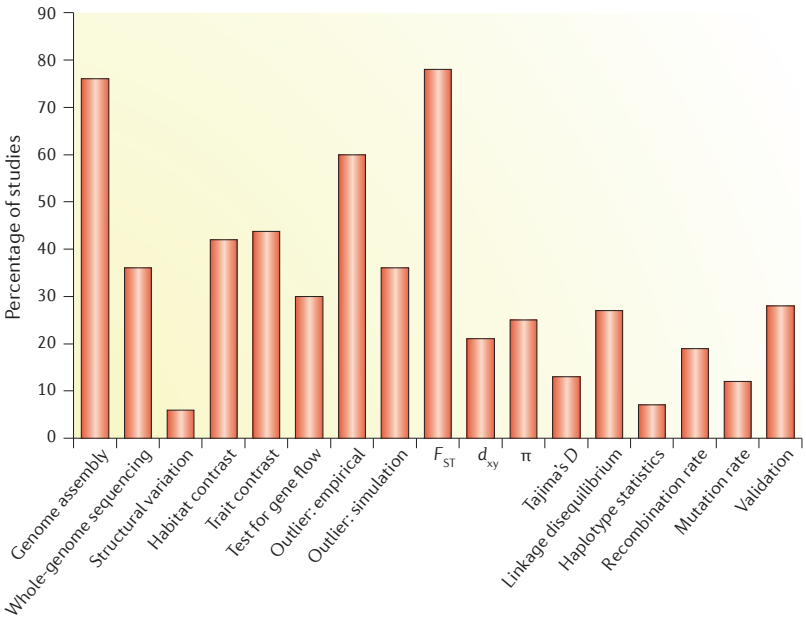

Figure 2 | **A summary of central aspects of our literature survey on speciation genomic studies.** The figure shows the percentage of studies in our data set that included central aspects of speciation genomic analyses.

availability of methods to infer the presence and/or amount of gene flow from multi-locus data[67], more effort needs to be devoted to this aspect. Given the difficulty of separating migration from incomplete lineage sorting[19], it is strongly advised that information from several sources is combined; these sources include field evidence of naturally occurring backcrosses, ABBA-BABA tests, admixture analyses and model-based inference on isolation-with-migration models[9,34,35,51].

Gene flow is not the only component of population history to consider when interpreting genome-wide patterns of genetic variation. Theory predicts that demographic perturbations such as population bottlenecks increase the variance in the time to the most recent common ancestor and thus have the potential to introduce significant heterogeneity in genetic differentiation across the genome[68,69]. Another factor to consider is the systematic effects for hemizygous sex chromosome sequences; compared with the diversity of autosome sequences, sex chromosome sequence diversity is relatively more affected upon population reduction and growth[70]. In this context it is worth noting that even with a constant population size, the lower N_e of the X (or the Z) chromosome will inevitably lead to faster lineage sorting and hence higher F_{ST} of X- or Z-linked sequences than of autosomal sequences[32]. The higher rate of differentiation in sex chromosomes is even further accentuated in the case of the non-recombining region of the Y (or the W) chromosome. Under random mating and assuming neutrality, N_e of the Y (or the W) chromosome is only one-quarter that of autosomes, and in practice the difference is much more pronounced due to the effects of selection in the absence of recombination. From a targeted assembly of the W chromosome in flycatcher species, Smeds *et al.*[71] recently demonstrated nearly full lineage sorting of W-linked sequences (F_{ST} = 0.96 – 1.00) despite only moderate levels of autosomal differentiation (F_{ST} = 0.27 – 0.40). Moreover, going beyond demographic effects, sex chromosomes are known to be important in speciation. Examples are Haldane's rule (the observation that in hybrids the sex that is most often sterile or inviable is the heterogametic sex) and the large X effect (the disproportionate involvement of sex-linked genes in reproductive isolation)[72,73]. Therefore, autosomes and sex chromosomes need to be explicitly separated in genomic analyses.

In general, under a simplistic null model of demographic stability (which is often accompanied by the expectation of homogeneous N_e across the genome), regional genomic inference on selection between populations will at best be obscured, if not falsely suggested[74]. However, only a small minority of studies provide background information on the demographic history of the populations under investigation. Despite explicit models for simultaneously estimating demography, gene flow and selection still being in their infancy[74], the potential impact of demographic perturbation should at least be qualitatively assessed. Explorative methods extracting demographic information from zygosity disequilibrium[75] or the single-nucleotide polymorphism (SNP) frequency spectrum[76] can be readily used on whole-genome data and can provide insight into the putative impact of changes in population size and structure[9,10,49].

Generating and analysing data

Data generation

Reference genome. Genome scans can, in principle, be conducted using a set of (unlinked) genetic markers in the absence of positional information from a reference genome. This may be potentially useful in the context of preselected candidate genes for phenotypes with a presumably simple genetic architecture. However, the full potential of genome scans can only be exploited with reference to an assembled genome backbone, which takes the non-independence of markers into account. Assembly of gigabase-sized genomes still requires some effort, and the resulting draft genomes vary substantially in quality. Most of the speciation genomic studies in our data set (Supplementary information S1 (table)) had access to a draft assembly of the genome, and in several cases the draft genomes were specifically assembled for the purpose of studying speciation in the system being investigated[32,35,58,77,78]. This clearly demonstrates that speciation genomic studies no longer rely on resources based on data from genetic model organisms. The use of a high-quality genome assembly should thus be considered current state of the art. Using reference genomes of closely related species may be a viable option[33,47], although this requires consideration of mapping biases and disruption of general patterns through rearrangements.

Ideally, speciation genomic studies ought to be based on chromosome-level genome assemblies. In several studied systems scaffolds have at least partially been ordered and oriented into longer sequence blocks on

Lineage sorting
The process by which alleles segregating in the common ancestor converge to the overall phylogeny of diverging lineages.

Backcrosses
Crosses between hybrids and one of the parents or a genetically similar individual from the parental population.

ABBA-BABA tests
A type of statistical test for introgression.

Admixture analyses
Tests for introgression of alleles between hybridizing populations to establish individual ancestries of individuals and/or genomic regions.

Isolation-with-migration
The process of population divergence in the presence of gene flow.

Population bottlenecks
Sharp reductions in the size of a population.

Box 3 | **Summary statistics of genetic variation**

Given a high-quality reference genome and adequate sequencing depth, whole-genome re-sequencing data can, in principle, provide a complete account of genetic variation in a sample population. If the sample is sufficiently large, the distribution of allele frequencies (the allele-frequency spectrum (AFS)) across all loci within the sample allows inference on demographic processes and selective forces of the population as a whole. Historically, a number of summary statistics have been derived, each drawing from different aspects of the AFS or assaying the correlation of AFSs among populations (known as the joint AFS). In this Box, we describe the commonly used metrics that we deem to be most relevant. Note that in the vast majority of studies, conclusions are qualitatively derived from estimated summary statistics, rather than from direct inference of population genetic parameters and processes specified in model-based inference.

Inter-population statistics

F_{ST}. This is a standardized measure of allele frequency differences between populations[145,146] and is used most pervasively in the context of genome scans[108,147]. Under the (often reasonable) assumption that the rate of mutation is low relative to migration, F_{ST} provides a basic measure of genetic drift. For neutral genetic variation F_{ST} is directly proportional to the level of gene flow under conditions of migration-drift equilibrium. However, there are several complicating factors that warrant caution when interpreting levels of F_{ST} at face value, and these caveats are shared in part by F_{ST}-like statistics such as G_{ST}[148], G_{ST}' (REF. 149) or D[150]. F_{ST} is an upward-bounded measure, which limits its sensitivity in those regions of the genome with high or intermediate levels of differentiation. As an indicator of regional genomic effective population size (N_e), F_{ST} is sensitive to both selection and gene flow (see FIG. 1). Moreover, it is influenced by the way it is calculated or averaged across loci[147,151]. Importantly, as a relative measure of differentiation F_{ST} co-varies with the level of within-population diversity[102,152], such that low diversity will generally result in a signal of increased differentiation. Normalized versions of F_{ST} (such as ZF_{ST} (REF. 153) and F_{ST}' (REF. 49), which are z-transformed) have been proposed; such versions of F_{ST} express differentiation in standard deviations and hence allow the quantification of relative differences in peak amplitude among population comparisons.

Population branch statistics. The population branch statistic (PBS) is a variant of the F_{ST} statistic that quantifies population-specific change in allele frequency from the point of population split. It is considered to have strong power to detect recent selective events and is attractive as it pinpoints lineage-specific rather than pairwise signals of differentiation[154].

d_{xy}. This is an absolute measure of population divergence that captures the average number of nucleotide differences among populations. It is proportional to the mutation rate (μ), time since divergence (t) and ancestral levels of diversity ($d_{xy} = 2\mu t + 4N_e\mu$). It is not confounded by within-population polymorphism and is thus more sensitive to gene flow than to recent selection in the sampled populations (but is sensitive to selection in the ancestor[19]). It was originally formulated by Nei[155] and has also been referred to as π_{XY} (REF. 156) and π_B (REF. 102).

d_a. d_a is also found as D_a or D_m in the literature, and was originally defined as δ[156]. It reflects the net divergence between populations since their split, assuming neutrality and no gene flow ($d_a = 2\mu t$). In contrast to d_{xy}, it removes the component of ancestral variation before the split. In practice, ancestral variation is generally approximated by averaging across the diversity of contemporary populations, which renders d_a dependent on intra-population diversity levels.

d_f. d_f is a less commonly used measure that quantifies the number of nucleotide differences that are fixed between populations, and it is standardized per base pair of sequence unit under consideration (known as the 'density of fixed differences' (REF. 32)). Despite its intuitive appeal, the number of fixed differences between populations is not merely governed by mutation rate and time since divergence, but is also influenced by coalescence times within populations[157]. Confident inference on fixation of alleles also requires a sufficiently large sample of chromosomes.

Intra-population statistics

θ_π and θ_S. Genetic diversity (θ) within a population is a fundamental statistic that is derived from the AFS. It can either be calculated as the mean difference between all possible pairs of nucleotide sequences present in a population (θ_π; generally referred to as nucleotide diversity (π)[155]), or by counting the number of segregating variants normalized by sample size (Watterson's estimator (θ_S))[158]. Under conditions specified by the neutral theory, these measures are expected to be of equal magnitude, influenced only by the mutation rate and the N_e (that is, the population mutation rate $\theta \sim N_e\mu$).

Tajima's D. Whereas θ_π captures variation of medium frequency alleles, θ_S is sensitive to rare alleles. The Tajima's D metric essentially compares the difference between the two ($\theta_\pi - \theta_S$) standardized by its standard deviation[159], with a negative value indicating an excess of rare variants. As the proportion of rare alleles is influenced both by demographic processes and by selection, their respective contribution to the metric is difficult to distinguish. In addition, background selection and positive selection qualitatively influence the measure in a similar way. Nevertheless, Tajima's D is a central test statistic in examining deviation from neutrality.

Fay and Wu's H. This is a summary statistic that measures departures from neutrality that are reflected in the difference between intermediate-frequency and high-frequency variants ($\theta_\pi - \theta_H$). Thus, Fay and Wu's H is less sensitive to rapid changes in population size than Tajima's D, which capitalizes on rare-frequency variants. Comparison of the two metrics may help to distinguish demographic processes from selection.

Fu and Li's D. This statistic extends the above-mentioned metrics by polarizing changes (derived or ancestral allele) with respect to an outgroup[160]. This makes Fu and Li's D suitable when investigating the importance of derived mutations in population-specific (hard) sweeps.

Linkage disequilibrium. This is a basic measure of allelic association between loci. Being influenced by many processes — including recombination, selection, admixture and their complex inter-relationships — it can be difficult to interpret. The extent of linkage disequilibrium in a population, and in specific regions of the genome, is crucial to any type of genome scan (that is, not only in the context of measuring diversity within or divergence between populations), as it directly affects how far along the chromosome signals (from selection, for example) will be detectable. Methodological approaches making full use of genome-wide linkage disequilibrium signatures in the context of population divergence are gradually emerging[161].

EHH, iHH, EHHS, iES, XP-EHH and Rsb. Haplotype statistics reflecting the decay of linkage disequilibrium[162,163] along the genome are useful measures that contain information about selection. Extended haplotype homozygosity (EHH) is the probability that two chromosomes are identical by descent for a given interval between allelic variants of two core single-nucleotide polymorphisms (SNPs). It detects the transmission of extended haplotypes decaying monotonically to zero with increasing distance from the focal SNPs. The integrated EHH (iHH) quantifies the area under this curve against map position, thus providing a measure of 'haplotype length' on a SNP-by-SNP basis. The EHHS and iES statistics are analogous to the EHH and iHH statistics, respectively, and provide a weighted average of both alleles of a focal SNP. Contrasting haplotype length of ancestral (iHHa) and derived (iHHd) alleles, the iHS statistic provides a measure of recent positive selection events. Statistics such as XP-EHH and Rsb allow for comparisons of haplotype length between populations.

Coalescent hidden Markov models. Coalescent hidden Markov models (HMMs)[164] examine the dependence of genealogies between neighbouring nucleotides as a function of coalescence and recombination. The explicit modelling of the history of recombination holds the potential to quantify ancestry along the genome and thereby detect differential introgression. Similar methods exploiting HMMs to account for local non-independence of markers are emerging[165].

the basis of genetic maps (for example, sticklebacks, flycatchers, sunflowers, *Heliconius* butterflies and stick insects). When genetic maps are beyond reach, long contiguous sequences can still serve as a valuable backbone, and new sequencing technologies providing long reads will help to obtain such sequences.

Whole-genome versus reduced-representation re-sequencing. The terms 'genome-wide' or 'genome-scale' do not necessarily imply that every base pair of each individual is analysed. For one thing, very few eukaryotic genomes have been fully sequenced and assembled to date. For example, highly repetitive and heterochromatic regions are extremely difficult to decode, and in a strict sense the term 'genome-wide' is inevitably limited to 'genome-assembly-wide'. Of more practical relevance, the majority of the studies in our data set based their conclusions on a (sometimes only small) subset of the genome in the form of transcriptomes, preselected SNP arrays, or genotyping-by-sequencing methods with several orders of magnitude fewer SNPs than are genotyped in whole-genome re-sequencing efforts.

Although they are generally more cost-effective, approaches based on subsets of the genome can be problematic and, strictly speaking, may not warrant the label genome-wide. For example, reduced-representation data preclude the use of many informative summary statistics and can risk not capturing the distribution of genetic variation at sufficient resolution. Localized signatures may be either missed or incorporated into seemingly large regions of interest. Information on linkage disequilibrium between markers is crucial in this context, although it is often not explicitly addressed. Several studies discuss candidate genes in the vicinity of outlier regions suggested by one of a few genetic markers without formally establishing the degree of linkage between the marker and the candidate gene. On the basis of these considerations, we advocate the use of genome-wide approaches that are based on assembled genome sequences and polymorphism data obtained from whole-genome re-sequencing. In fact, speciation genomic research should ultimately aim to characterize genetic variation on the basis of individual assemblies for each sampled specimen to make full use of the variation that is not accessible when using a single reference genome[79].

The use of pools versus individuals for sequencing. Although most studies in our data set sequenced individually barcoded specimens, ten chose to sequence pools of individuals. Pooled sequencing (also known as Pool-seq) is a cost-effective alternative that, in principle, allows for population allele frequencies to be inferred directly from read counts[80]. However, deviations from equimolar DNA contributions, sampling variance (even for large samples), PCR biases and impeded sequencing error detection can lead to bias in the estimation of summary statistics. Moreover, pooled sequencing precludes the use of haplotype-based summary statistics that can leverage important information about admixture and selection (BOX 3). Simply put, sequencing individuals (at sufficient depth) increases the accuracy and power of the data, and is increasingly becoming possible owing to the decreasing cost of sequencing.

Data analysis

Source of genetic variation. SNPs are still by far the prevailing category of mutations considered in the studies forming our data set, and only a small proportion of studies (16%) explicitly included information on structural genetic variation (Supplementary information S1 (table)). The focus on SNPs has some limitations because structural genomic variation — including chromosomal rearrangements, insertions, deletions and duplications — is known to have an impact on phenotypic diversity and may contribute to reproductive isolation.

Dissection of the genetic basis of several distinct traits of domestic animals provides illustrative examples of how complex structural variation can have important phenotypic effects[81]. Likewise, chromosomal rearrangements have been experimentally shown to affect gene expression and fitness[82]. Several studies of natural populations of non-model species also report that segregating phenotypes can be explained both by chromosomal inversions[83–86] and by recurrent deletions[57] or insertions of transposable elements[87]. When structural variation underlies variation in traits that are under divergent selection among populations, it may contribute to reproductive isolation in ways that are not fundamentally different from SNPs in coding or regulatory sequences. Structural variation in the form of chromosomal rearrangements has also been the focus of much research in speciation (the field of 'chromosomal speciation') that more generally explores effects on fitness ('underdominance') and the effects of linked selection of allelic variants across many genes[4,88]. Theoretical models often focus on inversions and differ in their views on the contribution that they make to inter-population barriers against gene flow[88–91], and empirical evidence is also mixed. Several studies provide evidence (although often indirect) that is consistent with a role of inversions in population divergence[64,92,93]. Others find no effect of inversions on genetic differentiation under controlled conditions[94], and genetic variation present in inverted genomic regions can likewise diverge over long timescales and still segregate without promoting reproductive isolation, which is reminiscent of sex chromosomes[95]. The question of how much structural variation contributes to speciation requires more quantitative empirical input, and model-based approaches predicting genome-wide patterns caused by structural genomic variation will also be needed[96,97].

Sequencing data from long-read technology or methods such as optical mapping contain relevant information on structural variation, including duplications, deletions and copy-number variation. It may be necessary to make separate *de novo* assemblies of the lineages under comparison because structural variation can easily be missed if reads from different lineages are mapped to a single reference genome[79]. This rather novel type of data may not only add the dimension of structural variation

Reduced-representation data
Genetic data from a defined subset of the genome.

Linkage disequilibrium
The non-random, statistical association between alleles at different loci.

Underdominance
Fitness reduction of heterozygous genotypes at a bi-allelic locus.

Long-read technology
Sequencing technologies that generate relatively long stretches of DNA sequence per read. The recent development of long-single-molecule sequencing (> 20 kb) blurs the initial dichotomy of short reads (from sequencing-by-synthesis technology) versus long reads (~ 1 kb Sanger reads).

Optical mapping
A technique for constructing high-resolution restriction maps from single molecules.

to genome scans but may also affect population genetic inference by allowing the assembly of genomic regions that would otherwise remain unanalysed[98].

Summary statistics. Traditionally, in genome scans F_{ST} is the measure of choice for quantifying the genetic distance between populations along the genome (BOX 3). Although certainly useful, it is important to keep in mind that F_{ST} is a relative measure of differentiation that is dependent on the underlying intra-population genetic diversity. If differentiation is uniform across the genome but diversity levels vary, differentiation as estimated by F_{ST} would appear to be heterogeneous. Absolute distance measures such as d_{xy} are, in principle, better suited to differentiate between evolutionary scenarios that entail locally elevated differentiation (reviewed in REF. 19). It is thus surprising that only 21% of studies in our data set considered d_{xy} at all. Similarly, population branch statistics (PBSs) that allow inference on population-specific selection versus selection already acting in the ancestor are only rarely used outside the context of human population genetics.

Most studies in our data set drew conclusions from a narrow set of summary statistics generally including F_{ST} (78%) and nucleotide diversity (θ_π; 51%), which by themselves are unsuitable to identify the underlying evolutionary processes. Both F_{ST} and θ_π are also sensitive to regional variation in mutation rate (μ), proxies of which have only explicitly been addressed in 12% of the studies. Summary statistics that probe different aspects of the allele frequency spectrum — such as Tajima's *D*, Fay and Wu's *H*, and Fu and Li's *D* — that could provide more insight into population-specific selection were hardly used. Linkage disequilibrium was estimated in 27% of the studies, but the use of powerful haplotype statistics (BOX 3) was the exception rather than the rule (only being used in 7% of studies). Outgroup species polarizing ancestral and derived variants — and hence enabling more powerful summary statistics regarding positive selection events to be obtained — were generally absent.

A parameter that deserves particular attention is the recombination rate. It has a central role in theoretical work on speciation[99] but also directly influences the extent of linked selection[18]. The traditional way of generating recombination rate data has been via linkage analysis based on marker genotyping in pedigrees. For use in population genomic analyses, this almost necessitates that a genome assembly is available such that recombination fractions can be translated into rates of recombination (r) per physical unit DNA (often centimorgans per megabase). In cases for which recombination data from external sources are not available, the population recombination rate $\delta = 4N_e r$ might be used as a proxy on the basis of analyses of linkage disequilibrium[13,100]. However, although δ is generally well correlated with r, it is sensitive to selection reducing the contribution of N_e. A possible alternative is to find proxies for N_e to solve r from $\delta = 4N_e r$ (REFS 32,50). Nevertheless, this approach is not without its problems. As r mediates the effects of selection on N_e, estimating r through N_e is somewhat circular.

The effects of linked selection will be particularly pronounced in low-recombining regions, where it will leave signatures that can be approximated by reduced N_e (REFS 18,101), and will cause accelerated lineage sorting and hence elevated rates of genetic differentiation[102], irrespective of the amount of gene flow among populations. Without knowledge of the landscape of recombination rate variation, regions of elevated differentiation can therefore be mistakenly interpreted as signals of divergent selection or reduced gene flow. Only a subset of studies in our data set (19%) had access to external recombination data. For instance, Burri *et al.*[47] concluded that heterogeneous differentiation landscapes are better explained by linked selection in regions of low recombination and high gene density rather than by divergent selection against gene flow. Consistent with this interpretation, Roesti *et al.*[103] also found reduced nucleotide diversity (and hence increased F_{ST}) in areas of low recombination. In general, linked selection seems to contribute substantially to variation in nucleotide diversity (but does not fully explain it[104]). In a meta-analysis, Corbett-Detig *et al.*[16] suggested that nucleotide diversity is affected by the local recombination environment across a large taxonomic range of species. Using model-based approaches to assess the general processes acting genome-wide before considering outlier peaks[105] promises to be a fruitful way forward, as high-resolution recombination data become available for an increasing number of taxa.

Outlier tests. Genome-wide scans rely on the assumption that for most of the genome genetic diversity reflects neutrality. Anomalous patterns of genetic diversity — usually referred to as 'outliers' — are then thought to indicate selection. However, as mentioned above, factors such as mutation, demographic perturbation or recombination rate variation, in conjunction with linked selection, may compromise the validity of this approach[20]. Importantly, outlier tests are only sensitive to large-effect loci and thus have an inherent ascertainment bias in detecting traits under putative selection that have a simple genetic architecture (BOX 2). It is often forgotten that genome scans are effectively blind to selection on small-effect polygenes or when epistasis is involved[106,107].

Accepting the general logic of genome scans, outliers can be inferred empirically on the basis of extreme percentiles of the distribution (with an arbitrary threshold) for a given summary statistic, usually F_{ST}. Alternatively, outliers can be inferred by contrasting the observed distribution of the summary statistic to an expected distribution generally derived by simulation under a certain demographic scenario[108]. In the studies forming our data set, 54% took an exclusively 'empirical approach' and 36% used model-based inference (mostly following the methods in REFS 109,110), with the remainder not addressing outliers. Model-based approaches have the clear advantage of incorporating heterogeneity in the distribution of genetic variation caused by demographic history. However, they are often computationally intense, which was reflected by the fact that model-based approaches were almost restricted to studies using data

from reduced-representation libraries. The empirical approach might result in more false-positive outlier regions, but F_{ST} has been shown to be rather robust to demographic effects[108].

Even when explicit model-based approaches are amenable, multi-way comparisons across populations of different connectivity are a promising way forward to separate processes acting similarly on all populations from processes that are specific to a single population or population comparison (see above). The few studies that have actively pursued this idea strongly suggest that linked selection, that is, not related to population-specific selection against gene flow, is pervasive across the genome[33,47,111]. In the future we clearly need explicit null models and simulation tools that take common heterogeneity of differentiation into account to separate population-specific effects[74,112]. As a simple, empirical way forward it has been suggested that standardized F_{ST} values in orthologous genomic regions of control comparisons (for example, allopatric or no contrast speciation settings) are essentially subtracted from the focal comparisons, yielding a measure of net differentiation (such as $\Delta F_{ST}'$ (REF. 49) or Δ divergence[113,114]). Regions classified for F_{ST}, but not for the Δ statistic, can then be interpreted as genomic regions that are subject to shared selection pressures, whereas regions classified as outliers for the Δ statistic are potentially affected by selection pressures that are specific to the target comparisons. Similarly, Roesti *et al.*[111] suggested the use of 'residual F_{ST}', which controls for systematic variation in F_{ST} as a function of distance from the centromere. Formal development of *ad hoc* ΔF_{ST} statistics would constitute an important step forward.

Overall, despite obvious limitations, F_{ST}-based genome scans are a useful exploratory tool but need to be complemented with additional information from other summary statistics (see above). A posteriori searches for candidate genes in outlier regions[20] can narrow in on regions of interest, but ultimately functional validation is necessary to support conclusions on selection that are inferred from genetic diversity data.

Functional validation

Associations between a priori ecological or phenotypic contrasts and a genetic signal do not prove causality, even if the contrast has been demonstrated to differentially affect fitness between diverging populations. In the context of speciation studies, genomic regions with elevated differentiation could possibly harbour loci that are involved in reproductive isolation, but, as described above, such signals can also be due to processes that are not related to speciation. Moreover, highly differentiated regions usually contain several annotated genes plus an unknown amount of sequence of unknown function. Searching for candidates within such regions can therefore only indicate, not demonstrate, causality. Functional validation is thus a central component of genomic studies of adaptation or speciation in natural populations.

One relatively straightforward, though not definitive, way to functionally validate a genetic signal is to carry out gene expression analyses. For example, differential gene expression between grey-coated hooded crows and all-black carrion crows has been shown to largely be confined to genes of the melanogenesis pathway, and genes that are crucial for the regulation of these pigmentation genes have been linked to major differentiation peaks of the two crow taxa[35,115]. In this case, there is a biologically plausible link between gene expression and genomic differentiation, yet this does not fully establish causality. In other cases in which a more complex genetic architecture is expected for a trait under selection the link will be even less obvious, and caution is warranted in immediately interpreting the differential expression of genes located within highly differentiated regions as evidence for a role in speciation. As for any type of variation, standing *cis*-regulatory variation for gene expression in the ancestral species is more likely to sort by drift between diverging lineages in regions of low N_e than elsewhere in the genome. This makes differential gene expression that is not related to the ecological or phenotypic contrast more likely in such regions than in other regions of the genome (everything else being equal).

Another way forward is validation using information from QTL mapping from genetic crosses or natural pedigrees[116], or from admixture mapping using genome-wide data from naturally hybridizing populations[117,118]. Colocalization of QTLs of relevant characters and differentiation islands clearly strengthens the interpretation of islands representing regions under divergent selection in association with population divergence[22]. This combined approach is currently the most commonly used method of validation: examples include pea aphids, sticklebacks, *Heliconius* butterflies, whitefish and *Drosophila* spp.

In molecular studies of model organisms, the traditional way of demonstrating that a particular mutation or gene underlies a phenotype has been to use transgenic technologies (for example, knockouts) and RNA interference (RNAi). The introduction of the CRISPR–Cas9 system as a versatile tool for genome editing enables much simplified possibilities for such functional studies in both model and non-model organisms[119]. Despite still being in its infancy in terms of applications in evolutionary and ecological research[120,121], there should be tremendous potential in the use of CRISPR–Cas9 for revealing whether mutations in candidate genes lead to reproductive incompatibility. Again, however, caution is necessary here, as functional genomic validation is only likely to be successful for traits with simple genetic architectures (such as those with an oligogenic background and little G × E interaction).

Conclusions and future directions

Summarizing the conclusions from the 67 studies forming the core data set for this Review (Supplementary information S1 (table)) is challenging. The studies used different sets of methodological approaches (FIG. 2), genomes and life histories differ among taxa in important ways, and the studies differ in focus. The seemingly simple task of comparing the relative number and extent of 'differentiation peaks' is in fact nearly impossible.

Different window sizes are used, and only a few studies are designed so as to enable meaningful comparisons in genetic map units rather than physical distance. Additionally, outliers are defined either as tails of distributions or by means of different, necessarily simplistic demographic models. Some studies find few discrete regions of genetic differentiation, whereas others find pervasive peaks across the genome; in both settings the findings are often related to subjective a priori contrasts.

Box 4 | **Relevant aspects to consider in speciation genomic studies**

Demographic history
As a basis for all further investigation, the demographic history of populations needs to be reconstructed with due care, in particular the level and direction of gene flow between target populations need to be determined.

Alternative processes
Alternative processes to divergent selection against gene flow need to be considered. They may have no bearing on speciation but still generate similar patterns of locally elevated differentiation. The role of linked selection even in the absence of gene flow is crucial in this context.

Theoretical null models
On the theoretical side, null models need to be refined so that they more completely describe how genetic variation is expected to segregate across genomes that are under the influence of (linked) selection to provide a basis for separating between alternatives.

Genetic maps
Recombination is a central factor in determining which traces selection will leave in patterns of genetic variation in the genome. Genome scans should preferably be made with reference to genetic maps rather than physical maps. Moreover, comparisons of genetic and physical maps will uncover regions of reduced recombination rate.

Summary statistics
There is a need to move away from the F_{ST}-centric view and adopt a multivariate perspective that integrates several different (often correlated) summary statistics (BOX 3). Genome scans do not necessarily have to start with F_{ST}. Ideally, the common practice of hierarchical, sequential use of different summary statistics can be abandoned for the sake of integrated, model-based analyses that directly estimate the parameter of interest.

Type of genetic variation
Increased attention should be paid to genetic variation other than single-nucleotide variation, in particular structural genomic variation. Technology is available for detection of structural variation, but new models and summary statistics might be needed.

Divergence continuum
Even if genome scans may identify regions that are good candidates for driving population differentiation, their relevance to reproductive isolation is often unclear. Replicate sampling of populations is important to assess the generality of identified candidates in different genotypic backgrounds. It can also help to judge whether differentiation builds up around these loci at increasing levels of population divergence. However, the contribution of shared ancestry between population comparisons compromising independence needs to be taken into account.

Outlier plausibility
Theory tells us that genome scans are most sensitive to large-effect loci and blind to polygenes. The functional interpretation of genome scans needs to accommodate this concept.

Functional validation
As for all types of association study, functional validation should eventually be strived for, when possible. CRISPR–Cas9 genome editing is a promising tool for use in molecular ecology.

Meta-scale analyses
In addition to well-established ecological models, in which detailed and replicate inference is possible, we need data from a broader variety of taxa, spanning wide ranges of N_e, recombination and mutation rate. Comparative analyses will prove useful to extract commonalities.

Despite apparent heterogeneity in the types of system, the approaches and the findings, most studies in our data set suggest divergent selection against gene flow as the main process generating peaks of elevated genetic differentiation, mostly within the framework of isolation-by-ecology. However, in line with Cruickshank and Hahn[19], we point out that caution is warranted when generalizing this interpretation. The field is currently dominated by a few model species, by a community with an often adaptationist perspective (depending on the organism studied, this mindset is possibly justified), and not least by the powerful metaphor of 'speciation islands' that has been successfully spread in the context of ecological speciation. However, alternative explanations — such as background selection, recurrent, parallel meiotic or centromeric drive, or increased variance due to demographic perturbations — are rarely considered as coequal processes that contribute to heterogeneity in genetic differentiation. Moreover, a multitude of additional factors complicate the expected patterns of genetic variation and differentiation along the genome, making it difficult to infer the underlying evolutionary process from patterns alone. These include mutational variation, biogeographic population history, temporal fluctuation in gene flow, strength and timing of selection, genetic architecture of traits under selection (such as dominance, pleiotropy and epistasis) and their interaction[31,122]. The existence of these additional factors leads researchers to interpret data in line with their own preconceptions rather than formally testing alternative hypotheses of equal value. Re-analysis and re-interpretation of published material in the spirit of community exercises such as the Assemblathon[123] would be highly welcome and are expected to invigorate the field. Moreover, marked peaks of differentiation are only expected for traits under strong selection that have a simple genetic architecture. For divergent selection on quantitative traits with a polygenic architecture and for which epistatic interactions may have a role[61], marked F_{ST} peaks are not expected[106]. Together, this indicates that we are not yet in a position to be able to generalize broad-scale patterns of how genomic divergence relates to the speciation process across groups of organisms. In an attempt to provide recommendations for how the field should move forward we have summarized ten central aspects that we suggest should be considered in future speciation genomic work (BOX 4).

A decade ago there was much enthusiasm about finding 'speciation genes' in genomic regions of elevated differentiation. As the field has matured, it has become apparent that expectations may need to be tempered. This does not mean that analyses of genetic variation segregating in natural populations do not contain useful information about the micro-evolutionary processes contributing to population divergence, and eventually promoting speciation. Work published in recent years has documented an intriguing, previously unrecognized heterogeneity in genomic differentiation during lineage divergence, and this heterogeneity requires explanation. To advance our knowledge of how these patterns relate to reproductive barriers encoded in the genome, future studies will need to be more comprehensive in several aspects.

1. Darwin, C. & Wallace, A. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J. Proc. Linn. Soc. Zool.* **3**, 45–62 (1858).
2. Dobzhansky, T. G. *Genetics and the Origin of Species* (Columbia Univ. Press, 1937).
3. Mayr, E. & Provine, W. B. *The Evolutionary Synthesis: Perspectives on the Unification of Biology* (Harvard Univ. Press, 1998).
4. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, 2004).
5. Presgraves, D. C. The molecular evolutionary basis of species formation. *Nat. Rev. Genet.* **11**, 175–180 (2010).
6. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* genetic reference panel. *Nature* **482**, 173–178 (2012).
7. Wolf, J. B. W., Lindell, J. & Backstrom, N. Speciation genetics: current status and evolving approaches. *Phil. Trans. R. Soc. B Biol. Sci.* **365**, 1717–1733 (2010).
8. Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192 (2014).
This Review resulted from a workshop and discusses genomic approaches in speciation at an advanced level.
9. Foote, A. *et al.* Genome-culture coevolution promotes rapid divergence in the killer whale. *Nat. Commun.* **7**, 11693 (2016).
10. Nadachowska-Brzyska, K., Burri, R., Smeds, L. & Ellegren, H. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol. Ecol.* **25**, 1058–1072 (2016).
11. Lawrie, D. S. & Petrov, D. A. Comparative population genomics: power and principles for the inference of functionality. *Trends Genet.* **30**, 133–139 (2014).
12. Comeron, J. M., Ratnappan, R. & Bailin, S. The many landscapes of recombination in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1002905 (2012).
13. Singhal, S. *et al.* Stable recombination hotspots in birds. *Science* **350**, 928–932 (2015).
14. Mugal, C. F., Weber, C. C. & Ellegren, H. GC-biased gene conversion links the recombination landscape and demography to genomic base composition. *BioEssays* **37**, 1317–1326 (2015).
15. Romiguier, J. *et al.* Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* **515**, 261–263 (2014).
16. Corbett-Detig, R. B., Hartl, D. L. & Sackton, T. B. Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.* **13**, e1002112 (2015).
17. Noor, M. A. F. & Bennett, S. M. Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* **103**, 439–444 (2009).
18. Cutter, A. D. & Payseur, B. A. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**, 262–274 (2013).
19. Cruickshank, T. E. & Hahn, M. W. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157 (2014).
This paper provides a good introduction to the processes by which genetic differentiation can be locally elevated.
20. Haasl, R. J. & Payseur, B. A. Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.* **25**, 5–23 (2016).
21. Wu, C. I. The genic view of the process of speciation. *J. Evol. Biol.* **14**, 851–865 (2001).
This influential paper provides a conceptual link between (Darwinian) selection acting on single loci and Mayr's concept of cohesive, genome-wide reproductive isolation under the biological speciation concept.
22. Feder, J. L., Egan, S. P. & Nosil, P. The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012).
23. Nosil, P. & Feder, J. L. Genome evolution and speciation: toward quantitative descriptions of pattern and process. *Evolution* **67**, 2461–2467 (2013).
24. Barton, N. & Bengtsson, B. O. The barrier to genetic exchange between hybridising populations. *Heredity* **57**, 357–376 (1986).
25. McDermott, S. R. & Noor, M. A. F. The role of meiotic drive in hybrid male sterility. *Phil. Trans. R. Soc. B* **365**, 1265–1272 (2010).
26. Zanders, S. E. *et al.* Genome rearrangements and pervasive meiotic drive cause hybrid infertility in fission yeast. *eLife* **3**, e02630 (2014).
27. Harr, B. Genomic islands of differentiation between house mouse subspecies. *Genome Res.* **16**, 730–737 (2006).
28. Turner, T. L., Hahn, M. W. & Nuzhdin, S. V. Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**, 1572–1578 (2005).
This influential paper was the first to interpret islands of differentiation as 'speciation islands'.
29. Pennisi, E. Disputed islands. *Science* **345**, 611–613 (2014).
This editorial piece provides a historical perspective on the interpretation of genomic regions with elevated differentiation and includes illustrative examples.
30. Yeaman, S. Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proc. Natl Acad. Sci. USA* **110**, E1743–E1751 (2013).
31. Feder, J. L., Flaxman, S. M., Egan, S. P., Comeault, A. A. & Nosil, P. Geographic mode of speciation and genomic divergence. *Annu. Rev. Ecol. Evol. Syst.* **44**, 73–97 (2013).
32. Ellegren, H. *et al.* The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760 (2012).
This is one of the first genome-wide re-sequencing studies to demonstrate marked heterogeneity in the level of differentiation with few clear peaks per chromosome.
33. Renaut, S. *et al.* Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nat. Commun.* **4**, 1827 (2013).
This study provides an important empirical demonstration that genomic islands of elevated differentiation emerge between populations in a similar way across a variety of geographical contexts that differ in the presumed amount of gene flow.
34. Martin, S. H. *et al.* Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* **23**, 1817–1828 (2013).
This study quantifies the level of gene flow during species divergence.
35. Poelstra, J. W. *et al.* The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**, 1410–1414 (2014).
This empirical study provides evidence for highly localized genomic selection against introgression and includes functional analyses.
36. Soria-Carrasco, V. *et al.* Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**, 738–742 (2014).
37. Marques, D. A. *et al.* Genomics of rapid incipient speciation in sympatric threespine stickleback. *PLoS Genet.* **12**, e1005887 (2016).
38. Via, S. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Phil. Trans. R. Soc. B* **367**, 451–460 (2012).
39. Nachman, M. W. & Payseur, B. A. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Phil. Trans. R. Soc. B* **367**, 409–421 (2012).
This empirical study has a solid conceptual introduction and highlights the importance of linked selection in genomic regions of low recombination.
40. Smith, J. M. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35 (1974).
41. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**, 909–919 (2000).
42. Charlesworth, B., Morgan, M. T. & Charlesworth, D. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303 (1993).
43. Charlesworth, B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. *J. Hered.* **104**, 161–171 (2013).
44. Stukenbrock, E. H. in *Advances in Botanical Research* (ed. Martin, F.) **70**, 397–423 (Academic Press, 2014).
45. Dettman, J. R., Sirjusingh, C., Kohn, L. M. & Anderson, J. B. Incipient speciation by divergent adaptation and antagonistic epistasis in yeast. *Nature* **447**, 585–588 (2007).
46. Shaw, K. L. & Mullen, S. P. Speciation continuum. *J. Hered.* **105**, 741–742 (2014).
47. Burri, R. *et al.* Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* **25**, 1656–1665 (2015).
48. Andrew, R. L. & Rieseberg, L. H. Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution* **67**, 2468–2482 (2013).
49. Vijay, N. *et al.* Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nat. Commun.* (in the press).
50. Feulner, P. G. D. *et al.* Genomics of divergence along a continuum of parapatric population differentiation. *PLoS Genet.* **11**, e1004966 (2015).
51. Malinsky, M. *et al.* Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350**, 1493–1498 (2015).
52. Via, S. & West, J. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Mol. Ecol.* **17**, 4334–4345 (2008).
53. Via, S. Natural selection in action during speciation. *Proc. Natl Acad. Sci. USA* **106**, 9939–9946 (2009).
54. Nadeau, N. J. *et al.* Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol. Ecol.* **22**, 814–826 (2013).
This empirical study demonstrates the power of study design in the interpretation of outlier genomic regions.
55. Kronforst, M. R. *et al.* Hybridization reveals the evolving genomic architecture of speciation. *Cell Rep.* **5**, 666–677 (2013).
56. Nadeau, N. J. *et al.* Population genomics of parallel hybrid zones in the mimetic butterflies. *H. melpomene* and *H. erato*. *Genome Res.* **24**, 1316–1333 (2014).
57. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a *Pitx1* enhancer. *Science* **327**, 302–305 (2009).
58. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
59. Roesti, M., Kueng, B., Moser, D. & Berner, D. The genomics of ecological vicariance in threespine stickleback fish. *Nat. Commun.* **6**, 8767 (2015).
60. Savolainen, O., Lascoux, M. & Merilä, J. Ecological genomics of local adaptation. *Nat. Rev. Genet.* **14**, 807–820 (2013).
61. Mossman, J. A., Biancani, L. M. & Rand, D. M. Mitonuclear epistasis for development time and its modification by diet in *Drosophila*. *Genetics* **203**, 463–484 (2016).
62. Slatkin, M. Inbreeding coefficients and coalescence times. *Genet. Res.* **58**, 167–175 (1991).
63. Kulathinal, R. J., Stevison, L. S. & Noor, M. A. F. The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. *PLoS Genet.* **5**, e1000550 (2009).
64. McGaugh, S. E. & Noor, M. A. F. Genomic impacts of chromosomal inversions in parapatric *Drosophila* species. *Phil. Trans. R. Soc. B* **367**, 422–429 (2012).
65. Shafer, A. B. A. & Wolf, J. B. W. Widespread evidence for incipient ecological speciation: a meta-analysis of isolation-by-ecology. *Ecol. Lett.* **16**, 940–950 (2013).
66. Shafer, A. B. A., Northrup, J. M., Wikelski, M., Wittemyer, G. & Wolf, J. B. W. Forecasting ecological genomics: high-tech animal instrumentation meets high-throughput sequencing. *PLoS Biol.* **14**, e1002350 (2016).
67. Payseur, B. A. & Rieseberg, L. H. A genomic perspective on hybridization and speciation. *Mol. Ecol.* **25**, 2337–2360 (2016).
68. Hein, J., Schierup, M. H. & Wiuf, C. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory* (Oxford Univ. Press, 2005).
69. Gattepaille, L. M., Jakobsson, M. & Blum, M. G. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity* **110**, 409–419 (2013).
70. Pool, J. E. & Nielsen, R. Population size changes reshape genomic patterns of diversity. *Evolution* **61**, 3001–3006 (2007).
71. Smeds, L. *et al.* Evolutionary analysis of the female-specific avian W chromosome. *Nat. Commun.* **6**, 7330 (2015).
72. Presgraves, D. C. Sex chromosomes and speciation in *Drosophila*. *Trends Genet.* **24**, 336–343 (2008).
73. Qvarnström, A. & Bailey, R. I. Speciation through evolution of sex-linked genes. *Heredity* **102**, 4–15 (2009).

74. Bank, C., Ewing, G. B., Ferrer-Admettla, A., Foll, M. & Jensen, J. D. Thinking too positive? Revisiting current methods of population genetic selection inference. *Trends Genet.* **30**, 540–546 (2014).
75. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
76. Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559 (2015).
77. The Heliconius Genome Consortium. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* **487**, 94–98 (2012).
78. Gompert, Z. *et al.* Experimental evidence for ecological selection on genome variation in the wild. *Ecol. Lett.* **17**, 369–379 (2014).
79. Chaisson, M. J. P., Wilson, R. K. & Eichler, E. E. Genetic variation and the *de novo* assembly of human genomes. *Nat. Rev. Genet.* **16**, 627–640 (2015).
80. Schlötterer, C., Tobler, R., Kofler, R. & Nolte, V. Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. *Nat. Rev. Genet.* **15**, 749–763 (2014).
81. Bed'hom, B. *et al.* The *lavender* plumage colour in Japanese quail is associated with a complex mutation in the region of *MLPH* that is related to differences in growth, feed consumption and body temperature. *BMC Genomics* **13**, 442 (2012).
82. Avelar, A. T., Perfeito, L., Gordo, I. & Ferreira, M. G. Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. *Nat. Commun.* **4**, 2235 (2013).
83. Schwander, T., Libbrecht, R. & Keller, L. Supergenes and complex phenotypes. *Curr. Biol.* **24**, R288–R294 (2014).
84. Küpper, C. *et al.* A supergene determines highly divergent male reproductive morphs in the ruff. *Nat. Genet.* **48**, 79–83 (2016).
85. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88 (2016).
86. Kirubakaran, T. G. *et al.* Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. *Mol. Ecol.* **25**, 2130–2143 (2016).
87. Saenko, S. V. *et al.* Amelanism in the corn snake is associated with the insertion of an LTR-retrotransposon in the *OCA2* gene. *Sci. Rep.* **5**, 17118 (2015).
88. Hoffmann, A. A. & Rieseberg, L. H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? *Annu. Rev. Ecol. Evol. Syst.* **39**, 21–42 (2008).
89. Rieseberg, L. H. Chromosomal rearrangements and speciation. *Trends Ecol. Evol.* **16**, 351–358 (2001).
90. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **173**, 419–434 (2006).
91. Faria, R. & Navarro, A. Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends Ecol. Evol.* **25**, 660–669 (2010).
92. Navarro, A. & Barton, N. H. Chromosomal speciation and molecular divergence-accelerated evolution in rearranged chromosomes. *Science* **300**, 321–324 (2003).
93. Lohse, K., Clarke, M., Ritchie, M. G. & Etges, W. J. Genome-wide tests for introgression between cactophilic *Drosophila* implicate a role of inversions during speciation. *Evolution* **69**, 1178–1190 (2015).
94. Huang, Y., Wright, S. I. & Agrawal, A. F. Genome-wide patterns of genetic variation within and among alternative selective regimes. *PLoS Genet.* **10**, e1004527 (2014).
95. Tuttle, E. M. *et al.* Divergence and functional degradation of a sex chromosome-like supergene. *Curr. Biol.* **26**, 344–350 (2016).
96. Guerrero, R. F., Rousset, F. & Kirkpatrick, M. Coalescent patterns for chromosomal inversions in divergent populations. *Phil. Trans. R. Soc. B Biol. Sci.* **367**, 430–438 (2012).
97. Feder, J. L., Nosil, P. & Flaxman, S. M. Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. *Front. Genet.* **5**, 295 (2014).
98. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
99. Felsenstein, J. Skepticism towards Santa Rosalia, or why are there so few kinds of animals. *Evolution* **35**, 124–138 (1981).
This seminal paper illustrates the antagonism between selection and recombination for coupling loci that convey reproductive isolation in a two-allele model with gene flow.
100. Auton, A. *et al.* A fine-scale chimpanzee genetic map from population sequencing. *Science* **336**, 193–198 (2012).
101. Gossmann, T. I., Woolfit, M. & Eyre-Walker, A. Quantifying the variation in the effective population size within a genome. *Genetics* **189**, 1389–1402 (2011).
102. Charlesworth, B. Measures of divergence between populations and the effect of forces that reduce variability. *Mol. Biol. Evol.* **15**, 538–543 (1998).
103. Roesti, M., Moser, D. & Berner, D. Recombination in the threespine stickleback genome — patterns and consequences. *Mol. Ecol.* **22**, 3014–3027 (2013).
This empirical study highlights the dependence of allele frequency shifts between populations on the genome-wide distribution of broad-scale recombination rates and chromosomal features such as centromeres.
104. Coop, G. Does linked selection explain the narrow range of genetic diversity across species? Preprint at *bioRxiv* http://dx.doi.org/10.1101/042598 (2016).
105. Reed, F. A., Akey, J. M. & Aquadro, C. F. Fitting background-selection predictions to levels of nucleotide variation and divergence along the human autosomes. *Genome Res.* **15**, 1211–1221 (2005).
106. Rockman, M. V. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. *Evolution* **66**, 1–17 (2012).
107. Le Corre, V. & Kremer, A. The genetic differentiation at quantitative trait loci under local adaptation. *Mol. Ecol.* **21**, 1548–1566 (2012).
This meta-analyses reviews expectations for allelic differentiation at QTLs and highlights the limitations of F_{ST}-based genome scans.
108. Beaumont, M. A. Adaptation and speciation: what can F_{st} tell us? *Trends Ecol. Evol.* **20**, 435–440 (2005).
109. Foll, M. & Gaggiotti, O. A. Genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* **180**, 977–993 (2008).
110. Beaumont, M. A. & Nichols, R. A. Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Lond. B Biol. Sci.* **263**, 1619–1626 (1996).
111. Roesti, M., Hendry, A. P., Salzburger, W. & Berner, D. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Mol. Ecol.* **21**, 2852–2862 (2012).
112. Zeng, K. A coalescent model of background selection with recombination, demography and variation in selection coefficients. *Heredity* **110**, 363–371 (2013).
113. Roesti, M., Gavrilets, S., Hendry, A. P., Salzburger, W. & Berner, D. The genomic signature of parallel adaptation from shared genetic variation. *Mol. Ecol.* **23**, 3944–3956 (2014).
114. Berner, D. & Salzburger, W. The genomics of organismal diversification illuminated by adaptive radiations. *Trends Genet.* **31**, 491–499 (2015).
115. Poelstra, J. W., Vijay, N., Hoeppner, M. P. & Wolf, J. B. W. Transcriptomics of colour patterning and coloration shifts in crows. *Mol. Ecol.* **24**, 4617–4628 (2015).
116. Laporte, M. *et al.* RAD-QTL mapping reveals both genome-level parallelism and different genetic architecture underlying the evolution of body shape in lake whitefish (*Coregonus clupeaformis*) species pairs. *G3 (Bethesda)* **5**, 1481–1491 (2015).
117. Winkler, C. A., Nelson, G. W. & Smith, M. W. Admixture mapping comes of age. *Annu. Rev. Genom. Hum. Genet.* **11**, 65–89 (2010).
118. Gompert, Z. & Buerkle, C. A. A powerful regression-based method for admixture mapping of isolation across genome hybrids. *Mol. Ecol.* **18**, 1207–1224 (2009).
119. Bono, J. M., Olesnicky, E. C. & Matzkin, L. M. Connecting genotypes, phenotypes and fitness: harnessing the power of CRISPR/Cas9 genome editing. *Mol. Ecol.* **24**, 3810–3822 (2015).
120. Hall, A. B. *et al.* A male-determining factor in the mosquito *Aedes aegypti*. *Science* **348**, 1268–1270 (2015).
121. Markert, M. J. *et al.* Genomic access to Monarch migration using TALEN and CRISPR/Cas9-mediated targeted mutagenesis. *G3 (Bethesda)* **6**, 905–915 (2016).
122. Strasburg, J. L. *et al.* What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Phil. Trans. R. Soc. B* **367**, 364–373 (2012).
123. Earl, D. *et al.* Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res.* **21**, 2224–2241 (2011).
124. Darwin, C. *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life* (John Murray, 1859).
125. Kohn, D. in *The Cambridge Companion to the "Origin of Species"* (eds Ruse, M. & Richards, R. J.) 87–108 (Cambridge Univ. Press, 2008).
126. Mallet, J. Mayr's view of Darwin: was Darwin wrong about speciation? *Biol. J. Linn. Soc.* **95**, 3–16 (2008).
127. Mayr, E. *Systematics and the Origin of Species* (Columbia Univ. Press, 1942).
128. Orr, H. A. The population genetics of speciation: the evolution of hybrid incompatibilities. *Genetics* **139**, 1805–1813 (1995).
129. Muller, H. J. Isolating mechanisms, evolution and temperature. *Biol. Symp.* **6**, 71–125 (1942).
130. Bateson, W. in *Darwin and Modern Science* (ed. Seward, A. C.) 85–101 (Cambridge Univ. Press, 1909).
131. Oka, H.-I. Genic analysis for the sterility of hybrids between distantly related varieties of cultivated rice. *J. Genet.* **55**, 397–409 (1957).
132. Smadja, C. M. & Butlin, R. K. A framework for comparing processes of speciation in the presence of gene flow. *Mol. Ecol.* **20**, 5123–5140 (2011).
133. Dieckmann, U. & Doebeli, M. On the origin of species by sympatric speciation. *Nature* **400**, 354–357 (1999).
134. Barluenga, M., Stolting, K. N., Salzburger, W., Muschick, M. & Meyer, A. Sympatric speciation in Nicaraguan crater lake cichlid fish. *Nature* **439**, 719–723 (2006).
135. Papadopulos, A. S. T. *et al.* Speciation with gene flow on Lord Howe Island. *Proc. Natl Acad. Sci. USA* **108**, 13188–13193 (2011).
136. Roux, C. *et al.* Shedding light on the grey zone of speciation along a continuum of genomic divergence. Preprint at *bioRxiv* http://dx.doi.org/10.1101/059790 (2016).
137. Doebeli, M. & Dieckmann, U. Speciation along environmental gradients. *Nature* **421**, 259–264 (2003).
138. Flaxman, S. M., Wacholder, A. C., Feder, J. L. & Nosil, P. Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* 4074–4088 (2014).
139. Gavrilets, S. Models of speciation: where are we now? *J. Hered.* **105**, 743–755 (2014).
140. Abbott, R. *et al.* Hybridization and speciation. *J. Evol. Biol.* **26**, 229–246 (2013).
This perspective article summarizes important aspects of speciation with gene flow.
141. Flaxman, S. M., Feder, J. L. & Nosil, P. Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution* **67**, 2577–2591 (2013).
142. van Doorn, G. S., Edelaar, P. & Weissing, F. J. On the origin of species by natural and sexual selection. *Science* **326**, 1704–1707 (2009).
143. Servedio, M. R., Doorn, G. S. V., Kopp, M., Frame, A. M. & Nosil, P. Magic traits in speciation: 'magic' but not rare? *Trends Ecol. Evol.* **26**, 389–397 (2011).
144. Thompson, M. J. & Jiggins, C. D. Supergenes and their role in evolution. *Heredity* **113**, 1–8 (2014).
145. Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931).
146. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).
147. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nat. Rev. Genet.* **10**, 639–650 (2009).
148. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl Acad. Sci. USA* **70**, 3321–3323 (1973).
149. Hedrick, P. W. A standardized genetic differentiation measure. *Evolution* **59**, 1633–1638 (2005).
150. Jost, L. G_{ST} and its relatives do not measure differentiation. *Mol. Ecol.* **17**, 4015–4026 (2008).
151. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting F_{st}: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
152. Jakobsson, M., Edge, M. D. & Rosenberg, N. A. The relationship between F_{ST} and the frequency of the most frequent allele. *Genetics* **193**, 515–528 (2013).
153. Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).

154. Yi, X. *et al.* Sequencing of fifty human exomes reveals adaptation to high altitude. *Science* **329**, 75–78 (2010).
155. Nei, M. *Molecular Evolutionary Genetics.* (Columbia Univ. Press, 1987).
156. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl Acad. Sci. USA* **76**, 5269–5273 (1979).
157. Hey, J. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Genetics* **128**, 831–840 (1991).
158. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).
159. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
160. Fu, Y. X. & Li, W. H. Statistical tests of neutrality of mutations. *Genetics* **133**, 693–709 (1993).
161. Kemppainen, P. *et al.* Linkage disequilibrium network analysis (LDna) gives a global view of chromosomal inversions, local adaptation and geographic structure. *Mol. Ecol. Resour.* **15**, 1031–1045 (2015).
162. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837 (2002).
163. Sabeti, P. C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).
164. Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G. & Schierup, M. H. Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLoS Genet.* **7**, e1001319 (2011).
165. Zamani, N. *et al.* Unsupervised genome-wide recognition of local relationship patterns. *BMC Genomics* **14**, 347 (2013).

Acknowledgements
The authors are grateful to members of their laboratory groups and to many external visitors for years of stimulating discussions on the subject. Their research is supported by the European Research Council, the Swedish Research Council, and the Knut and Alice Wallenberg Foundation.

Competing interests statement
The authors declare no competing interests.

SUPPLEMENTARY INFORMATION
See online article: S1 (table)
ALL LINKS ARE ACTIVE IN THE ONLINE PDF