Trends in Genetics

CellPress

Review

The long and short of hyperdivergent regions

Nicolas D. Moya[1,2], Stephanie M. Yan[1,2], Rajiv C. McCoy[1,2,*], and Erik C. Andersen[1,2,*]

The increasing prevalence of genome sequencing and assembly has uncovered evidence of hyperdivergent genomic regions – loci with excess genetic diversity – in species across the tree of life. Hyperdivergent regions are often enriched for genes that mediate environmental responses, such as immunity, parasitism, and sensory perception. Especially in self-fertilizing species where the majority of the genome is homozygous, the existence of hyperdivergent regions might imply the historical action of evolutionary forces such as introgression and/or balancing selection. We anticipate that the application of new sequencing technologies, broader taxonomic sampling, and evolutionary modeling of hyperdivergent regions will provide insights into the mechanisms that generate and maintain genetic diversity within and between species.

Highlights

Sequencing of diverse *Caenorhabditis elegans* samples revealed punctuated genomic regions with excess genetic diversity, notable because most of the *C. elegans* genome exhibits low diversity caused by self-fertilization.

Hyperdivergent regions have also been documented in the genomes of humans and other mammals, such as the MHC locus, which encodes essential components of the adaptive immune system.

Recent sequencing projects have uncovered additional examples of hyperdivergent loci across the tree of life, including in *Capsella* plants, sunflowers, and parasitic nematodes.

Hyperdivergent regions are likely generated and maintained by mechanisms such as introgression from diverged lineages, hypermutability, long-term balancing selection, and/or local suppression of recombination.

More comprehensive evolutionary models are needed to determine the mechanisms that explain hyperdivergent regions.

Genomic hotspots of genetic variation

The advent of massively parallel short-read sequencing (SRS) technologies revolutionized comparison of genomes within and between species. One result of this development has been the discovery of genomic regions with extreme genetic diversity compared to the remainder of the genome [1–6]. However, discovery of such hyperdivergent regions (Box 1) is easily confounded by technical artifacts such as sequencing or genome assembly errors. In this review, we discuss hyperdivergent regions in *Caenorhabditis elegans*, *Homo sapiens*, and other organisms, focusing on their identification, prevalence across the tree of life, and possible evolutionary origins. First, we describe the discovery of hyperdivergent regions in *C. elegans*, as well as methods for calling these regions from SRS and long-read sequencing (LRS) data. Next, we address how **balancing selection** (see Glossary), **introgression**, suppressed recombination, and other mechanisms could contribute to hyperdivergent regions, reviewing examples from nematodes, plants, and humans. Last, we discuss methods for distinguishing these evolutionary mechanisms on the basis of their unique genomic signatures.

The discovery of *C. elegans* hyperdivergent regions

The free-living nematode *C. elegans* has served as a keystone model in virtually all fields of biology. *C. elegans* populations primarily comprise hermaphrodites that reproduce by **self-fertilization (selfing)** with rare **cross-fertilizing (outcrossing)** males that spontaneously arise from meiotic nondisjunction of X chromosomes. Evolutionary theory suggests that the transition to selfing and the associated reduction in effective population size lead to reductions in genetic diversity, including allelic extinction and reduced frequency of heterozygous genotypes. The resulting genome-wide homozygosity renders recombination within selfing species largely ineffective at generating new **haplotypes** [7–10]. Long blocks of linked variants are then subject to intense background selection, whereby negative selection on linked deleterious variation further reduces diversity [11,12]. As low genetic diversity limits the potential to adapt to changing environments, selfing has traditionally been viewed as an 'evolutionary dead end' [13,14].

Early population genetic studies showed that genetic diversity in *C. elegans* is lower than its outcrossing relatives *Caenorhabditis brenneri* [15] and *Caenorhabditis remanei* [16]. Historical

[1] Department of Biology, Johns Hopkins University, Baltimore, MD, USA
[2] All authors contributed equally to this article

*Correspondence:
rajiv.mccoy@jhu.edu (R.C. McCoy) and
erik.andersen@gmail.com (E.C. Andersen).

Box 1. How do we define hyperdivergent regions?

Hyperdivergent regions have been defined as punctate regions of extreme genetic diversity. In *C. elegans*, these regions were quantitatively defined by combining sequential intervals of the genome that show high densities of SNVs (>16 variants per kilobase) and gaps in coverage (0–35% coverage fraction) [3]. These thresholds were selected from a wide range of values and optimized to generate a set of hyperdivergent regions that best matched a manually curated truth set of regions defined from pairwise comparison of LRS genomes.

The punctate nature of *C. elegans* hyperdivergent regions is largely explained by the adoption of selfing as a primary mating strategy, where regions of the genome that are not protected by mechanisms that maintain genetic diversity become homogeneous over time. As a result, hyperdivergent regions are most evident in homogeneous populations where there are substantial local differences in the amplitude of variant density across their genomes. However, many of the studies described in this review involve populations with varying degrees of selfing or inbreeding, vastly different effective population sizes and structures, and other differences in their demographic histories. By consequence, a higher level of neutral genetic variation in species with high rates of outcrossing and large effective populations could minimize the signal of local deviations of polymorphism density and obfuscate the detection of hyperdivergent regions. As a result, the thresholds of SNV density and coverage used to detect hyperdivergent regions in *C. elegans* are unlikely to be informative when studying other species. This limitation motivates the development of generalizable definitions based on genomic signatures that are produced by the mechanisms that generate and maintain hyperdivergent regions.

Coalescent theory helped establish models to study the ancestry of a population, and recent advances expanded these models to repeatable simulations that can include more complex population parameters (such as recombination rate, population structure, and migration) [57]. Efficient new methods for forward simulation offer additional flexibility, such as by incorporating various modes of selection and even ecological interactions among species [58]. Such simulations can help establish null expectations of divergence among the genome sequences of a population, which can be used to identify outlier loci that violate these expectations (such as hyperdivergent regions). However, such simulations require *a priori* knowledge of demographic history, mutation, and recombination rates. Although many of these parameters can be inferred from genomic data itself, inferences can be confounded by unmodeled phenomena such as linked selection [59], especially for self-fertilizing organisms where linked selection is intensified.

Alternatively, when comparing closely related taxa, *trans*-specific polymorphism may be sufficient to define hyperdivergent regions with increased coalescence time, assuming recombination has not broken up the ancestral haplotypes. In cases where closely related sister species are available, *trans*-specific polymorphisms (and closely linked sites) may help distinguish between deeply coalescing alleles from more recent introgression and gene flow events. Extensive sampling of individual genomes within a population could also reveal hyperdivergent haplotypes associated with structural variation, which could help justify a model based on separate evolutionary trajectories by recombination suppression.

Our review provides several examples of organisms across the tree of life with hyperdivergent regions. These regions can be generated and maintained by various evolutionary mechanisms with complex and sometimes indistinguishable genomic signatures. Additionally, these mechanisms are unlikely to be isolated and can jointly contribute to the evolution of hyperdivergent regions. As a result, a definition rooted in such evolutionary mechanisms and their genomic signatures is challenging and requires further study of natural genetic variation.

episodes of strong positive selection are thought to have purged genetic variation from entire chromosomes in the worldwide *C. elegans* population [17]. Over the past few decades, the nematode research community has amassed an ever-growing collection of wild strains and genome sequences of *C. elegans* from around the world, yielding a powerful platform to further investigate natural genetic diversity of this genetic model. To date, the *Caenorhabditis* Natural Diversity Resource (CaeNDR) contains samples and genome sequences of over 1500 *C. elegans* wild strains, including strains from all continents with the exception of Antarctica, as well as numerous oceanic islands [18]. The addition of new strains in more recent population genetic studies has revealed individuals with unexpectedly high levels of genetic variation that might represent ancestral diversity predating recent selective sweeps [3,19].

Because of the low levels of genetic diversity and established homozygosity, *C. elegans* offers an ideal case to investigate the genomic distribution and patterns of variation. This variation is not evenly distributed across the *C. elegans* genome, but condensed in punctuated hyperdivergent regions that harbor extremely high levels of single-nucleotide variation (SNV) and structural variation (SV) [3]. Accurately identifying hyperdivergent regions using SRS alone is complicated by the high

Glossary

Balancing selection: an umbrella term that describes various modes of natural selection that favor maintenance of multiple alleles within a population (see Box 4 in the main text). Local genealogies of loci under long-term balancing selection might exhibit deep coalescence times. Over long timescales, recombination events that flank the causal locus break down linkage disequilibrium, causing short haplotypes.

Haplotype: a set of physically and genetically linked alleles (i.e., an entire chromosome or portion of a chromosome) that are inherited together. Linkage between the alleles on a haplotype is broken down by recombination over time.

Introgression: gene flow from one divergent lineage (e.g., species or subspecies) to another by interbreeding. Recent introgression events create long, divergent haplotypes within the recipient population, as the introgressed haplotypes have had little time to be broken down by recombination.

Outcrossing (cross-fertilization): a reproductive strategy where fertilization occurs between gametes produced by different individuals.

Selfing (self-fertilization): a reproductive strategy where hermaphroditic individuals fertilize their own gametes. Some species engage in both selfing and outcrossing. Selfing is distinct from asexual reproduction, where offspring are produced only by mitosis and are genetically identical (clonal) to their parent.

***Trans*-species polymorphism:** one possible signature of balancing selection, whereby two or more species share haplotypes or alleles by descent. This phenomenon occurs when the polymorphism originates in the ancestral population and is maintained within each lineage after species divergence.

Box 2. Variants cannot run, but they can hide

Hyperdivergent regions contain a high density of both SNVs and SVs, complicating accurate variant calling using SRS. Conventional approaches for variant discovery and genotyping based on SRS depend on initial alignment of reads to a reference genome, and reads that align with low fidelity are often discarded. As a result, the accuracy and completeness of the reference genome, as well as its genetic similarity with the target sample, impact variant calling accuracy – a problem termed 'reference bias' [60,61]. Insertions, deletions, and inversions (i.e., SVs), can also produce large genomic differences between the reference and target sample. These SVs are also difficult to discover using SRS, which fails to identify approximately 50% of deletions and 80% of insertions [62].

Several studies of hyperdivergent regions instead use an alternative approach of LRS to comprehensively assemble and compare hyperdivergent haplotypes [5,24]. This assembly approach is especially well suited for sequences that are highly divergent from the reference genome, because it enables more accurate comparison of SV- and SNV-dense sequences that would otherwise be lost among the set of unmapped reads. As LRS declines in cost, we expect that semiautomated *de novo* assembly of complete or near-complete genomes will become an increasingly accessible approach for genetic studies [63–65] and will reveal additional hyperdivergent regions across the tree of life. However, even whole-genome alignment is not perfect, and complex genomic rearrangements (e.g., nested SVs and segmental duplications [31]) and highly repetitive regions still pose a challenge for variant discovery [66,67].

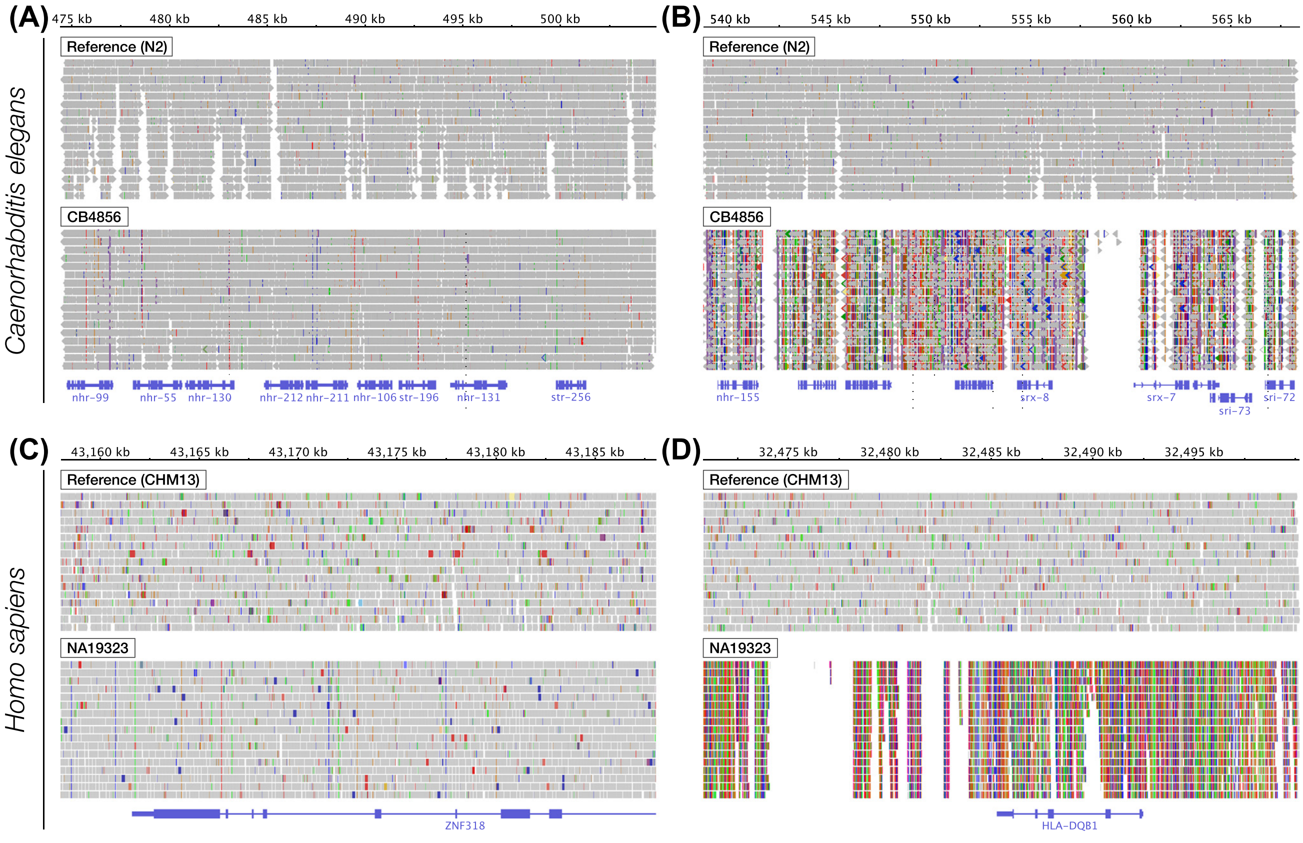
variant density and lack of homology that distinguish these sequences (Box 2). Consequently, the accurate sequences and diversity of hyperdivergent haplotypes will likely be revealed only with *de novo* assembly guided by LRS or alternative approaches that leverage high-quality assemblies such as emerging pan-genome methods (Box 3). Pairwise alignment of 15 long-read sequenced genome assemblies against the existing *C. elegans* reference genome were used to define and manually curate a set of high-confidence hyperdivergent regions. This truth set was then used to optimize the coverage and SNV count thresholds required to call hyperdivergent regions using SRS data from 328 wild strains. A total of 366 regions that spanned approximately 20% of the *C. elegans* reference genome were classified as hyperdivergent in at least one strain [3]. Within these regions, SRS alignments from diverse *C. elegans* wild strains possess SNV density that is 16.6-fold the genome-wide average – comparable with levels of sequence divergence observed between *Caenorhabditis* species. These peaks of high genetic diversity often contained large gaps in coverage that could reflect SVs (e.g., insertions, deletions, and inversions) interspersed with the SNVs (Figure 1).

Comparison of LRS genomes revealed that hyperdivergent haplotypes often contained protein-coding genes with substantial divergence in their amino acid sequences (<95% identity), as well as novel genes that were absent from the reference genome. Haplotype calling approaches showed that as many as seven distinct haplotypes could be identified in a given hyperdivergent region across only 15 LRS assemblies [3]. In addition, hyperdivergent regions were enriched for

Box 3. The future of hyperdivergent region discovery

As an alternative to whole-genome LRS, targeted LRS can be used to select for reads specific to a region of interest, reducing the cost of sequencing and assembly. The most common methods for targeted LRS involve using a primer [68] or Cas9 guide RNA [69] to enrich libraries for a specific sequence, followed by sequencing with either Oxford Nanopore Technologies (ONT) or PacBio. An alternative method uses ONT to identify and selectively eject reads in real time, thus enriching for sequencing reads of interest [70]. Targeted sequencing has been used to generate local assemblies of the IGH locus, creating a customized reference sequence that improves subsequent variant calling with short reads [24].

LRS approaches maximize structural variant discovery, with estimated recall rates of 70–90% compared with 10–70% for short reads [71,72]. However, these methods are not always feasible in terms of cost and have rarely been applied to population-scale samples. Long-read SV discovery followed by short-read genotyping, achieved via graph or other pangenome-based methods [73,74], offers a promising middle ground that can combine the resolution of long reads with the higher throughput of short reads. These methods have been used to identify novel combinations of MHC class II gene types [73], as well as selection on immunoglobulin genes [26], both of which are known to be highly polymorphic (i.e., hyperdivergent) in humans. Given sufficient long-read data, these hybrid approaches provide a mechanism that can bridge the gap between long- and short-read methods and apply them to larger sets of samples.



Trends in Genetics

Figure 1. Visualization of hyperdivergent regions with short-read sequencing. Short-read alignments from *Caenorhabditis elegans* and *Homo sapiens* individuals at hyperdivergent loci. (A) Top and bottom panels show read alignments from a normal genomic region in the N2 (Bristol, UK) and CB4856 (Hawaii, USA) *C. elegans* strains, respectively, aligned to the N2 reference genome. (B) Same as in (A), but for a hyperdivergent genomic region. The CB4856 strain shows extensive polymorphism, including multiple gaps where reads fail to align, overlapping with a series of genes encoding for G protein–coupled receptors (GPCRs), nuclear hormone receptors, and serpentine receptors. (C) Top and bottom panels show read alignments from a normal genomic region in the CHM13 and NA19323 (Luhya in Webuye, Kenya) *H. sapiens* individuals, respectively, aligned to the CHM13 reference genome. (D) Same as in (C), but for a hyperdivergent genomic region. The NA19323 sample shows extensive polymorphism, including multiple deletions, overlapping with the *HLA-DQA1* and *HHLA-DQB1* genes in the human MHC locus.

genes associated with environmental responses such as olfaction, xenobiotic stress, and pathogen defense [3]. Along with *C. elegans*, analysis of a small panel of 35 *Caenorhabditis briggsae* SRS genomes identified punctuated hyperdivergent regions that were found in shared genomic locations across the genomes of members of temperate and tropical geographic clades [3]. In parallel, genome-wide scans of nucleotide diversity in *Caenorhabditis tropicalis* also identified punctuated peaks of extremely high diversity in contrast to low levels of diversity elsewhere in the genomes [4].

The existence of hyperdivergent regions that display extreme haplotype diversity and are enriched for environmental response genes might reflect adaptations that enable *C. elegans* (and possibly other selfing *Caenorhabditis* species) to survive and respond to varying environments. Although the evolutionary origin of hyperdivergent regions in *C. elegans* is unknown, it was hypothesized that their high levels of genetic variation have been maintained by long-term balancing selection since the evolution of selfing in this species [3]. Alternatively, the authors proposed that shared local environments that harbor multiple *Caenorhabditis* species could have permitted introgression events giving rise to hyperdivergent regions. Together, these results show that regions of high genetic diversity can account for a large proportion of the genome in a species and that these hyperdivergent haplotypes might persist because they provide advantages for responding to environmental changes.

Mechanisms that generate and maintain hyperdivergent regions

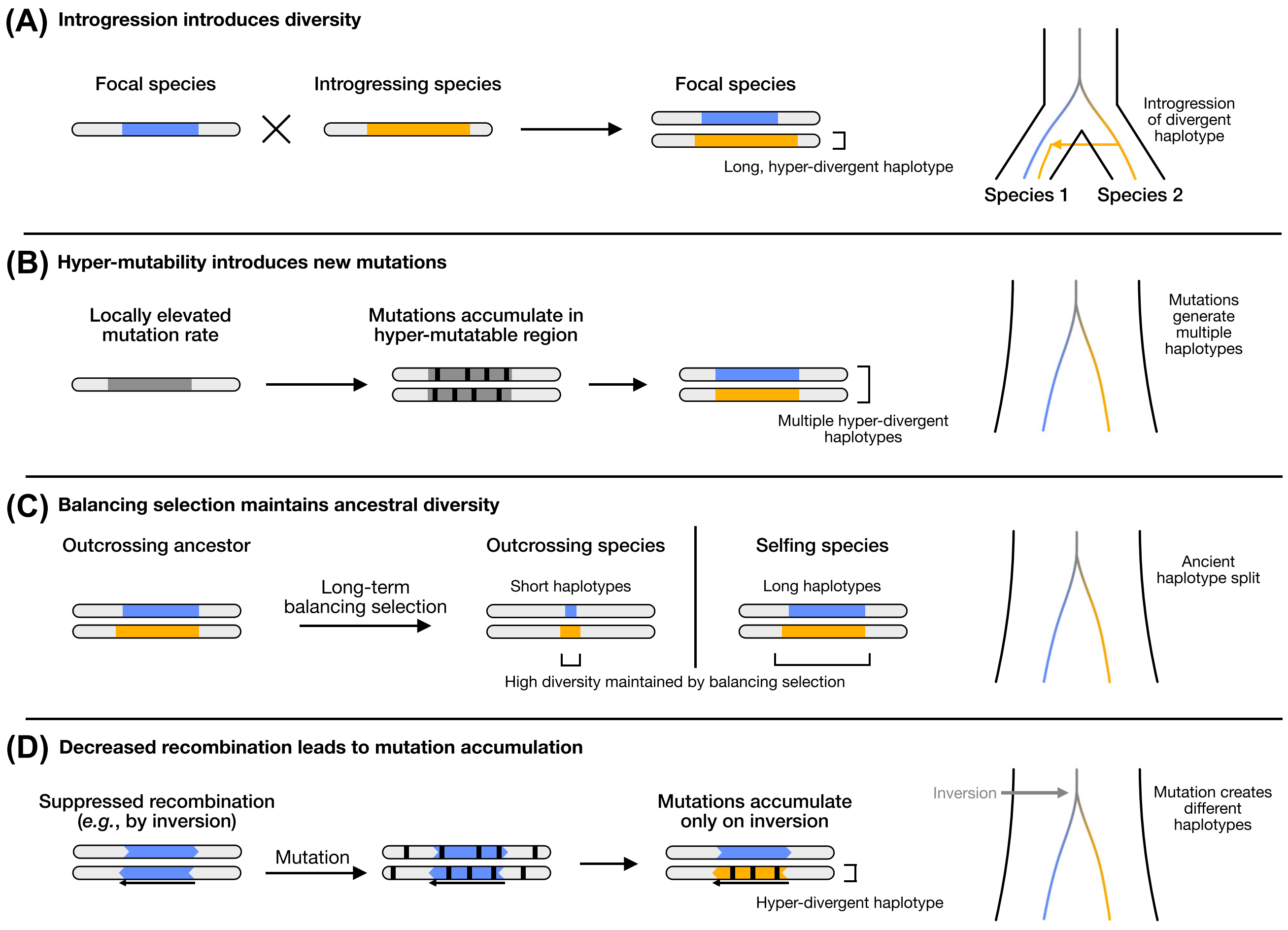
Multiple mechanisms have been proposed to explain how hyperdivergent regions are generated (e.g., via an influx of diversity from introgression or locally elevated mutation rate) and maintained (e.g., via balancing selection or suppressed recombination, which separates haplotypes and allows them to accumulate different mutations). The case studies in the following sections represent a few select species with robust evidence of hyperdivergent regions from population genomic data but are by no means comprehensive. Although we classify these examples on the basis of whether they are thought to involve introgression, hypermutability, balancing selection, or suppressed recombination, we note that these mechanisms are not mutually exclusive and that many hyperdivergent regions could be generated and maintained by a combination of factors, including some that are likely undiscovered.

An origin by introgression

The origin of hyperdivergent regions can be explained by hybridization and exchange of genetic material between divergent lineages (i.e., introgression), introducing haplotypes with high levels of nucleotide diversity into a recipient species (Figure 2A). Adaptively introgressed (i.e., beneficial) alleles would be less likely to drift out of the population, possibly aided by recurrent gene flow or recombination suppression [20].

A recent study identified 37 regions, ranging from 1 to 100 megabase pairs (Mbp) in size, with distinct population structure in three sunflower (*Helianthus*) species. Across the three species, these regions cover 4–16% of the genome and display highly divergent haplotypes with sequence identity as low as 94% in some cases (in contrast to a genome-wide average of 99.4% identity). Divergent haplotypes within these regions were associated with traits related to local adaptation, such as differences in flowering time between coastal island and inland populations of *Helianthus argophyllus* or seed size between dune and non-dune populations of *Helianthus petiolaris*. Importantly, four regions of the *H. argophyllus* genome possess haplotypes that are phylogenetically closer to *Helianthus annuus* haplotypes than they are to other haplotypes at the same locus, suggesting the haplotype diversity in these regions might have originated by introgression. The authors also showed evidence that these regions were frequently associated with SVs that suppress recombination and thus preserve linkage disequilibrium (LD) among adaptive alleles [2].

Another example of genetic divergence by introgression is ancient hybridization between modern humans and two archaic hominin groups, the Neanderthals and Denisovans [21,22]. By consequence of these introgression events, Neanderthal and Denisovan sequences comprise an average of 2–6% of the genome of contemporary non-African individuals. Archaic introgression explains some of the variation in the immunoglobulin heavy chain (IGH) locus in humans, a hyperdivergent region containing genes that contribute to the adaptive immune response. These genes undergo somatic mutation and recombination in B cells to generate epitopes for antigen recognition, producing high levels of polymorphism within individuals [23]. IGH haplotypes are highly variable between populations and harbor a high density of germline structural variants – including whole-gene deletions and insertions – likely driven by the repeat content of the IGH locus, which causes an increased rate of segmental duplication [24,25]. Yan *et al.* identified signatures of strong positive selection on a Neanderthal-introgressed haplotype, encompassing multiple IGH genes and at least two SVs, which is nearly fixed in two populations in southeast Asia but at low frequencies outside of Asia [22,26]. Although archaic introgression only accounts for one of


(A) Introgression introduces diversity
Focal species
Introgressing species
Focal species
Long, hyper-divergent haplotype
Introgression of divergent haplotype
Species 1
Species 2
(B) Hyper-mutability introduces new mutations
Locally elevated mutation rate
Mutations accumulate in hyper-mutatable region
Multiple hyper-divergent haplotypes
Mutations generate multiple haplotypes
(C) Balancing selection maintains ancestral diversity
Outcrossing ancestor
Long-term balancing selection
Outcrossing species
Short haplotypes
Selfing species
Long haplotypes
High diversity maintained by balancing selection
Ancient haplotype split
(D) Decreased recombination leads to mutation accumulation
Suppressed recombination (e.g., by inversion)
Mutation
Mutations accumulate only on inversion
Hyper-divergent haplotype
Inversion
Mutation creates different haplotypes

Trends in Genetics

Figure 2. Evolutionary mechanisms for generating and maintaining hyperdivergent regions. The right panels show simplified trees that summarize each evolutionary scenario. In all cases, haplotypes must remain separate from each other for an extended period of time in order to accumulate mutations and become distinct from each other (yellow and blue bars). This separation can occur due to the following. (A) Introgression. Recent introgression after an ancient population split creates long, divergent haplotypes in the focal species. (B) Hypermutability. A genomic region with increased mutation rate generates divergent haplotypes through mutation accumulation over time. (C) Long-term balancing selection. In an outcrossing species such as *Homo sapiens*, extended periods of balancing selection result in accumulated recombination events that create very short balanced haplotypes; in a selfing species, this signature is masked by high levels of homozygosity, resulting in the appearance of long haplotypes. (D) Decreased recombination. An inversion (bottom chromosome) suppresses recombination within the inverted sequence, causing mutations to accumulate only on the inversion haplotype (rather than being removed by recombination and drift).

the haplotypes segregating at the IGH locus, the combined evidence of somatic mutation, recombination, positive selection, and introgression in this genomic region suggests a history of recurrent diversifying selection targeting IGH genes, generating and maintaining genetic diversity within a single individual, between individuals, and between populations.

An origin by hypermutability

An elevated mutation rate in specific genomic regions, caused by features such as microsatellite or mononucleotide repeats, recombination hotspots, or DNA fragility, could also create punctate regions of excess nucleotide diversity (Figure 2B). For example, at microsatellite loci such as the trinucleotide expansions that cause Huntington's disease in humans, replication errors generate a wide range of haplotypes in the human population, ranging from 10 to 35 repeats of the trinucleotide sequence for non–disease carriers [27]. Mutation 'hotspots' have also been identified as drivers of adaptation in stickleback fish [28,29] and bacteria [30], although most such examples involve directional selection where one haplotype sweeps to fixation within a population. Finally, LRS of human genomes has recently revealed that nonallelic homologous recombination between segmental duplications or other repeats can induce recurrent inversions, contributing to the formation of hyperdivergent loci [31].

Maintenance by long-term balancing selection

The *C. elegans* examples that we discussed in the first section, as well as several other examples in the literature, invoke balancing selection as a mechanism for maintaining hyperdivergent haplotypes (Figure 2C). Balancing selection preserves genetic variation at loci where the fixation of one allele leads to a reduction in fitness, and several modes of balancing selection have been described (Box 4) [32]. One potential signature of long-term balancing selection is ***trans*-species polymorphism**, where ancestral variation is maintained within each of two or more lineages after their divergence from a common ancestor [33].

For example, *Capsella rubella*, a selfing plant thought to have recently diverged from its outcrossing sister species *Capsella grandiflora* (estimated split 170 000 generations ago), possesses numerous *trans*-species polymorphisms. Variant calling within natural populations of these two species revealed that over half of segregating variants in *C. rubella* (>800 000 SNPs) were shared with *C. grandiflora*. Regions with high density of *trans*-species SNPs were correlated with regions of high genetic diversity in *C. rubella*, and immune response loci [such as *Arabidopsis thaliana* nucleotide-binding leucine-rich repeat (NLR) homologs] were highly enriched for *trans*-species SNPs. Importantly, when incorporating a more distant selfing lineage, *Capsella orientalis* (estimated split from *C. grandiflora* over 1.8 million generations ago), thousands of polymorphic sites shared among the three species were identified. Such magnitude of shared polymorphism between distant taxa is highly unlikely under neutral evolution, suggesting that ancient alleles have likely been maintained since the evolution of selfing in *Capsella* [1].

Depending on the timescale, history of recombination, and architecture of causal mutations, *trans*-species polymorphisms might extend beyond individual SNPs to longer shared haplotypes. In outcrossing parasitic nematodes of the *Heligmosomoides* genus, comparisons of genome sequences of *Heligmosomoides bakeri* individuals revealed hundreds of punctuated hyperdivergent regions with high SNP density. These regions span 9.6% of the reference genome and are enriched for genes associated with parasitism, suggesting that recurrent selection to evade the host immune

Box 4. Modes of balancing selection

Balancing selection acts to preserve genetic variation at loci where the fixation of one allele leads to a reduction in fitness [32]. Several modes of balancing selection have been described, including:

- Heterozygote advantage, also called overdominance, where heterozygotes possess higher fitness than individuals with either homozygous genotype [75]. A classic example of heterozygote advantage in humans is the hemoglobin beta gene, where a sickle cell allele confers malaria resistance in heterozygotes but causes severe disease in homozygotes [75,76]. Heterozygote advantage is expected to maintain two balanced alleles at a locus.
- Negative frequency-dependent selection, where rare alleles possess a fitness advantage. Negative frequency-dependent selection has been observed in host–pathogen dynamics where pathogens are more likely to infect hosts that carry common immune alleles [77]. Because hosts with rare alleles are less likely to be infected, these alleles will increase in frequency over time [78]. This mode of selection can cause multiple distinct alleles to persist in a population.
- Spatially or temporally varying selection, also called fluctuating selection, where fitness of an allele depends on the environment, season, or time period. Local adaptation is a common form of spatially varying selection that maintains genetic variation at the species level while reducing genetic diversity at the population level. In addition, temporally varying selection has been observed in some wild populations of *Drosophila melanogaster*, where the frequency of ancient balanced alleles in the population fluctuates predictably with season or weather [79–81]. Like negative frequency-dependent selection, these modes of balancing selection can maintain multiple distinct alleles at each selected locus.

system has produced high levels of genetic diversity in these regions. Phylogenetic trees built from the protein sequences of genes within *H. bakeri* hyperdivergent regions and their orthologs in *Heligmosomoides polygyrus* revealed hundreds of cases – nearly half of all genes within hyperdivergent regions – where *H. bakeri* and *H. polygyrus* did not form separate clades. Importantly, the synonymous site divergence of genes within shared haplotypes did not differ significantly from the average genome-wide divergence, which suggests that shared haplotypes did not originate from recent gene flow. The prevalent haplotype sharing with *H. polygyrus* indicates that *H. bakeri* hyperdivergent regions likely represent ancient genetic diversity that has been maintained since the last common ancestor of these two species.

Within vertebrates, genome-wide scans in humans and chimpanzees have identified more than 100 loci with evidence of *trans*-species polymorphisms, enriched for genes involved in host–pathogen interactions [34]. The best-characterized examples of *trans*-species polymorphisms lie in the MHC locus, which encodes antigen-presenting proteins in T cells and is an essential component of the adaptive immune system [35]. When MHC proteins were first identified in the 1900s, it was discovered that the levels of divergence between human MHC alleles rivals the divergence observed between species for other proteins [36,37]. Recent analyses support high heterozygosity and deep coalescence structure at classic MHC genes, including the most extreme case of *HLA-DQB1*, where polymorphism has been maintained for more than 45 million years, predating the divergence of humans and New World monkeys (Figure 1B) [38]. In addition to SNVs, sequencing and assembly of MHC haplotypes across vertebrate species revealed both fixed and polymorphic gene duplications of up to 193 copies in some organisms [29–31]. The evolutionary origins of this hyperdivergent region have been attributed to diverse modes of balancing selection, including heterozygote advantage, negative frequency-dependent selection, and spatially varying selection (Box 4) [39].

Finally, we note that selfish maternal [4] or paternal [40] genetic elements can maintain hyperdivergent regions – likely by balancing selection, although it is uncertain which mode of balancing selection best explains this phenomenon. Noble *et al.* [4] and Rockman [41] suggested that hyperdivergent regions in *C. tropicalis* persist because of maternal Medea elements, which kill offspring that are homozygous for a non-Medea haplotype. In this model, Medea alleles arise on existing haplotypes and subsequently force the Medea haplotype to fixation within a selfing population. The fixation of different haplotypes in different populations (i.e., spatially varying selection), coupled with strong LD in selfing species that links a large region of the chromosome to the Medea element, creates the appearance of hyperdivergence at the species-wide level [41].

Maintenance by suppressed recombination

Suppression of recombination, often through chromosomal rearrangements, is another mechanism by which hyperdivergent regions can be maintained. For example, an inversion can generate one haplotype that does not recombine within a population or mutations that impact the recombination rate (e.g., by disrupting recombination hotspots) can limit recombination locally [42].

Although initially neutral, an inverted haplotype that captures locally adapted alleles can quickly establish in a population by preventing the formation of maladaptive recombinants [43]. As a result, the evolutionary trajectories of alternative chromosomal arrangements are split, allowing the haplotypes to accumulate mutations separately and to cause hyperdivergent haplotypes (Figure 2D). However, the reduction in effective population size for each arrangement can limit the effectiveness of purifying selection, and deleterious mutation accumulation can cause degeneration of either haplotype and loss of beneficial alleles [44,45]. Supergenes – tightly linked alleles with the pattern of inheritance of a single Mendelian locus, often associated with inversion polymorphism – can potentially escape this fate. Berdan *et al.* proposed a model where the accumulation of mutations in supergene arrangements can also lead to the emergence of alleles that are lethal when homozygous. Individuals that possess both arrangements show higher fitness, and the supergene polymorphism is maintained through overdominance (Box 4), balancing the lethal allele (i.e., associative overdominance) [46].

Another example of hyperdivergence caused by suppressed recombination is the evolution of heteromorphic chromosomes (i.e., chromosomes that are different in size or content but still pair during meiosis – most notably, sex chromosomes such as the X and Y in humans or the Z and W in birds) [47]. Suppression of recombination between sex chromosomes is advantageous because it forces separation of sex-determining genes onto different haplotypes, which can then specialize by accumulating sex-specific genes or mutations [48]. Selection for a suppressed recombination rate is thought to cause mutation accumulation and sequence divergence, creating two genetically distinct haplotypes. Eventually, this lack of recombination tends to have a deleterious effect, leading to the accumulation of transposable elements and ampliconic sequences, followed by degeneration and the loss of functional genes, as on the human Y chromosome [49].

Distinguishing evolutionary mechanisms for generating hyperdivergent regions

Determining the evolutionary mechanisms that generated a hyperdivergent region can be challenging because balancing selection, introgression, and suppressed recombination leave similar genomic signatures (or might even be indistinguishable) (Figure 2). Inversions that suppress recombination are likely the most straightforward mechanism to identify, because several methods exist for detecting these SVs from high-throughput sequencing or assembled haplotypes [2].

A common approach for identifying evidence of introgression is the *D* statistic (also called the ABBA-BABA test) and related statistics [50]; the ABBA-BABA test was originally developed to identify Neanderthal-introgressed variants in human populations [51]. This statistic, which requires sequencing data from the introgressing species, an outgroup, and two populations in the recipient species (one of which has not experienced introgression), reveals introgression based on an excess of allele sharing between the introgressing species and one recipient species population. Because balancing selection in particular lineages could also produce an excess of allele sharing in theory, it is unclear whether this statistic would be appropriate for distinguishing introgression from balancing selection.

Introgression and balancing selection can sometimes be differentiated using haplotype length: balanced alleles are maintained in the population over long periods of time, causing very short haplotypes due to the accumulation of recombination events across many generations (Figure 2C). By contrast, recently introgressed haplotypes have not yet been broken down by recombination and should exhibit long tracts of LD (Figure 2A). This haplotype length decay has been used to pinpoint the timing of archaic introgression into human populations [52]. However, one complication in species such as *C. elegans* is that excess haplotype length might not be unique to introgressed sequences, because LD is extensive throughout the genome because of reproduction by selfing (Figure 2C) [53,54].

As an alternative to LD-based approaches, Koenig *et al.* [1] and Stevens *et al.* [5] distinguished balancing selection from introgression by comparing hyperdivergent haplotypes between two species that shared *trans*-species polymorphisms. If these haplotypes were ancestral (i.e., maintained by ancient balancing selection), hyperdivergent haplotypes in the two species should be as different as normal regions of the genome. If they were inherited by recent introgression from one species, the hyperdivergent haplotypes should be more closely related than other regions of the genome [5].

Several tools exist for determining whether balancing selection, introgression, or suppressed recombination generated a hyperdivergent region. These methods include direct discovery of inversions or introgressed alleles from sequencing data, contrasting the expected lengths of haplotypes under introgression and balancing selection, and comparing divergence within and outside of hyperdivergent regions. Simulations are another fairly underexplored method that could provide additional understanding of how evolutionary phenomena alter genetic and haplotypic diversity in selfing organisms [41,55,56]. Further development of evolutionary simulations, as well as corresponding population genetic statistics that predict how these mechanisms should behave under different scenarios, could help identify additional genomic signatures that distinguish introgression, balancing selection, and suppressed recombination, providing valuable insights into the evolutionary origins of hyperdivergent regions.

Concluding remarks

The past two decades of genome sequencing, combined with the assembly of high-quality reference genomes for many species, has revealed that organisms across the tree of life carry hyperdivergent regions with extreme levels of genetic variation. This observation of excess genetic diversity has been attributed to introgression, hypermutability, historical balancing selection, suppressed recombination, or a combination of multiple mechanisms. However, confidently distinguishing these mechanisms is challenging, especially in selfing species, where common population genetic expectations about LD and genotype ratios are violated (see Outstanding questions).

As LRS becomes more widespread and is applied across additional species, we believe that hyperdivergent regions will be revealed to be more common than is currently appreciated. We anticipate that in coming years, the combination of additional genomes and novel population genetic methods will transform our ability to study hyperdivergent regions, advancing understanding of how genetic diversity is maintained across the tree of life.

Outstanding questions

As genomes of more species are sequenced and assembled, will hyperdivergent regions be ubiquitous or isolated to specific taxa?

What percentage of hyperdivergent regions are generated by introgression, hypermutability, balancing selection, and suppressed recombination? Are there additional undiscovered mechanisms?

What could explain the differences in the proportion of the genome that is impacted by hyperdivergent regions in different species?

Can evolutionary simulations accurately distinguish signatures of balancing selection and introgression, especially within selfing species?

What mechanisms besides archaic introgression and local adaptation drive the germline hyperdivergence of the IGH locus in humans? Does the elevated recombination rate in this region lead to higher rates of germline mutation?

To what extent does intragenomic conflict, such as the selfish Medea or peel elements reported in *Caenorhabditis*, contribute to the maintenance of hyperdivergent regions?

Aside from inversions, what types of mutations (e.g., destruction of recombination hotspots) can suppress recombination?

How does variation and evolution of mutation and recombination contribute to patterns of hyperdivergence? For example, can a locally reduced recombination rate cause haplotype-specific mutation accumulation?

Acknowledgments

E.C.A. was funded by a Human Frontiers Science Program (RGP0001/2019) award and a National Science Foundation (NSF) CAREER award. N.D.M. was supported in part by a National Institutes of Health (NIH) Training Grant (T32 GM008449) through Northwestern University's Biotechnology Training Program. S.M.Y. is supported by NIH National Human Genome Research Institute (NHGRI) F31HG012495. R.C.M. and S.M.Y. are supported by NIH National Institute of General Medical Sciences (NIGMS) R35GM133747 and NIH OT2OD034190. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Declaration of interests

The authors have no conflicts of interest to declare.

References

1. Koenig, D. *et al.* (2019) Long-term balancing selection drives evolution of immunity genes in *Capsella*. *Elife* 8, e43606
2. Todesco, M. *et al.* (2020) Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* 584, 602–607
3. Lee, D. *et al.* (2021) Balancing selection maintains hyperdivergent haplotypes in *Caenorhabditis elegans*. *Nat. Ecol. Evol.* 5, 794–807
4. Noble, L.M. *et al.* (2021) Selfing is the safest sex for *Caenorhabditis tropicalis*. *Elife* 10, e62587
5. Stevens, L. *et al.* (2023) Ancient diversity in host-parasite interaction genes in a model parasitic nematode. *Nat. Commun.* 14, 7776
6. Cole, R. *et al.* (2023) The parasitic nematode *Strongyloides ratti* exists predominantly as populations of long-lived asexual lineages. *Nat. Commun.* 14, 6427
7. Pollak, E. (1987) On the theory of partially inbreeding finite populations. I. Partial selfing. *Genetics* 117, 353–360
8. Charlesworth, D. and Charlesworth, B. (1995) Quantitative genetics in plants: the effect of the breeding system on genetic variability. *Evolution* 49, 911–920
9. Nordborg, M. (2000) Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. *Genetics* 154, 923–929
10. Charlesworth, D. and Wright, S.I. (2001) Breeding systems and genome evolution. *Curr. Opin. Genet. Dev.* 11, 685–690

11. Kaplan, N.L. *et al.* (1989) The 'hitchhiking effect' revisited. *Genetics* 123, 887–899
12. Cutter, A.D. and Payseur, B.A. (2003) Selection at linked sites in the partial selfer *Caenorhabditis elegans*. *Mol. Biol. Evol.* 20, 665–673
13. Stebbins, G.L. (1957) Self fertilization and population variability in the higher plants. *Am. Nat.* 91, 337–354
14. Igic, B. and Busch, J.W. (2013) Is self-fertilization an evolutionary dead end? *New Phytol.* 198, 386–397
15. Dey, A. *et al.* (2013) Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11056–11060
16. Teterina, A.A. *et al.* (2023) Genomic diversity landscapes in outcrossing and selfing *Caenorhabditis* nematodes. *PLoS Genet.* 19, e1010879
17. Andersen, E.C. *et al.* (2012) Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nat. Genet.* 44, 285–290
18. Crombie, T.A. *et al.* (2024) CaeNDR, the *Caenorhabditis* natural diversity resource. *Nucleic Acids Res.* 52, D850–D858
19. Crombie, T.A. *et al.* (2019) Deep sampling of Hawaiian *Caenorhabditis elegans* reveals high genetic diversity and admixture with global populations. *Elife* 8, e50465
20. Zhang, W. *et al.* (2016) Genome-wide introgression among distantly related *Heliconius* butterfly species. *Genome Biol.* 17, 25
21. Nielsen, R. *et al.* (2017) Tracing the peopling of the world through genomics. *Nature* 541, 302–310
22. Browning, S.R. *et al.* (2018) Analysis of human sequence data reveals two pulses of archaic denisovan admixture. *Cell* 173, 53–61.e9
23. Watson, C.T. *et al.* (2017) The individual and population genetics of antibody immunity. *Trends Immunol.* 38, 459–470
24. Rodriguez, O.L. *et al.* (2023) Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.* 14, 4419
25. Pramanik, S. *et al.* (2011) Segmental duplication as one of the driving forces underlying the diversity of the human immunoglobulin heavy chain variable gene region. *BMC Genomics* 12, 78
26. Yan, S.M. *et al.* (2021) Local adaptation and archaic introgression shape global diversity at human structural variant loci. *Elife* 10, e67615
27. Nesta, A.V. *et al.* (2021) Hotspots of human mutation. *Trends Genet.* 37, 717–729
28. Chan, Y.F. *et al.* (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* 327, 302–305
29. Xie, K.T. *et al.* (2019) DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science* 363, 81–84
30. Moxon, R. *et al.* (2006) Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* 40, 307–333
31. Porubsky, D. *et al.* (2022) Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. *Cell* 185, 1986–2005.e26
32. Charlesworth, D. (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2, e64
33. Klein, J. *et al.* (1998) Molecular trans-species polymorphism. *Annu. Rev. Ecol. Evol. Syst.* 29, 1–21
34. Leffler, E.M. *et al.* (2013) Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science* 339, 1578–1582
35. Klein, J. *et al.* (1993) The molecular descent of the major histocompatibility complex. *Annu. Rev. Immunol.* 11, 269–295
36. Klein, J. (1987) Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. *Hum. Immunol.* 19, 155–162
37. Raymond, C.K. *et al.* (2005) Ancient haplotypes of the HLA class II region. *Genome Res.* 15, 1250–1257
38. Fortier, A.L. and Pritchard, J.K. (2022) Ancient trans-species polymorphism at the major histocompatibility complex in primates. *bioRxiv*, Published online September 17, 2024. https://doi.org/10.1101/2022.06.28.497781
39. Takahata, N. and Nei, M. (1990) Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics* 124, 967–978
40. Long, L. *et al.* (2023) A toxin-antidote selfish element increases fitness of its host. *Elife* 12, e81640
41. Rockman, M.V. (2024) Parental-effect gene-drive elements under partial selfing, or why do *Caenorhabditis* genomes have hyperdivergent regions? *Genetics* 30, iyae175
42. Jay, P. *et al.* (2022) Sheltering of deleterious mutations explains the stepwise extension of recombination suppression on sex chromosomes and other supergenes. *PLoS Biol.* 20, e3001698
43. Kirkpatrick, M. and Barton, N. (2006) Chromosome inversions, local adaptation and speciation. *Genetics* 173, 419–434
44. Stevison, L.S. *et al.* (2011) Effects of inversions on within- and between-species recombination and divergence. *Genome Biol. Evol.* 3, 830–841
45. Berdan, E.L. *et al.* (2021) Deleterious mutation accumulation and the long-term fate of chromosomal inversions. *PLoS Genet.* 17, e1009411
46. Berdan, E.L. *et al.* (2022) Mutation accumulation opposes polymorphism: supergenes and the curious case of balanced lethals. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 377, 20210199
47. Checchi, P.M. and Engebrecht, J. (2011) Heteromorphic sex chromosomes: navigating meiosis without a homologous partner. *Mol. Reprod. Dev.* 78, 623–632
48. Charlesworth, D. *et al.* (2005) Steps in the evolution of heteromorphic sex chromosomes. *Heredity* 95, 118–128
49. Bachtrog, D. (2013) Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat. Rev. Genet.* 14, 113–124
50. Patterson, N. *et al.* (2012) Ancient admixture in human history. *Genetics* 192, 1065
51. Green, R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328, 710
52. Sankararaman, S. *et al.* (2016) The combined landscape of Denisovan and Neanderthal ancestry in present-day humans. *Curr. Biol.* 26, 1241–1247
53. Cutter, A.D. (2006) Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics* 172, 171–184
54. Burgarella, C. and Glémin, S. (2017) Population genetics and genome evolution of selfing species. *eLS*, Published online January 16, 2017. https://doi.org/10.1002/9780470015902.a0026804
55. Hartfield, M. and Bataillon, T. (2020) Selective sweeps under dominance and inbreeding. *G3* 10, 1063–1075
56. Smith, M.L. and Hahn, M.W. (2024) Selection leads to false inferences of introgression using popular methods. *Genetics* 227, iyae089
57. Baumdicker, F. *et al.* (2022) Efficient ancestry and mutation simulation with msprime 1.0. *Genetics* 220, iyab229
58. Haller, B.C. and Messer, P.W. (2023) SLiM 4: Multispecies eco-evolutionary modeling. *Am. Nat.* 201, E127–E139
59. Schrider, D.R. *et al.* (2016) Effects of linked selective sweeps on demographic inference and model selection. *Genetics* 204, 1207–1223
60. Chen, N.-C. *et al.* (2021) Reference flow: reducing reference bias using multiple population genomes. *Genome Biol.* 22, 8
61. Sirén, J. *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* 374, abg8871
62. Chaisson, M.J.P. *et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* 10, 1784
63. Liao, W.-W. *et al.* (2023) A draft human pangenome reference. *Nature* 617, 312–324
64. Nurk, S. *et al.* (2022) The complete sequence of a human genome. *Science* 376, 44–53
65. Makova, K.D. *et al.* (2024) The complete sequence and comparative analysis of ape sex chromosomes. *Nature* 630, 401–411
66. Ebert, P. *et al.* (2021) Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 372, eabf7117
67. Jain, C. *et al.* (2022) Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat. Methods* 19, 705–710
68. Rodriguez, O.L. *et al.* (2020) A novel framework for characterizing genomic haplotype diversity in the human immunoglobulin heavy chain locus. *Front. Immunol.* 11, 2136
69. Gilpatrick, T. *et al.* (2020) Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* 38, 433–438

70. Kovaka, S. *et al.* (2021) Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nat. Biotechnol.* 39, 431–441
71. Smolka, M. *et al.* (2024) Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* 42, 1571–1580
72. Sedlazeck, F.J. *et al.* (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* 15, 461–468
73. Chin, C.-S. *et al.* (2023) Multiscale analysis of pangenomes enables improved representation of genomic diversity for repetitive and clinically relevant genes. *Nat. Methods* 20, 1213–1221
74. Chen, S. *et al.* (2019) Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* 20, 291
75. Sellis, D. *et al.* (2011) Heterozygote advantage as a natural consequence of adaptation in diploids. *Proc. Natl. Acad. Sci. U. S. A.* 108, 20666–20671
76. Allison, A.C. (1954) Protection afforded by sickle-cell trait against subtertian malareal infection. *Br. Med. J.* 1, 290–294
77. Christie, M.R. and McNickle, G.G. (2023) Negative frequency dependent selection unites ecology and evolution. *Ecol. Evol.* 13, e10327
78. Koskella, B. and Lively, C.M. (2009) Evidence for negative frequency-dependent selection during experimental coevolution of a freshwater snail and a sterilizing trematode. *Evolution* 63, 2213–2221
79. Bergland, A.O. *et al.* (2014) Genomic evidence of rapid and stable adaptive oscillations over seasonal time scales in Drosophila. *PLoS Genet.* 10, e1004775
80. Johnson, O.L. *et al.* (2023) Fluctuating selection and the determinants of genetic variation. *Trends Genet.* 39, 491–504
81. Machado, H.E. *et al.* (2021) Broad geographic sampling reveals the shared basis and environmental correlates of seasonal adaptation in *Drosophila*. *Elife* 10, e67577