Trends in Ecology & Evolution

CellPress REVIEWS

Opinion

Unifying Theoretical and Empirical Perspectives on Genomic Differentiation

Georgy A. Semenov,[1,*,@] Rebecca J. Safran,[1] Chris C.R. Smith,[1] Sheela P. Turbek,[1] Sean P. Mullen,[2] and Samuel M. Flaxman[1]

Differentiation is often heterogeneous across the genomes of diverging populations. Despite substantial recent progress, much work remains to improve our abilities to connect genomic patterns to underlying evolutionary processes. Crosstalk between theoretical and empirical research has shaped the field of evolutionary genetics since its foundation and needs to be greatly enhanced for modern datasets. We leverage recent insights from theoretical and empirical studies to identify existing gaps and suggest pathways across them. We stress the importance of reporting empirical data in standardized ways to enable meta-analyses and to facilitate parameterization of analyses and models. Additionally, a more comprehensive view of potential mechanisms – especially considering variable recombination rates and ubiquitous background selection – and their interactions should replace common, oversimplified assumptions.

Explaining Patterns of Heterogeneous Genomic Differentiation

Decades of empirical studies have revealed that genetic variation within and between populations is distributed in a heterogeneous fashion across the genome. These discoveries raised substantial interest in connecting observed patterns of genomic differentiation to underlying processes, due to appreciation of their potential to inform us about the mechanistic basis of divergent evolution [1–4]. As genomic sequencing technology has continued to develop, focus has shifted from studying single markers to understanding: (i) how sites and regions of the genome evolve with varying degrees of independence from or dependence upon each other; (ii) how continuous the process of differentiation is and whether it contains distinct stages; (iii) how different evolutionary processes may promote or prevent transitions between these stages; and (iv) how evolutionary processes may be inferred from population genetic data.

It is now possible to obtain large amounts of genome-wide data from virtually any organism, yet we continue to struggle to confidently infer process from pattern. Here we highlight some of the gaps that remain in evolutionary genomics to help construct a roadmap for productive future work. We focus on ways to enhance synergy between theoretical and empirical approaches to: (i) make inferences about evolutionary processes and transitions from single-locus to multilocus processes during population divergence; and (ii) explain patterns of **heterogeneous genomic differentiation** (**HGD**; see Glossary) (Figure 1, Key Figure).

Contemporary studies of evolutionary genomics have been profoundly shaped by the **genic view** of genomic evolution [5]. This framework highlighted the primary role of interactions between divergent **selection**, recombination, and gene flow in creating HGD and raised enthusiasm about finding loci relevant for local adaptation and reproductive isolation. More recently, theoretical work (e.g., [6,7]) and empirical data (e.g., [8,9]) have increasingly emphasized caution by showing that the interpretation of HGD is less straightforward than a heuristic genic

Highlights

There is a growing need to enhance crosstalk between theoretical and empirical research in evolutionary genomics.

The lack of a unified framework for quantitative characterization of population genomic divergence hampers rigorous hypothesis testing about evolutionary processes. We propose steps toward creating one.

Common assumptions of theoretical studies and empirical analysis methods neglect consequential variation in recombination and mutation rates, gene density, spatial arrangements of populations, and population size.

Future empirical studies should incorporate recombination rates, population size variation, and explicit genomic estimates of admixture whenever feasible.

Theory provides clear predictions about evolutionary mechanisms when they act alone; predicting the combined effects of multiple mechanisms in empirically testable ways is a challenge for the future.

Reporting genomic analyses in a more standardized way could enhance future meta-analyses.

[1]University of Colorado Boulder,
Boulder, CO, USA
[2]Boston University, Boston, MA, USA

*Correspondence:
georgy.semenov@colorado.edu
(G.A. Semenov).
@Twitter: @georgy_semenov
(G.A. Semenov).

https://doi.org/10.1016/j.tree.2019.07.008

view of speciation would suggest. The current consensus is that various mechanisms, on their own or in concert, can lead to peaks and troughs of differentiation, leading to much more nuance and uncertainty than verbal predictions tend to represent (Figure 1). Recent reviews have focused on interpreting HGD in terms of speciation processes [3,10] or providing an overview of how signatures of divergent selection and reproductive isolation can be distilled from noise created by other mechanisms [4,11,12]. Of particular note, these recent syntheses highlight that in empirical studies it is rarely the case that predictions of multiple hypotheses are simultaneously explored. Instead, patterns are detected, and possible explanations are offered but not thoroughly tested. While theory can provide a framework for testing these distinct alternatives, the modeling software and approaches that are employed by theoretical studies often rely on assumptions and parameters that are not met empirically.

Comparing Recent Theoretical and Empirical Advances in Understanding HGD Reveals Existing Gaps

HGD can be quantitatively described in terms of: (i) the heights and widths of peaks and troughs of differentiation; (ii) distances between divergent sites; and (iii) the magnitude of **linkage disequilibrium (LD)** (Figure 1). Our review of recent theoretical and empirical studies (Table S1 in the supplemental information online) revealed that, despite the fact that, superficially, the same summary statistics are commonly used across theoretical and empirical studies to characterize genomic patterns (Box 1), they are reported in disparate ways. For example, there is little standardization with respect to the filtering criteria used to generate the summary statistics including the handling of missing data and the size and overlap of sliding windows, despite the broad recognition that there are substantial effects of these parameters for the quantification of HGD [13–16]. Further, different mechanisms (Figure 1C) have received uneven attention, with the majority of studies – both empirical and theoretical – focusing on signatures of divergent selection and gene flow.

Theoretical models predict clear distinctions between the statistical properties of divergently selected and neutral loci [17]. Hence, loci with exceptionally elevated differentiation (outliers) between populations or ecotypes are frequently inferred to be candidate genomic regions associated with adaptation and reproductive isolation. However, a growing body of recent theory indicates that multiple mechanisms have the potential to obscure true peaks or generate false ones, suggesting reasons to question those inferences (Figure 1). As a consequence, the cautionary conclusion of recent studies is that many commonly used empirical analyses are likely to have reported large fractions of false positives and false negatives [6,7,18–20]. These statistical problems are especially hard to avoid when divergent evolution has a multigenic basis [21] and particularly when it is acting on complex gene regulatory networks [22]. These issues underscore the need for greatly enhanced exchange between empirical and theoretical work to test hypotheses about HGD. Which pathways are likely to be the most fruitful toward that end? To answer this question, we first briefly note key considerations of empirical and theoretical approaches related to the major evolutionary factors highlighted in Figure 1.

Selection

Recent theory has focused on the combination of direct and indirect effects of selection in two main ways: (i) the magnitude of **hitchhiking** effects [23]; and (ii) the reliability of statistical methods for detecting selected loci (noted above). On the empirical side, making inferences about divergent selection is frequently one of the main goals of many studies. In systems with detailed knowledge of genotype-to-phenotype relationships, and in systems where carefully designed selection experiments can be conducted, causal relationships between genetic variants, fitness, and differentiation have received strong empirical support (Box 2). However, in many ‘nonmodel’

Glossary

Absolute divergence (d_{XY}): the number of sequence differences observed between two samples.

Causal variants: genetic variation underlying differences in phenotypes and/or fitness.

Coupling: multiple loci having joint effects that lead to much stronger reproductive isolation than any individual locus would on its own.

Effective population size (N_e): the number of breeding individuals in an idealized population.

Genic view: evolutionary forces act heterogeneously across the genome on a locus-by-locus basis; the locus as the functional unit of evolution.

Genomic scans for differentiation: statistical methods used to find signatures of selection in the genome. Typically, a fraction of genomic variation is characterized as statistical outliers.

Heterogeneous genomic differentiation (HGD): variability in the relative and absolute magnitude of divergent genomic regions, their span along chromosomal segments, their patterns of distribution within chromosomes, and the number of such sites or regions within and between chromosomes.

Hitchhiking: indirect effects of selection on patterns of allelic variation in the genome.

Linkage disequilibrium (LD): nonrandom associations between alleles at different loci (which may be located on the same or different chromosomes).

Locus clustering: colocalization (on the same chromosome) of loci relevant for local adaptation or reproductive isolation.

Missing heritability: the portion of phenotypic variance unexplained by associated loci in genome-wide association studies.

Selection: divergent selection arises due to distinct evolutionary pressures in different populations. Background selection is selection against unconditionally deleterious mutations. For any type of selection, it is common to distinguish between direct selection acting on a specific locus and indirect selection affecting loci indirectly with strength dependent on their LD with a locus under direct selection.

Stages of divergence: over the course of divergent evolution, genomic differentiation between populations may organisms, inferences about selection are made primarily by identifying outliers in **genomic scans for differentiation**. These tests often rely on consequential assumptions about mutation rates, recombination, and demographic history that are nearly impossible to validate in the absence of rich genomic resources (e.g., a high-quality reference genome).

Genetic Architecture of Adaptation and Reproductive Isolation

Genetic architecture refers to the effect size (on fitness) of a locus as well as the number and genomic distribution of loci that underlie a phenotype targeted by selection. In a homogeneous genomic background, adaptation underlain by a few genes of large effect is predicted to lead to strong differentiation and easier empirical detection [24]. At the same time, simple architectures have less potential to promote genome-wide reductions of gene flow, and hence genome-wide differentiation, than polygenic traits [25]. This leads to an ascertainment bias toward traits with simple genetic architectures (Box 2) and implies greater difficulty when seeking additional **causal variants** and working out **missing heritability**.

Recombination

Theory has illuminated the key role of the ratio of selection to recombination in determining whether **locus clustering** [26] and **coupling** [27–29] could be stable. Recent theory highlights the role of intragenomic heterogeneity in recombination rates in generating HGD [30] and, specifically, large 'islands' of differentiation, even when those islands do not contain barrier loci [20]. In sum, any mechanism that lowers the recombination rate in one region of the genome (especially when recombination is strongly suppressed; e.g., in large chromosomal rearrangements) can increase differentiation and clustering in that region by a variety of mechanisms, particularly background selection (e.g., [12,31]). Also, crossover frequency varies with regard to location along the chromosome (e.g., center vs periphery), chromosome size, gene density [32], chromatin modifications [33], and structural variants [34]. Hence, recombination is increasingly recognized as a potentially powerful factor shaping HGD [20,35–37]. However, few theoretical studies have included realistically variable genetic maps and – due to the difficulty of generating genetic maps in nonmodel organisms – recombination rates rarely inform hypothesis testing in most empirical studies (Box 3). We suggest that this is one of the most important areas for the development of future theory and empirical analyses in concert.

Gene Flow

Early in the process of differentiation, theory predicts that gene flow makes substantial, sustained differentiation difficult or impossible at neutral and weakly selected sites [38–43]. Furthermore, even low levels of ongoing gene flow cause loss of differentiation at neutral sites in the long run. Thus, in the absence of mitigating factors (Figure 1), theory predicts that persistent gene flow should make it easier to detect which sites in the genome play a role in local adaptation and reproductive isolation [44–47]. Empirical studies [48–50] provide data consistent with a role for gene flow coupled with genomically localized selection in promoting a few narrow and strong divergent peaks surrounded by the near absence of differentiation throughout the rest of genome (Box 2). However, a growing number of studies suggest that it may often be more difficult to discern the role of gene flow in shaping HGD [25]. Furthermore, detection becomes more difficult if strong multilocus coupling has occurred, causing a transition to a state of high LD (Figure 1).

Population Size Changes

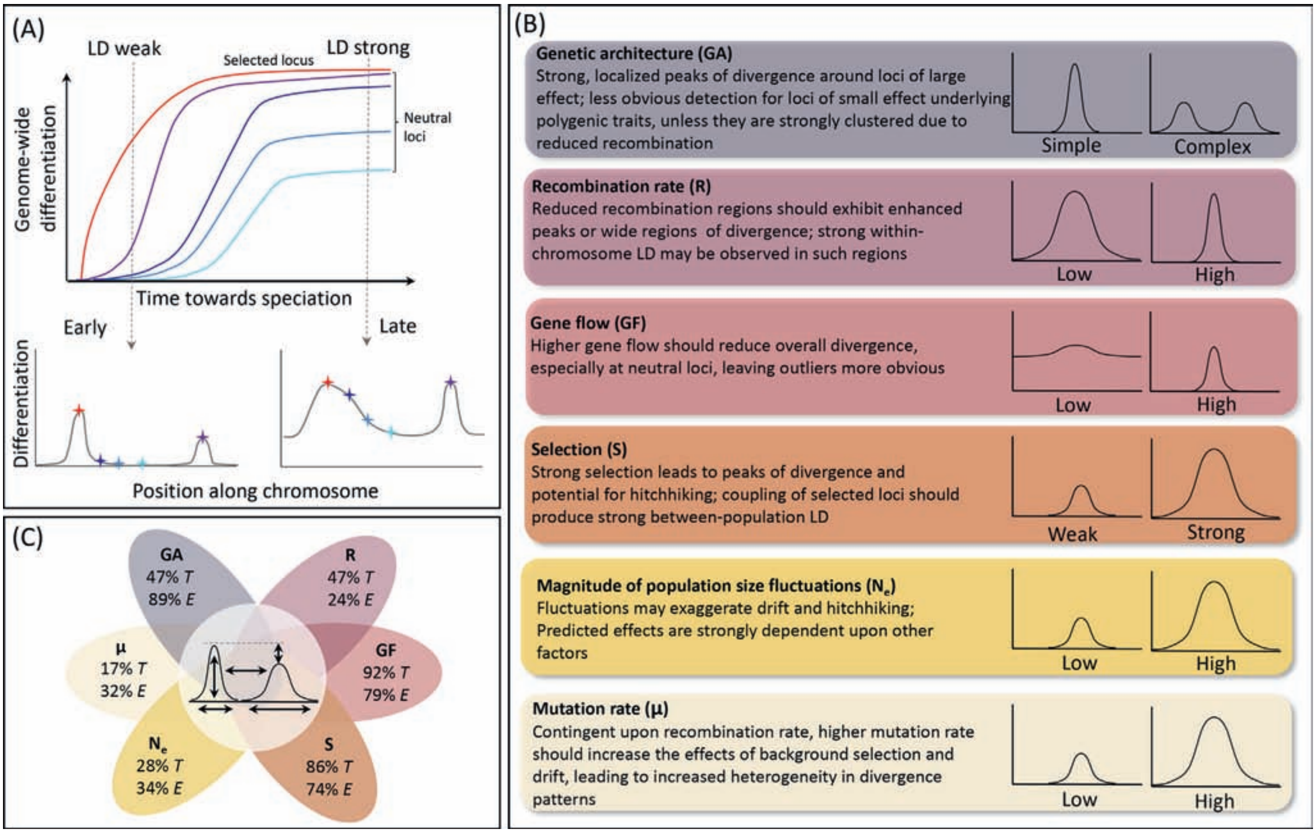
Effective population size (N_e) can vary due to historical dynamics of population abundance and can also vary across the genome (e.g., sex chromosomes vs autosomes). Both theory and empirical studies demonstrate that an understanding of N_e dynamics is essential to interpret

begin from a few narrow regions and expand toward substantial genome-wide differentiation through a series of intermediate steps. This process is often viewed as a continuum; however, the changes need not be linear in time.

Summary statistics: scalar metrics that characterize genomic variation through calculations that are usually in some way dependent on allele frequencies (Figure 1). To reduce noise, summary statistics are often averaged across genomic 'windows' spanning from a few hundred to tens of thousands of base pairs.

Key Figure

Conceptual Overview of the Process of Genomic Differentiation


(A)
LD weak
Selected locus
LD strong
Neutral loci
Genome-wide differentiation
Time towards speciation
Early
Late
Differentiation
Position along chromosome
(B)
Genetic architecture (GA)
Strong, localized peaks of divergence around loci of large effect; less obvious detection for loci of small effect underlying polygenic traits, unless they are strongly clustered due to reduced recombination
Simple
Complex
Recombination rate (R)
Reduced recombination regions should exhibit enhanced peaks or wide regions of divergence; strong within-chromosome LD may be observed in such regions
Low
High
Gene flow (GF)
Higher gene flow should reduce overall divergence, especially at neutral loci, leaving outliers more obvious
Low
High
Selection (S)
Strong selection leads to peaks of divergence and potential for hitchhiking; coupling of selected loci should produce strong between-population LD
Weak
Strong
Magnitude of population size fluctuations (N_e)
Fluctuations may exaggerate drift and hitchhiking; Predicted effects are strongly dependent upon other factors
Low
High
Mutation rate (μ)
Contingent upon recombination rate, higher mutation rate should increase the effects of background selection and drift, leading to increased heterogeneity in divergence patterns
Low
High
(C)
GA 47% T 89% E
R 47% T 24% E
μ 17% T 32% E
GF 92% T 79% E
N_e 28% T 34% E
S 86% T 74% E

Trends in Ecology & Evolution

Figure 1. Genomic differentiation arises from mechanisms that create allele frequency differences and linkage disequilibrium (LD) between loci in a pair of populations (A). During the early **stages of divergence**, selection can promote local allele frequency sweeps (shown in red) that affect allele frequencies of loci in close physical proximity (shades of blue) due to indirect selection. The magnitude of indirect selection can increase over evolutionary time due to locus clustering or due to coupling, which can include loci that are far apart on the same chromosome or on different chromosomes (red and purple), leading to patterns that are more idiosyncratic than simple verbal models have implied. Temporally, the strength of these associations can change nonlinearly during divergence. Various mechanisms (B,C) can ultimately determine the absolute and relative size, width, and arrangement of divergent regions in the genome [double-sided arrows in (C)]. There has been substantial bias in research focus toward particular mechanisms as shown in (C), where numbers indicate the percentage of theoretical (*T*, n = 36) and empirical (*E*, n = 38) studies considering each mechanism from a set of recent papers we surveyed (Table S1 in the supplemental information online). The cartoons in panel (B) are intentionally drawn simplistically to underscore two key points: (i) theory provides clear predictions for simplistic scenarios in which one factor is considered in isolation and all others are ignored; and (ii) different mechanisms can produce similar patterns, confounding hypothesis testing. Predictions become nuanced when multiple factors operate in concert.

HGD, because fluctuations in N_e strongly impact allele frequencies and hitchhiking, and thus most summary statistics [51]. Accordingly, N_e is widely appreciated as a crucial parameter for deriving neutral expectations [7]. Further, a number of empirical studies have considered variation in N_e as a null hypothesis and proceeded by asking how much of the observed variation in differentiation can be explained by fluctuations in N_e alone, with selection being invoked only when observed patterns are far outside those plausibly attributed to changes in N_e [46]. Nonetheless, only about one-third of empirical studies (Figure 1C) accounted for N_e while interpreting genomic patterns.

Box 1. Promising Pathways Forward for Characterizing and Explaining HGD

Differentiation between genomic samples is typically quantified via summary statistics, most of which rely on patterns of nucleotide diversity and allele frequencies [3]. Multiple metrics analyzed simultaneously can provide better inferences about underlying processes, but interpretation of patterns can be fraught with the issues discussed in the main text. Recent developments in supervised machine learning show great promise in overcoming these challenges due to their ability to accommodate complex patterns and search for signals of local adaptation and reproductive isolation while accounting for other factors, such as nonequilibrium demography and recombination rates [58]. Continuing to explore the properties of both new and old summary statistics – such as absolute allele frequency differences [16] – is important as well.

In addition to methods for individual systems, another major challenge remains with respect to conducting comparative studies using published data. One solution for this hurdle could be in reporting called genotype data in a standardized format that would account for the most sought-after types of information. First, we suggest that unfiltered VCFs should always be archived, along with metadata on how they were produced. Although large to store, VCFs represent the results of valuable workflows in which expert knowledge of the specific study system and analytical pipelines has been applied. Second, to facilitate meta-analyses and theoretical–empirical crosstalk in population genomics, we suggest the creation and archiving of a plain-text table containing a basic summary of each SNP in each population, such as the following columns of data: SNP ID, scaffold ID, location on scaffold, reference allele, alternative allele, sequence quality metrics, minor allele frequency in population 1, sample size in population 1, minor allele frequency in population 2, etc. Recombination map information could be added if available. The benefits of such a format would be several: (i) the file is likely to be much smaller than a VCF; (ii) it would be easily usable by anyone with basic, open-source analysis software programs; (iii) it could be easily used by theoreticians to create realistic variation in models; (iv) statistics from different study systems could be easily aggregated and compared; and (v) consistent filtering strategies and windowed averages of statistics could be easily applied if desired. We provide a toy example in Table S2 in the supplemental information online, which was produced from a previously published VCF [17]. We seek to more fully develop this in the near future.

Mutation Rate (μ)

Especially in coalescent models, the mutation rate μ (along with N_e) is a key parameter for predicting numbers of segregating sites, levels of **absolute divergence (d_{xy})**, and heterozygosity [52]. In forward-time simulations, these statistics are affected by choice of μ as well, but μ may be a choice of convenience: because forward-time simulations are often computationally expensive, theoreticians sometimes shorten run times by elevating μ above empirically reasonable

Box 2. Lessons from Empirical Studies and Remaining Gaps

Case studies linking patterns of genomic variation and underlying processes ultimately require functional validation of candidate mechanisms [59]. Such validation can come in the form of physiological or gene expression assays, experimental genetics, and/or field data directly measuring dispersal, mate choice, or fitness in natural settings. Strong evidence supports the role of divergent selection alone or in combination with gene flow in promoting a few prominent peaks corresponding to phenotypic targets of selection (e.g., [60–62]). Many studies to date have focused on traits with simple monogenic or oligogenic architectures, which are expected to leave the most prominent genomic signatures and hence are comparatively easier to dissect genomically. Nonetheless, in the majority of such 'easy' cases, the mechanistic relationships between genetic variation and evolutionary processes remain poorly understood and need to be further studied. The situation is even more complicated with polygenic traits, as selection on many small-effect loci may not be detectable using genomic scans alone [63–65], and methods such as admixture mapping should be used to corroborate connections between phenotypic and/or experimental data and HGD [59].

Another theme of many empirical studies is parallel evolution, which relates to the repeatability and predictability of evolution [66,67]. A number of studies have documented that differentiation accumulates repeatedly in the same genomic regions across closely related lineages, between populations connected by substantial gene flow as well as between closely related biological species [68–70]. Are these regions: repeated targets of recent divergent selection; selection in ancestral populations; due to background selection modulated by reduced crossover rate; regions with higher densities of selection targets; or all of the above?

Finally, a key consideration for hypothesis testing is the use of null or default models. Neutral processes have long been considered a kind of 'null' model of evolution, but recent data suggest that background selection may have abundant genome-wide effects on genetic variation, especially over long timescales [71]. Hence, in the future it is important that the default assumptions used in modeling and analysis software incorporate background selection as well as the potential for variable recombination rates (Box 3). Furthermore, it is often the case that divergent (e.g., ecological) selection operates in a study system of interest but the genetic architecture of selection targets is unknown. In such cases, a null model with only neutral processes and background selection may be not sufficient [72].

Box 3. Accounting for Variable Crossover Rates

Despite a longstanding appreciation that variation in recombination rate can modulate the effects of selection and gene flow [34,35,45], recombination maps have been accounted for in relatively few empirical studies. Likewise, in most theoretical models recombination is assumed to be either uniform or absent. Recombination patterns can be estimated using cytological methods [73] or linkage mapping [74] and from LD patterns by leveraging high-density genomic data [75]. Cytological methods allow mapping of recombination locations in meiotic cell spreads via immunostaining of proteins involved in the formation of synaptonemal complexes, centromeres, and crossovers. Cytological methods provide a straightforward way to compare broad-scale recombination patterns across taxa; however, they offer only limited opportunity to match with genome-wide sequencing data, involving the use of costly fluorescence *in situ* hybridization probes. Linkage mapping localizes crossover breakpoints by tracing parent–offspring relationships or multigenerational pedigrees. The obvious limitation of this approach is the need for information about parent–offspring relationships, which restricts its application among nonmodel organisms to those that are tractable. Statistical approaches utilizing patterns of LD are based on the negative relationship between recombination frequency and LD, and such methods have been increasingly applied in recent years. However, areas of the genome under strong divergent selection may also show strong LD (see Figure 1 in main text) even if they are in a region with high crossover frequency, therefore biasing recombination estimates. Some studies indicate good correspondence between linkage maps and LD-based methods on a genome-wide scale [9], but the field of evolutionary genomics will benefit from more studies addressing this question, especially at finer scales.

Newer analytical pipelines allow estimation of crossover frequency by sequencing sperm samples [76,77] or a single pair of genomes [78] and have the potential to revolutionize recombination mapping in nonmodel organisms. There is, however, compelling evidence that recombination patterns may vary broadly between individuals of the same sex and between males and females [79], although there are some emerging common patterns in the crossover landscape among species at a larger physical scale [31]. Overall, we stress that intragenomic recombination-rate variation should be considered in default theoretical and empirical models of HGD.

estimates. In empirical studies, estimations of μ are highly affected by uncertainty in generation times, divergence times between populations, and ancestral population sizes. Hence, μ is often either derived from organismal systems for which there are a long history of work and many genetic resources (e.g., data on substitution rates and appropriate outgroups available) or estimated as part of a compound parameter, $4N_e\mu$. A few studies have measured μ and demonstrated that it can vary substantially [53], but in general little attention is given to intragenomic variation in μ as a possible explanation for HGD.

Concluding Remarks and Future Perspectives

Recent theoretical and empirical studies have revealed several emergent trends in understanding the mechanistic basis of HGD. First, the extent to which recombination rate modulates the effects of selection, drift, and gene flow – and hence promotes HGD – has not been widely acknowledged until recently and is a crucial emphasis for the future. Second and related, the potential for chromosomal rearrangements to promote localized peaks of genomic differentiation is known yet mostly ignored when we focus overwhelmingly on SNPs. Third, the detection of barrier loci is much more difficult than suggested by the predictions of many theoretical models; this discrepancy is due in part to common modeling assumptions that happen to lead to especially clear predictions. For example, false positives will be less common in the familiar spatial assumption of two parapatric demes than in other spatial scenarios [7]. Simple models of a complex world are frequently useful, but not when they lead to false confidence in inferences.

Predictions about HGD become nuanced when one simultaneously considers realistic combinations of different factors (Figure 1); assuming that one factor can be ignored or held constant will have a large impact on predictions. Some of the modifying effects of factors on one another have long been appreciated [40]; for example, divergence with gene flow should be more difficult under high rates of recombination and gene flow, and when a diffuse genetic architecture underlies selected traits. Furthermore, the effects of different mechanisms can be strongly correlated in a genome, making it difficult to distinguish between cause and consequence. For example,

Outstanding Questions

Will evolutionary genomics remain a historical science, with very limited predictability even in the best-known organismal systems? Can we learn enough about the effects of evolutionary processes on HGD to infer evolutionary processes from empirical patterns or will patterns in HGD rarely serve to evaluate *a priori* hypotheses, instead producing *post hoc* hypotheses to be evaluated using simulations?

Can the field of evolutionary genomics reach the lofty goal of rigorous testing of multiple hypotheses? Multiple evolutionary processes can leave similar patterns in genome-scan data. Can more nuanced models and approaches help to disentangle alternative scenarios?

Is it possible to make generalizations that go beyond particular study systems? Some extreme cases (e.g., few peaks of differentiation with other parts of the genome remaining homogeneous) seem to be driven by the same processes across some study systems. Is it possible to derive such general conclusions for more complex patterns and across a broader range of organisms?

Why do some genomic regions appear to be involved repeatedly in differentiation? Are these regions consistently targets of divergent selection or, perhaps, of selection in ancestral populations? Are these frequent targets of background selection and reduced recombination? Do they have exceptional densities of selection targets? Is there some unique combination of factors operating on these regions?

What genomic resolution is optimal to infer the role that distinct mechanisms play in shaping HGD? How much detail is needed for analyses? The effects of variable N_e, mutation rate, migration, and recombination rates will differ depending on how much we ‘zoom into’ a particular genomic region. How do we choose optimal windows sizes for summarizing genomic patterns and eliminating noise created by single-locus estimates?

nucleotide composition (GC content) correlates with recombination rate [54]. Is recombination higher where there is more GC or is there more GC where the recombination rate is higher?

To create increased alignment and synergy between future theoretical and empirical work, we suggest the following as fruitful areas of exploration (see also Outstanding Questions). Realization of the complexity of mechanisms affecting HGD calls for more complex simulation approaches. This is especially timely because predictions from simplistic models have given empiricists confidence in findings that are likely to be either false positives or the exception rather than the rule (e.g., traits with simple architectures involving loci of large effect). Simulation studies should therefore aim to incorporate the range of variation in data from natural systems to generate refined predictions about the situations in which the best available outlier methods alone are likely to have high success. Conversely, this will help to identify situations where progress can be made only with complementary studies of natural history and experimental genetics and genomics. Some specific considerations for future theoretical studies should be spatial arrangements of populations that are more complex than the classical two-deme scenario, demographic changes, recombination-rate heterogeneity, and realistic mutation rates (possibly with intragenomic variation). Exploring such expansive parameter spaces is challenging and generalizations may be difficult. In some cases, it may be necessary to tailor simulations to specific systems under different scenarios rather than search for general predictions. It would also be useful to generate pseudodata in ways that mirror the uncertainty, limitations, data formats, and data-filtering strategies of empirical sequencing technologies.

On the empirical side, HGD is described in disparate ways, complicating comparisons between study systems. The now common standard of publishing raw sequence data is essential; the creation of additional standards would greatly facilitate comparative studies, meta-analyses, and synergy between theoretical and empirical approaches. The idea that results of different studies could be standardized might appear unrealistic but we suggest that it is well worth trying, and perhaps vital for making forward progress in our understanding of HGD. Although they are very large, we suggest that Variant Call Format (VCF) files should be publicly archived as part of publications and could be complemented by an additional common data format for easy analysis of HGD. We provide ideas in this direction in Box 1 and a small proof-of-concept example in Table S2 in the supplemental information online.

Additionally, a more comprehensive view of potential mechanisms should be taken whenever possible, accounting for recombination-rate heterogeneity, genomically localized patterns of gene flow, and variable N_e (e.g., [9,55]). Understanding the conditions under which multiple barrier loci become coupled with one another [28] – thereby acting almost like a single selected locus of very large effect – has been another focus of theory for several decades. Empirical studies of coupling remain limited (reviewed in [56]) and patterns of HGD are rarely placed into a framework that considers coupling and temporal transitions, but pursuing this is likely to provide a clearer context in which to interpret results. Further, comparisons of parallel divergence, where replicate sets of sister taxa have diverged in response to a similar form of divergent selection but vary in demographic history and/or levels of gene flow, can be more powerful than pairwise estimates of differentiation and would help to disentangle the effects of neutral processes from the drivers of adaptation and speciation (Box 2).

Finally, greater crosstalk can be achieved if theoretical studies strive to use empirically informed parameter space. We contend that the standards suggested above and in Box 1 would greatly facilitate this type of meaningful exchange. It is crucial for empiricists and theoreticians to understand what their counterparts need to know and use in their available methods. This will additionally include information that does not seem directly related to genomics but that turns out to be often relevant, such as generation times, age of reproduction, and other life-history traits. Otherwise, disconnection will further impede the development of the field of evolutionary genomics and will favor the continued coexistence of two 'parallel worlds' [57], hindering understanding of evolutionary processes.

Acknowledgments

We thank Marius Roesti, Mark Ravinet, Roger Butlin, and two anonymous reviewers for helpful comments on earlier drafts of this article. G.A.S. was supported by the National Science Foundation (NSF) DEB-CAREER 1149942 award to R.J.S. G.A.S. and S.M.F. were supported by NSF award 1627483 to S.M.F.

Supplemental Information

Supplemental information associated with this article can be found online https://doi.org/10.1016/j.tree.2019.07.008.

References

1. Butlin, R. *et al.* (2012) What do we need to know about speciation? *Trends Ecol. Evol.* 27, 27–39
2. Seehausen, O. *et al.* (2014) Genomics and the origin of species. *Nat. Rev. Genet.* 3, 176–192
3. Wolf, J.B. and Ellegren, H. (2017) Making sense of genomic islands of differentiation in light of speciation. *Nat. Rev. Genet.* 18, 87–100
4. Ravinet, M. *et al.* (2017) Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J. Evol. Biol.* 30, 1450–1477
5. Wu, C.-I. (2001) The genic view of the process of speciation. *J. Evol. Biol.* 14, 851–865
6. Cruickshank, T.E. and Hahn, M.W. (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23, 3133–3157
7. Lotterhos, K.E. and Whitlock, M.C. (2014) Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Mol. Ecol.* 23, 2178–2192
8. Vijay, N. *et al.* (2016) Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nat. Commun.* 7, 13195
9. Martin, S.H. *et al.* (2019) Recombination rate variation shapes barriers to introgression across butterfly genomes. *PLoS Biol.* 17, 1–28
10. Campbell, C.R. *et al.* (2018) What is speciation genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biol. J. Linn. Soc.* 124, 561–583
11. Hoban, S. *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397
12. Burri, R. (2017) Interpreting differentiation landscapes in the light of long-term linked selection. *Evol. Lett.* 1, 118–131
13. Lowry, D.B. *et al.* (2017) Breaking RAD: an evaluation of the utility of restriction site-associated DNA sequencing for genome scans of adaptation. *Mol. Ecol. Res.* 17, 142–152
14. Roesti, M. *et al.* (2012) Uninformative polymorphisms bias genome scans for signatures of selection. *BMC Evol. Biol.* 12, 94
15. Kofler, R. *et al.* (2016) Suitability of different mapping algorithms for genome-wide polymorphism scans with pool-seq data. *G3 (Bethesda)* 6, 3507–3515
16. Berner, D. (2019) Allele frequency difference AFD – an intuitive alternative to FST for quantifying genetic population differentiation. *Genes* 10, 308
17. Schilling, M.P. *et al.* (2018) Transitions from single- to multi-locus processes during speciation. *Genes* 9, 274
18. Narum, S.R. and Hess, J.E. (2011) Comparison of FST outlier tests for SNP loci under selection. *Mol. Ecol. Res.* 11, 184–194
19. Vilas, A. *et al.* (2012) A simulation study on the performance of differentiation-based methods to detect selected loci using linked neutral markers. *J. Evol. Biol.* 25, 1364–1376
20. Berner, D. and Roesti, M. (2017) Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. *Mol. Ecol.* 26, 6351–6369
21. Hermisson, J. and Pennings, P.S. (2017) Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods Ecol. Evol.* 8, 700–716
22. Boyle, E.A. *et al.* (2017) An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169, 1177–1186
23. Maynard Smith, J.M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35
24. Shaw, K.L. and Mullen, S.P. (2011) Genes versus phenotypes in the study of speciation. *Genetica* 139, 649–661
25. Flaxman, S.M. *et al.* (2014) Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Mol. Ecol.* 23, 4074–4088
26. Roesti, M. (2018) Varied genomic responses to maladaptive gene flow and their evidence. *Genes* 9, 298
27. Barton, N.H. (1983) Multilocus clines. *Heredity* 37, 454–471
28. Barton, N.H. and Bengtsson, B.O. (1986) The barrier to genetic exchange between hybridizing populations. *Heredity* 57, 357–376
29. Gavrilets, S. (2004) *Fitness Landscapes and the Origin of Species*, Princeton University Press
30. Lotterhos, K.E. (2019) The effect of neutral recombination variation on genome scans for selection. *G3 (Bethesda)* 9, 1851–1867
31. Burri, R. *et al.* (2015) 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
32. Haenel, Q. *et al.* (2018) Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. *Mol. Ecol.* 25, 238–259
33. Nicolas, A. (1998) Relationship between transcription and initiation of meiotic recombination: toward chromatin accessibility. *Proc. Natl. Acad. Sci. U. S. A.* 95, 87–89
34. Butlin, R.K. (2005) Recombination and speciation. *Mol. Ecol.* 14, 2621–2635
35. Roesti, M. *et al.* (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Mol. Ecol.* 21, 2852–2862
36. Noor, M.A. and Bennett, S.M. (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity* 103, 439–444
37. Aeschbacher, S. *et al.* (2017) Population-genomic inference of the strength and timing of selection against gene flow. *Proc. Natl. Acad. Sci. U. S. A.* 114, 7061–7066
38. Barton, N.H. (1979) The dynamics of hybrid zones. *Heredity* 43, 341–359
39. Barton, N.H. and Hewitt, G.M. (1985) Analysis of hybrid zones. *Annu. Rev. Ecol. Syst.* 16, 113–148

40. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17, 368–376

41. Gavrilets, S. (2003) Models of speciation: what have we learned in 40 years? *Evolution* 57, 2197–2215

42. Bolnick, D.I. and Fitzpatrick, B.M. (2007) Sympatric speciation: models and empirical evidence. *Annu. Rev. Ecol. Syst.* 38, 459–487

43. Pinho, C. and Hey, J. (2010) Divergence with gene flow: models and data. *Annu. Rev. Ecol. Syst.* 41, 215–230

44. Payseur, B.A. (2010) Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Mol. Ecol. Res.* 10, 806–820

45. Nachman, M.W. and Payseur, B.A. (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 367, 409–421

46. Kronforst, M.R. *et al.* (2013) Hybridization reveals the evolving genomic architecture of speciation. *Cell Rep.* 5, 666–677

47. Harrison, R.G. and Larson, E.L. (2014) Hybridization, introgression, and the nature of species boundaries. *J. Hered.* 105, 795–809

48. Martin, S.H. *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 23, 1817–1828

49. Toews, D.P. *et al.* (2016) Plumage genes and little else distinguish the genomes of hybridizing warblers. *Curr. Biol.* 26, 2313–2318

50. Knief, U. *et al.* (2019) Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone. *Nat. Ecol. Evol.* 3, 570–576

51. Ferchaud, A.L. and Hansen, M.M. (2016) The impact of selection, gene flow and demographic history on heterogeneous genomic divergence: three-spine sticklebacks in divergent environments. *Mol. Ecol.* 25, 238–259

52. Wakeley, J. (2009) *Coalescent Theory* (1st edn), W.H. Freeman

53. Hodgkinson, A. and Eyre-Walker, A. (2011) Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* 12, 756–766

54. Roesti, M. *et al.* (2013) Recombination in the threespine stickleback genome – patterns and consequences. *Mol. Ecol.* 22, 3014–3027

55. Samuk, K. *et al.* (2017) Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Mol. Ecol.* 26, 4378–4390

56. Butlin, R.K. and Smadja, C.M. (2018) Coupling, reinforcement, and speciation. *Am. Nat.* 191, 155–172

57. Fitzpatrick, C.L. *et al.* (2018) Theory meets empiry: a citation network analysis. *Bioscience* 68, 805–812

58. Schrider, D.R. and Kern, A.D. (2018) Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* 34, 301–312

59. Tiffin, P. and Ross-Ibarra, J. (2014) Advances and limits of using population genetics to understand local adaptation. *Trends Ecol. Evol.* 29, 673–680

60. Van Belleghem, S.M. *et al.* (2017) Complex modular architecture around a simple toolkit of wing pattern genes. *Nat. Ecol. Evol.* 1, 52

61. Poelstra, J.W. *et al.* (2014) The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* 344, 1410–1414

62. Lamichhaney, S. *et al.* (2015) Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* 518, 371–375

63. Corre, V.L. and Kremer, A. (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Mol. Ecol.* 21, 1548–1566

64. Yeaman, S. (2015) Local adaptation by alleles of small effect. *Am. Nat.* 186, S74–S89

65. Arnegard, M.E. *et al.* (2014) Genetics of ecological divergence during speciation. *Nature* 511, 307–311

66. Nosil, P. *et al.* (2018) Natural selection and the predictability of evolution in *Timema* stick insects. *Science* 359, 765–777

67. Bolnick, D.I. *et al.* (2018) (Non)Parallel evolution. *Annu. Rev. Ecol. Evol. Syst.* 49, 303–330

68. Campagna, L. *et al.* (2017) Repeated divergent selection on pigmentation genes in a rapid finch radiation driven by sexual selection. *Sci. Adv.* 3, e1602404

69. Van Doren, B.M. *et al.* (2017) Correlated patterns of genetic diversity and differentiation across an avian family. *Mol. Ecol.* 26, 3982–3997

70. Delmore, K.E. *et al.* (2018) Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. *Evol. Lett.* 2, 76–87

71. Pouyet, F. *et al.* (2018) Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *eLife* 7, e36317

72. Westram, A.M. *et al.* (2018) Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. *Evol. Lett.* 2, 297–309

73. Anderson, L.K. (1999) Distribution of crossing over on mouse synaptonemal complexes using immunofluorescent localization of MLH1 protein. *Genetics* 151, 1569–1579

74. Boopathi, N.M. (2013) Linkage map construction. In *Genetic Mapping and Marker Assisted Selection: Basics, Practice and Benefits*, pp. 81–108, Springer

75. Mueller, J.C. (2004) Linkage disequilibrium for different scales and applications. *Brief. Bioinform.* 5, 355–364

76. Hinch, A.G. *et al.* (2019) Factors influencing meiotic recombination revealed by whole genome sequencing of single sperm. *Science* 363, eaau8861

77. Dreau, A. *et al.* (2018) Genome-wide recombination map construction from single individuals using linked-read sequencing. *bioRxiv* Published online December 8, 2018. http://doi.org/10.1101/489989

78. Barroso, G.V. *et al.* (2018) Inference of recombination maps from a single pair of genomes and its application to archaic samples. *bioRxiv* Published online October 25, 2018. http://doi.org/10.1101/452268

79. Stapley, J. *et al.* (2017) Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 372, 20160455