MOLECULAR ECOLOGY

Molecular Ecology (2013) **22**, 2074–2095 doi: 10.1111/mec.12281

INVITED REVIEW AND META-ANALYSES

Molecular hyperdiversity and evolution in very large populations

ASHER D. CUTTER,*† RICHARD JOVELIN* and ALIVIA DEY*

Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada M5S 3B2, †Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario, Canada M5S 3B2

Abstract

The genomic density of sequence polymorphisms critically affects the sensitivity of inferences about ongoing sequence evolution, function and demographic history. Most animal and plant genomes have relatively low densities of polymorphisms, but some species are hyperdiverse with neutral nucleotide heterozygosity exceeding 5%. Eukaryotes with extremely large populations, mimicking bacterial and viral populations, present novel opportunities for studying molecular evolution in sexually reproducing taxa with complex development. In particular, hyperdiverse species can help answer controversial questions about the evolution of genome complexity, the limits of natural selection, modes of adaptation and subtleties of the mutation process. However, such systems have some inherent complications and here we identify topics in need of theoretical developments. Close relatives of the model organisms *Caenorhabditis elegans* and *Drosophila melanogaster* provide known examples of hyperdiverse eukaryotes, encouraging functional dissection of resulting molecular evolutionary patterns. We recommend how best to exploit hyperdiverse populations for analysis, for example, in quantifying the impact of noncrossover recombination in genomes and for determining the identity and micro-evolutionary selective pressures on noncoding regulatory elements.

Keywords: adaptation, nucleotide polymorphism, population genetics, population size, speciation

Received 26 October 2012; revision received 24 January 2013; accepted 29 January 2013

Introduction

A central aim of biology is to fully describe the processes that promote biodiversity, both in terms of taxonomic abundance and in terms of phenotypic and molecular variation within lineages. Explaining disparities amongst species in molecular diversity forms a long-standing problem in evolutionary genetics (Lewontin 1974; Leffler *et al.* 2012). As a result, population genetics and molecular evolutionary theory developed to explain such molecular diversity is now also a cornerstone in genome-wide trait association mapping, infection dynamics, molecular ecology and conservation genetics (Frankham 1995a; Milgroom *et al.* 1997; Knowles & Maddison 2002; Boyko *et al.* 2008; Ioannidis *et al.* 2009). However, the theoretical underpinnings of current applications of theory to empirical data often rely on assumptions that are violated in a growing number of systems (Desai & Plotkin 2008; Sella *et al.* 2009; Ninio 2010; Pritchard *et al.* 2010; Cutter *et al.* 2012). One such characteristic—hyperdiversity of molecular genetic variation within populations—traditionally has been regarded as relevant only to microbes and viruses. Yet, emerging data demonstrate that multicellular eukaryotic organisms also can harbour exceptional levels of molecular polymorphism, including close relatives to some of the classic model organisms *Caenorhabditis elegans* and *Drosophila melanogaster*.

Here, we consider the implications of hyperdiversity in eukaryotes for advancing our understanding of the evolutionary processes that govern the origin and maintenance of diversity within populations. Hyperdiverse eukaryotes can inform a broad range of issues in evolutionary genetics that are challenging or impossible to explore in traditional study systems. We show how investigating hyperdiverse taxa can help clarify such problems as the origins and evolution of genome complexity,

Correspondence: Asher D. Cutter, Fax: 416 978 5878;
E-mail: asher.cutter@utoronto.ca

© 2013 Blackwell Publishing Ltd

modes of adaptive evolution, the identification of functional regulatory elements, mutational dynamics and the dissection of the limits of natural selection. To do so, an explicit accounting of the complications relative to standard models of molecular evolution is the key to successfully exploiting hyperdiverse organisms.

Hyperdiverse eukaryotes in nature

What is hyperdiversity?

We consider populations to be hyperdiverse if, at the level of nucleotide sequence, a random pair of alleles differ on average by 5% or more of sites in genomic regions that putatively evolve without constraint (θ_{neu}). This 5% rule-of-thumb is motivated by the practical relevance of finite- versus infinite-sites models of mutation above this level of nucleotide polymorphism (Desai & Plotkin 2008). Historically, eukaryotes have been thought not to harbour such extensive genetic variation, but, as we document below, this view is changing. While many species comprised of subdivided populations might achieve species-wide hyperdiversity, it is the diversity within populations that concerns us here. By focusing on within-population variation, we clearly distinguish polymorphism from divergence, and target the measure of diversity that is most appropriate for comparing amongst species (Charlesworth 2009). In what follows, we focus on the situation in which large population size and high θ_{neu} are coincident, so that, presumably, long-term and short-term estimates of effective population size (N_e; Box 1) will be concordant. For our purposes, it is sufficient that θ_{neu} is very large and that N_e also is very large. We are interested in how extremely large N_e in eukaryotes can manifest as molecular hyperdiversity, how the presence of hyperdiversity can compromise conventional thought and analyses of population genetic phenomena, and how to exploit hyperdiversity to gain a deeper understanding of the evolutionary process. Some implications of these issues also are pertinent to species with high θ_{neu} induced by a high mutation rate, despite modest N_e, whereas other implications will also be relevant to species with high short-term population size despite only modest θ_{neu}.

Box 1 Timescales of N_e

A classic prediction of Kimura's (1968) neutral theory of molecular evolution is that the amount of standing genetic variation at sites unaffected by selection will equal the product of four times the neutral mutation rate and the effective population size (i.e., at equilibrium $\theta_{neu} = 4N_e\mu$) (Tajima 1983). Consequently, we might expect diversity in genomic regions presumed to be unconstrained to differ between species as a direct consequence of differences in mutation rate (μ) and/or effective population size (N_e). The effective population size is that central concept in population genetics introduced by Wright (1933) that describes the strength of genetic drift in populations (Charlesworth 2009).

Ongoing evolution in some species that have large present-day 'short-term' population sizes might not reflect the predictions derived from smaller 'long-term' population size estimates inferred from observed θ_{neu} (Karasov *et al.* 2010; Weissman & Barton 2012). Short-term effective population size, the most intuitive way of thinking about how N_e relates to genetic drift, is best inferred from direct measures of allele frequency change across recent generations for putatively neutral markers (variance effective size, $N_{e(V)}$) or by heterozygosity excess ($N_{e(H)}$). Singleton variants in very large population samples provide another means of estimating recent N_e from new mutational input (Keinan & Clark 2012). These measures indicate the *current* influence of genetic drift.

By contrast, segregating genetic variation essentially reflects the harmonic mean size over the population's entire coalescent history ($N_{e(\theta)}$) and is therefore very sensitive to any bottlenecks that may have occurred over this long span of time (Kimura & Crow 1963; Drummond *et al.* 2002; Wang 2005). This can potentially create appearance of a disconnect between a small coalescent effective size estimated from genetic diversity despite the small role of drift in a large present-day population. For some purposes, however, even $N_{e(\theta)}$ may not be long-term enough—for example, in deducing selection efficacy in the interval between the coalescent of population variation and the more ancient most recent common ancestor with other species. It is possible, however, to infer ancestral population sizes from the distribution of sequence divergence and gene-tree species-tree discordance (Rannala & Yang 2003; Zhou *et al.* 2007). Linkage disequilibrium-based estimators of N_e should provide a view of N_e intermediate between the long-term view afforded by $N_{e(\theta)}$ and very recent estimates of genetic drift from $N_{e(V)}$. Contrasting $N_{e(\theta)}$ values derived from molecular markers that mutate at different rates (e.g. SNPs vs. STRPs) might provide a supplementary way to assess N_e on different timescales (Payseur & Cutter 2006; Haasl & Payseur 2011). When census sizes (N) of a species can be very large and the species is subject to repeated, drastic bottlenecks or selective sweeps, then θ_{neu} implying a relatively small long-term N_e may fail to reflect the limited relevance of genetic drift to ongoing adaptation (Barton 2010; Karasov *et al.* 2010; Weissman & Barton 2012).

© 2013 Blackwell Publishing Ltd

How common is molecular hyperdiversity amongst eukaryotes?

Lynch (2006) and Leffler *et al.* (2012) surveyed a broad array of taxa for quantitative measures of nucleotide polymorphism, but the data are sparse for most of the species nominally conforming to the notion of hyperdiversity (θ_{neu} >5%). We consider cases with meagre population genetic data as preliminary (very few loci or individuals assessed), warranting further detailed confirmation. Several species, however, are convincing and merit enumeration (Fig. 1).

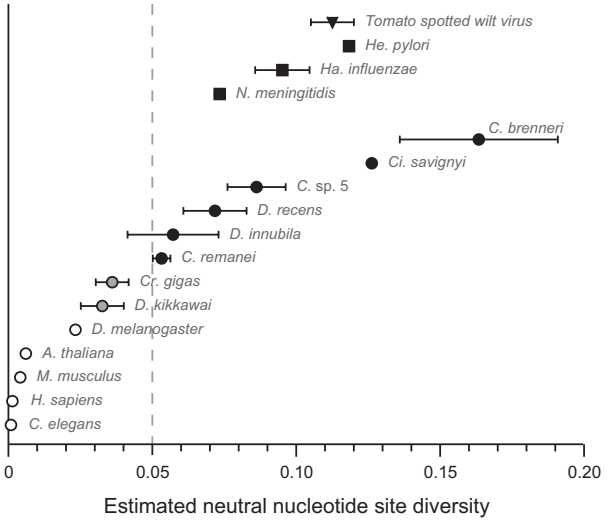


Fig. 1 Examples of hyperdiverse eukaryotes and other organisms. Estimates of neutral nucleotide diversity >5% are shown for a single population of eukaryotic species (black circles), with two species that approach this definition of hyperdiversity (light grey circles). Estimates of nucleotide diversity for model organisms (white circles) and a few examples of hyperdiverse viruses (black triangle) and bacteria (black squares) are also indicated for comparison. A Jukes-Cantor correction for multiple hits was applied to diversity values (black circles and *Haemophilus influenzae* only) and a correction for codon bias was applied to *Caenorhabditis* other than *C. elegans* and to *Ciona*. Values for *Drosophila recens* and *Drosophila innubila* derive from Watterson's θ estimator, whereas other values use average pairwise diversity (π). Data for *Caenorhabditis* and *Drosophila* include both autosomal and X-linked loci, with X-linked values unadjusted for any population sex ratio effects that would increase the values even further. Error bars indicate standard error of the mean across loci (values smaller than the symbol size not shown). References: *Arabidopsis thaliana* (Nordborg *et al.* 2005); *Caenorhabditis brenneri* (Dey *et al.* unpublished data); *C. elegans* (Andersen *et al.* 2012); *Caenorhabditis remanei* (Jovelin *et al.* 2003, 2009; Cutter *et al.* 2006a; Jovelin 2009; Dey *et al.* 2012); *C.* sp 5 (Cutter *et al.* 2012); *Ciona savignyi* (Small *et al.* 2007); *Crassostrea gigas* (Sauvage *et al.* 2007); *D. innubila* (Dyer & Jaenike 2004); *Drosophila kikkawai* (Goto *et al.* 2004); *Drosophila melanogaster* (Langley *et al.* 2012); *D. recens* (Dyer *et al.* 2007); *Ha. influenzae* (Maughan & Redfield 2009); *Helicobacter pylori* (Hughes *et al.* 2008); *Homo sapiens* (Perry *et al.* 2012); *Mus musculus* (Perry *et al.* 2012); *Neisseria meningitidis* (Hughes *et al.* 2008); TSWV (Tsompana *et al.* 2005).

First, the ascidian sea squirt *Ciona savignyi* exhibits polymorphism averaging 8% at four-fold degenerate codon sites (Small *et al.* 2007) (we calculate θ_{neu} = 0.126 after correcting for multiple hits and codon bias; Figs 1 and 2), and the related species *Ciona intestinalis* also appears to have very high diversity (Dehal *et al.* 2002; Nydam & Harrison 2010; Tsagkogeorga *et al.* 2012). These hermaphroditic marine chordates reproduce via broadcast spawning in shallow water and often achieve high densities (Jiang & Smith 2005; Therriault & Herborg 2008). Second, several species in the quinaria group of fungus-feeding *Drosophila* provide strong candidates for molecular hyperdiversity (Dyer & Jaenike 2004; Shoemaker *et al.* 2004; Dyer *et al.* 2007). Although only 5–6 nuclear loci were surveyed in each species, average silent-site diversity (θ_{si}) is ~5% in both *Drosophila innubila* and *Drosophila recens*. Some species of *Drosophila* appear near the cusp of hyperdiversity (Leffler *et al.* 2012), so tests for diversity-reducing effects of selection on synonymous sites should be considered carefully to determine θ_{neu}, as we will describe later. Third, polymorphism averages an astonishing 14% at synonymous sites (θ_{neu} = 0.16) in the nematode *Caenorhabditis brenneri* (Barriere *et al.* 2009; Jovelin 2009; Dey *et al.* unpublished data). At least two other species in this genus (*Caenorhabditis remanei*, *C.* sp. 5) also have very high diversity (Fig. 1) (reviewed in Cutter *et al.* 2009; Jovelin 2009; Wang *et al.* 2010b). Rapid decay of linkage disequilibrium (LD), gene genealogies, morphological criteria and experimental genetic crosses all support species cohesion within these groups. These ~1 mm long animals related to the classic animal model *C. elegans* are obligately outbreeding, capable of extremely rapid development and are thought to be able to disperse long distances during the stress-resistant 'dauer' phase of the life cycle (Kiontke *et al.* 2011; Lee *et al.* 2012). The nematode relatives of *C. elegans* and fly relatives of *D. melanogaster* provide particularly promising groups to study hyperdiversity, owing to the potential of porting model organism tools to these other species. *Ciona* also has a growing toolbox of experimental approaches (Lemaire 2011; Stolfi & Christiaen 2012).

We mention two final examples with extreme overall diversity, but that do not conform to our notion of within-population hyperdiversity. Although the monkeyflower *Mimulus guttatus* retains enormous polymorphism for two loci that have been examined in detail (θ_{si} ~ 7.7%), it has substantial population sub-structuring and allelic introgression from close relatives that contributes to much of the molecular diversity in this species complex (Sweigart & Willis 2003). These factors thus complicate inference of within-population effective size and hyperdiversity (Box 2). The free-living, nonparasitic protist *Paramecium aurelia* complex similarly does not have an easily interpretable evolutionary history for its high overall levels of polymorphism (Catania *et al.* 2009).

© 2013 Blackwell Publishing Ltd

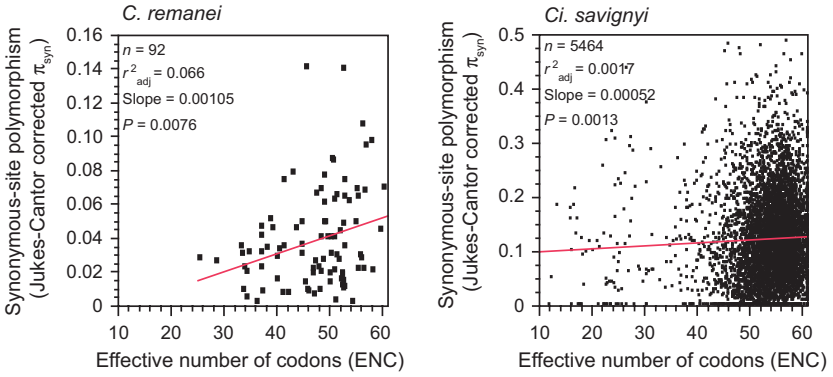


Fig. 2 Loci with stronger selection on synonymous sites for codon usage bias exhibit lower polymorphism. Data for *Caenorhabditis remanei* re-calculated from (Jovelin *et al.* 2003, 2009; Cutter *et al.* 2006a; Cutter 2008b; Jovelin 2009), plus 14 unpublished loci, using DnaSP and codonW with a manual Jukes-Cantor correction (Librado & Rozas 2009). Data for *Ciona savignyi* re-calculated from (Small *et al.* 2007), using Polymorphorama (Bachtrog & Andolfatto 2006) and codonW (14 loci with π_{syn} >0.5 excluded). Genes with stronger codon bias have two-fold more strongly depressed synonymous-site diversity in *C. remanei* than in *Ci. savignyi* (see Box 4). For *Ci. savignyi*, the estimate for θ_{neu} = 0.126 lies at the point on the regression line for an effective number of codons (ENC) value of 61, which corresponds to uniform codon usage.

Box 2 Ecological determinants of hyperdiversity

What factors contribute to a species being hyperdiverse? Organisms that capitalize on dense or widespread resources should themselves have high population abundances, with correspondingly greater N_e and genetic variation (Stevens *et al.* 2007). It is physically small organisms that can more plausibly achieve and maintain exceptionally large populations (Peters & Wassenberg 1983; Marquet *et al.* 2005). Moreover, small-bodied organisms tend to have more rapid turnover of generations (Blueweiss *et al.* 1978), and species with short generation times have larger population sizes (Chao & Carr 1993). Very small organisms (<1 mm) are also thought to have enormous inherent dispersal propensity and large geographic ranges (Finlay 2002), reinforcing the importance of small body size in hyperdiversity. This is encouraging, from the practical standpoint of studying hyperdiverse taxa in detail, because small organisms tend to make more tractable systems of laboratory interrogation.

Countervailing factors, however, may depress polymorphism in species that otherwise seem good candidates for high diversity. Subpopulations with restricted genetic exchange by migration can promote high species-wide diversity, but may largely reflect divergence between them and not particularly high intra-population variation (Nordborg *et al.* 2005; Ross-Ibarra *et al.* 2008). Turnover of subpopulations via extinction and re-colonization generally reduces N_e and population variation (Pannell & Charlesworth 1999). Unidirectional dispersal with source-sink dynamics, even in an unstructured species distribution, will diminish the expected amount of diversity that can be maintained, as may occur in some marine organisms with larval stages dispersed by ocean currents (Wares & Pringle 2008; Casteleyn *et al.* 2010). Temporal fluctuations in population size or high variance in reproductive success also will constrain genetic variation below what might be expected from census size (Hedgecock 1994; Hedrick 2005). Similarly, biparental inbreeding in long-lived organisms, biased sex ratios, and self-fertilization all will act to restrict the maintenance of polymorphism. The confluence of many of these factors, compounded by a bias towards study of species with body sizes >1 mm, may contribute to only a recent recognition of hyperdiversity amongst eukaryotes.

A greater variance in reproductive output amongst individuals in species that have large census sizes, or more drastically fluctuating populations, could explain a general tendency for highly abundant organisms to have smaller N_e/N ratios (Frankham 1995b; Pray *et al.* 1996; Hedrick 2005). Although high fecundity *per se* need not result in high fecundity variance (Nunney 1996), 'sweepstakes' effects and sexual selection that diminish the number of successfully breeding adults may be important in a variety of taxa, from plants to both terrestrial and aquatic invertebrates (Frankham 1995b; Pray *et al.* 1996; Boudry *et al.* 2002; Hedrick 2005); high reproductive skew changes expectations about how genetic drift works in relation to population size (Eldon & Wakeley 2006; Der *et al.* 2012). Genetic draft will further exacerbate small N_e/N ratios for abundant species (Gillespie 2001). For species with very low N_e/N ratios, N_e estimates may provide a poor predictor of the input of beneficial mutations available for screening by natural selection (e.g. Karasov *et al.* 2010).

© 2013 Blackwell Publishing Ltd

Discovering hyperdiverse species

As the sampling of new species grows with the increasing pace of molecular approaches in ecology (Box 2), it is essential that efforts akin to DNA barcoding mature to include multiple nuclear loci, as in multi-locus sequence typing (MLST) used commonly in bacteria (Maiden *et al.* 1998; Perez-Losada *et al.* 2006), so as to appropriately interrogate the evolutionary history of organisms (Chenuil *et al.* 2010). With hyperdiverse lineages, standard thresholds of divergence—particularly for mitochondrial loci—are unlikely to be useful yardsticks of species membership (Vences *et al.* 2005; Hajibabaei *et al.* 2006). Sampling of multiple nuclear loci permits inference of recombination in the recent history of putative con-specifics (via the decline of LD) and the uniformity of genealogical patterns across loci, which are much more powerful metrics of shared ancestry in a comparative population genetic framework (Edwards 2009; Fujita *et al.* 2012). The use of multi-locus, nuclear DNA barcode or tag-sequencing integrated with high-throughput next-generation sequencing platforms has the potential to provide an unprecedented view of population variation in poorly characterized systems, in addition to capturing raw biodiversity at the level of species delimitation (Hudson 2008; Hohenlohe *et al.* 2010). We anticipate that the accelerated application of molecular ecological approaches, including metagenomics and multi-locus DNA barcoding, will drastically increase the number of hyperdiverse organisms known to science in the coming years.

Re-focussing models of mutation and mutational implications

Infinite-site vs. finite-site models

A straight-forward implication of nucleotide hyperdiversity is that a non-negligible fraction of sites will have experienced more than one mutation since the coalescent time of a population sample. This violates the key assumption of the ubiquitous infinite-sites mutation model (Kimura 1969; Tajima 1996). For example, ~10% of segregating sites observed in *C. brenneri* have three or even four nucleotide variants (i.e. tri-allelic or tetra-allelic single nucleotide polymorphisms, SNPs) (Dey *et al.* unpublished data). Thus, it becomes critical to incorporate the effects of multiple hits via finite-sites models of molecular evolution into population genetic analyses (Ninio 2010). Finite-sites models have been developed to analyse LD, variant frequency spectra and estimators of θ (Fearnhead & Donnelly 2001; McVean *et al.* 2002; Desai & Plotkin 2008; Jenkins & Song 2011), which is preferable to simply applying a *post-hoc* Jukes-Cantor multiple hits correction. However, accessible software to estimate population genetic parameters in a finite-sites framework has not yet been deployed for all such metrics. Incorporation of finite-sites models into analyses of selection for hyperdiverse taxa is particularly important to arrive at appropriate interpretations about selection and demographic history (Desai & Plotkin 2008). Even in the absence of hyperdiversity, infinite-sites mutational models may handicap tests for adaptation; a 'fitness flux' perspective based on a finite-sites notion of evolution is potentially well-suited to both hyperdiverse species and taxa with typical levels of polymorphism (Mustonen & Lassig 2009).

Complex mutation

The exceptional density of heterozygosity in hyperdiverse genomes will also help to understand complex mutational patterns. The view that mutations can induce other mutations has emerged from observations that single nucleotide mutation rates are higher in the vicinity of insertion and deletions (indels) (Hardison *et al.* 2003; Longman-Jacobsen *et al.* 2003; Wetterbom *et al.* 2006; Tian *et al.* 2008; Zhu *et al.* 2009; De & Babu 2010; Hollister *et al.* 2010; McDonald *et al.* 2011), by the clustering of SNPs (Amos 2010; Schrider *et al.* 2011a) and from the higher-than-expected occurrence of tri-allelic variants (Hodgkinson & Eyre-Walker 2010). High heterozygosity will amplify these complex mutational patterns, with potentially major consequences for genome evolution by exacerbating heterogeneity in the distribution of nucleotide variation along chromosomes. Analyses of hyperdiverse species, with their high density of indels and tri-allelic SNPs, should thus help elucidate how the mutational process works and its attendant micro-evolutionary consequences.

Compensatory mutation

Deleterious mutations can impose a serious burden on finite populations. And yet, compensatory evolution can permit populations to recover fitness relatively quickly, given sufficiently large populations (Estes & Lynch 2003; Estes *et al.* 2004), and at equilibrium even a low rate of beneficial mutation can offset the effect of detrimental mutations to maintain high population mean fitness (Goyal *et al.* 2012). Fitness recovery can derive from new beneficial compensatory alleles (Denver *et al.* 2010) or from compensatory epistatic interactions between individually deleterious mutations (Kimura 1985). However, evolutionary responses for such epistatically interacting sites is less sensitive to N_e than are independently evolving sites (Piskol & Stephan 2011). A large population size also may increase the number of mutations at any locus, and standing variation can be very high, such

© 2013 Blackwell Publishing Ltd

that the raw material for compensatory and adaptive evolution may often not be limited by mutational input in hyperdiverse organisms (Hermisson & Pennings 2005; Ralph & Coop 2010). Should high μ cause a species' hyperdiversity, then compensatory evolution will be more rapid (Piskol & Stephan 2011).

Co-evolution between pairs of interacting amino acid residues is a good predictor of protein structure and has successfully predicted protein folds *de novo* (Marks *et al.* 2011; Hopf *et al.* 2012). This method stems from the principle that conservation of protein structure is essential to maintain proper function, and that this imposes selection for conservation of physical interactions between amino acid residues. Thus, efficient selection in hyperdiverse species with large populations could strengthen such patterns of intragenic co-evolution through compensatory mutation between interacting sites. Analysis of protein sequence evolution in *Drosophila* lends support for this view. Nonsynonymous substitutions cluster together along the protein backbone, preferentially within the same lineage, at least in part because of epistatic interactions between substitutions (Callahan *et al.* 2011). And, the most conserved genes, which experience the most effective selection, have greatest clustering of substitutions (Callahan *et al.* 2011).

Hyperdiverse species are particularly well suited for studying how compensatory mutation can mediate divergence under stabilizing selection, such as for RNA gene evolution (Piskol & Stephan 2011), transcription factor (TF)/target site co-evolution (Baker *et al.* 2011) and developmental system drift (DSD) (True & Haag 2001). DSD involves the dissociation between homology at the phenotypic and molecular levels (Dickinson 1995; Abouheif *et al.* 1997; Wray & Abouheif 1998), and it may be common in developmental networks (True & Haag 2001; Abouheif & Wray 2002; Kiontke *et al.* 2007). When directional and stabilizing selection act on different traits that have an overlapping genetic basis (Johnson & Porter 2007), mutations at a locus for the trait that experiences stabilizing selection can compensate for divergence at a pleiotropic locus. The result is trait conservation at the phenotypic level despite divergence amongst species in its genetic underpinnings: 'systems drift' (Johnson & Porter 2007). Because all forms of natural selection are expected to be more efficient in species with large N_e, DSD might be particularly widespread amongst taxa that include hyperdiverse members and could therefore complicate inferences about the genetic basis of parallel and convergent evolution.

Ancestral state inference and ancestral polymorphism

Hyperdiversity complicates the inference of ancestral states for individual nucleotide sites. This is so because the deep genealogical history of a given locus increases the potential for multiple mutations to have occurred; this is most acute at sites that are observed presently to have 3 or 4 segregating variants. Unfortunately, this can create problems because the evolutionary polarization of mutational changes, to infer ancestral and derived states, is central to a variety of analytical methods (Kreitman 2000; Nielsen 2005). Typically, a single copy of an outgroup reference sequence is used, or sometimes a single copy from each of several outgroups, for comparison with the multiple sequences from the focal species' population. But this approach neglects polymorphism within the outgroup taxa, which will underestimate the uncertainty in ancestral states, and potentially prove egregious, particularly if both focal and outgroup species are hyperdiverse. Coupled with the deep genealogical history of a locus, lineages corresponding to the closest outgroup species may share considerable ancestral polymorphism or, in the case of reciprocal monophyly, be very distantly related (Pamilo & Nei 1988; Clark 1997; Arbogast *et al.* 2002). Either of these scenarios will introduce marked uncertainty into the inference of ancestral states, owing to ancestral polymorphism or multiple mutational hits, respectively (Baudry & Depaulis 2003; Hernandez *et al.* 2007). Amino-acid changing sites will be less of a problem for inferring ancestral states than will synonymous or noncoding sites, partly due to purifying selection plausibly favouring the encoding of the same amino acid in multiple species and partly due to the empirical rarity of more than two variants at replacement-sites. Consequently, it will be critical to appropriately model the mutational process (and demography) of hyperdiverse lineages to capture the complexities of determining ancestral-derived variant relationships (Charlesworth *et al.* 2005; Hernandez *et al.* 2007), or instead to rely more heavily on finite-sites versions of 'folded frequency spectrum' approaches for population genetic analysis (Kreitman 2000; Nielsen 2005; Zeng 2010).

Whole-genome assembly, alignment and conflation of alleles with gene duplicates

As population genomics takes its foothold as a standard means of quantifying population genetic variation, hyperdiverse organisms are particularly formidable for making biological sense of the data abundance (Vinson *et al.* 2005). For *de novo* genome assembly, heterozygosity within a sample makes the computational process of creating contigs from short reads a serious challenge (Donmez & Brudno 2011; Huang *et al.* 2012). Given high heterozygosity, allelic variation can easily be confused as paralogous duplicate copies of a genomic segment. For example, despite 20 generations of laboratory sib-mating, 10–30% of the reference genomes of several species of *Caenorhabditis* retained residual heterozygosity that initially

© 2013 Blackwell Publishing Ltd

inflated genome assembly sizes and estimates of the number of genes (Barriere *et al.* 2009). Even given a reference genome, high population polymorphism will make any given individual's genome very different from the reference. Many read-mapping algorithms can tolerate few 'mismatches' between the sample and reference, which can compromise analyses from the very start: for $\theta_{neu} = 0.05$, the 95% binomial expected range of mismatches is 2–8 for a random 100 bp sequence read. For example, only ~30% of sequence reads from a population genomics survey for *Ci. intestinalis* were aligned to a reference cDNA collection, although the coverage was not high and the reference sequence is from a distinct subspecies (Tsagkogeorga *et al.* 2012). Fortunately, sequence coverage information, incorporation of long-read technologies, independent estimates of genome size (from e.g. flow cytometry), and more sophisticated haplotype-aware assemblers can help identify and alleviate these problems (Vinson *et al.* 2005; Kim *et al.* 2007; Barriere *et al.* 2009; Donmez & Brudno 2011; Lunter & Goodson 2011; Huang *et al.* 2012). In sequencing a heterozygous diploid genome, it is key to have sufficient coverage to sample both alleles at a locus to be able to confidently distinguish the alternative haplotypes (Zharkikh *et al.* 2008). Whole-genome sequencing of a haploid genome also provides a viable option in some taxa to circumvent heterozygosity within a sample (Langley *et al.* 2011).

Rate of adaptation

The primary aim of molecular population genetics is to infer and quantify adaptation in DNA sequences. Functional genetic variation is, of course, the raw material for adaptive evolution, and yet we have defined hyperdiversity in terms of neutral variation—what should we expect for how adaptive evolution leaves its imprint in the genomes of hyperdiverse species? The answers to this question intertwine intimately with the prevalence of different modes of adaptation: what is the relative incidence of 'hard sweeps,' 'soft sweeps' and partial sweeps involving polygenic selection on traits (Pritchard *et al.* 2010)? Our expectations also depend on whether we conceive of adaptation as any form of positive selection, including compensatory evolution and genomic conflict, or in the more restrictive sense as a response to extrinsic ecological and/or environmental changes (Mustonen & Lassig 2009).

Hard, soft and partial selective sweeps

Hard sweeps epitomize the standard population genetic notion of adaptation: a new beneficial mutation arises and increases in frequency to fixation, in the process dragging along with it by genetic hitchhiking any linked variants (Maynard Smith & Haigh 1974). By contrast, soft sweeps involve the selective fixation of alleles that pre-existed in the population as segregating genetic variation, or the simultaneous selection on alternative alleles with equivalent fitness effects until their joint frequency reaches 1 (Hermisson & Pennings 2005). This broad soft sweep definition includes substitutions of beneficial alleles that are identical by descent, but that have recombined onto different genetic backgrounds, as well as substitutions by multiple copies of equivalent beneficial alleles that were created by independent mutations ('parallel adaptation'; Ralph & Coop 2010). The segregating alleles are typically presumed to originally occur at mutation-drift or mutation-selection balance (Hermisson & Pennings 2005). For polygenic selection on standing variation, adaptation might not yield fixation of particular alleles at any locus at all or any obvious molecular signals of positive selection, despite a robust phenotypic evolutionary response to even very strong selection on the phenotype (Latta 1998; Le Corre & Kremer 2003; Kelly 2006; Chevin & Hospital 2008). This will yield incomplete 'partial' selective sweeps at the trait loci that simply reflects shifts in allele-frequency (Chevin & Hospital 2008; Pritchard *et al.* 2010). By considering only hyperdiverse species with both large N_e and high θ_{neu}, we can restrict the relevant portion of parameter space to evaluate the merits and implications of these alternative modes of adaptation.

When selection is strong and the population mutation rate is low, then a selectively favoured allele most likely has a single mutational origin (Hermisson & Pennings 2005). Hyperdiverse organisms, however, are unlikely to conform neatly to this scenario—regardless of whether the hyperdiversity results from large N_e or high mutation rate. This implies that adaptation in such species often will not be mutation-limited and that selection via soft sweeps involving multiple allelic origins might be particularly important, especially for loci subject to weak selection (Hermisson & Pennings 2005; Ralph & Coop 2010). In the event of a soft sweep originating from segregating variation, species with large populations, in particular, will often not register much of a classic signature of deviation from neutrality around the selective target—in terms of polymorphism, LD, or the variant frequency spectrum—even for a selectively favoured allele with a single mutational origin (Innan & Kim 2004; Hermisson & Pennings 2005; Przeworski *et al.* 2005; Olson-Manning *et al.* 2012). This is because the coalescent depth of the beneficial allele typically is much greater than when selection favours a new allele, with recombination having diminished correspondingly more LD between the beneficial allele and nearby neutral variants. Consequently, genomic scans for selection that presume signatures of 'hard' selective sweeps might generally be impotent to detect much of the positive selection operat-

© 2013 Blackwell Publishing Ltd

ing in hyperdiverse organisms (Pritchard *et al.* 2010; Storz & Wheat 2010). Methods are being developed that focus on soft sweep detection (Pritchard *et al.* 2010; Sattath *et al.* 2011); if hyperdiverse organisms provide greater power than low-diversity species to identify soft sweeps by virtue of the greater density of polymorphic sites, then they may nevertheless prove valuable in testing theory related to dominance, epistasis and the distribution of selective effects (Orr & Betancourt 2001; Orr 2003, 2010; Kassen & Bataillon 2006; Teshima & Przeworski 2006; Silander *et al.* 2007; Otto & Whitlock 2009; Takahasi 2009).

Should hard sweeps be relatively uncommon to soft sweeps in the adaptive process of hyperdiverse populations, then this holds important implications for genome-wide patterns of polymorphism in such taxa. Hard sweeps provide the conceptual basis to standard recurrent genetic hitchhiking models that predict reduced polymorphism levels in regions of the genome subject to infrequent crossover recombination (Wiehe & Stephan 1993; Stephan 2010). Hard sweeps also frame Gillespie's models of genetic draft (Gillespie 2000a,b, 2001), suggesting that the consequences of draft will not be as extreme as predicted if soft sweeps dominate, because soft sweeps do not depress linked variation as greatly (Innan & Kim 2004; Hermisson & Pennings 2005; Przeworski *et al.* 2005) (Box 3). When many beneficial mutations across the genome are present simultaneously, then interference can limit their density and make recombination the limiting factor in combining beneficial genetic backgrounds in adaptation (Weissman & Barton 2012). Further theory is needed to better understand the consequences of recurrent soft sweeps on genome-wide patterns of polymorphism in contrast to hard sweeps.

Even if soft sweeps are a common mode of adaptation in hyperdiverse taxa, surely some fraction of the episodes of positive selection will involve hard sweeps, and we may ask how hyperdiversity will affect the detection and localization of them in genomes. Indeed, recurrent hard sweep models applied to *Drosophila simulans*, a species that approaches our definition of being hyperdiverse, yields inference of frequent substitutions driven by positive selection (Begun *et al.* 2007; Macpherson *et al.* 2007; Andolfatto *et al.* 2011). When applied to hyperdiverse populations, their greater background levels of diversity and density of SNPs should provide more statistical power to detect shifts in polymorphism, LD and the site frequency spectrum. In part, scanning genomes for hard sweeps is a practical stance, as most methods for detecting the imprint of selection presume this mode of adaptation within a local population or across the species range—and the corresponding signature of selection will be easier to identify (Storz 2005; Pavlidis *et al.* 2008; Siol *et al.* 2010). While we must recognize that the loci pinpointed by such analysis may represent a biased subset of substitutions with respect to the genetic architecture of adaptation, nevertheless, they still may identify novel adaptive phenotypes (Storz & Wheat 2010).

Box 3 Drift, Draft and N_e

On the one hand, N_e and genetic drift are often argued to play a dominant role in effecting evolutionary change (Lynch & Conery 2003b; Lynch 2007); on the other hand, some propose that differences amongst species in N_e are largely irrelevant to evolution (Gillespie 2001). In the latter view, the hitchhiking effects of recurrent selective sweeps are predicted to counteract the build-up of neutral polymorphism in species with large population size by 'genetic draft,' through the influence of genetic linkage (Maynard Smith & Haigh 1974; Gillespie 2000a,b; Neher & Shraiman 2011; Frankham 2012). The limitation of a species' polymorphism by the draft effect will be most pronounced when selection pressures fluctuate dramatically (Barton 2000) and could lead to asymptotic θ_{neu} with increasing population size, approaching a value of $2\mu/a$ (for complete linkage, where a is the rate of substitution via 'hard sweeps' fixing new beneficial mutations) (Gillespie 2000a,b, 2001). Thus, the draft model predicts that species with populations of a size exceeding a modest threshold will not differ appreciably in diversity. Consequently, fewer species will achieve hyperdiversity under the draft model than might be expected from the traditional expectation of a linear association between N_e and θ_{neu}. Note that development of the draft model was motivated by the observation that measures of molecular diversity do not seem to scale linearly with population size: measures of population size across species appear to range over more orders of magnitude than do estimates of θ_{neu} (Lewontin 1974; Maynard Smith & Haigh 1974; Gillespie 2000a; Lynch 2007; Leffler *et al.* 2012). Selection on cell-biological processes also might act to limit hyperdiversity, owing to base-pairing requirements of recombination and DNA repair (Stephan & Langley 1992; Harfe & Jinks-Robertson 2000). To some extent, the limiting effect of genetic draft on a species' polymorphism must occur, but, as discussed in the main text, the increasing relative importance in adaptation of 'soft sweeps' over 'hard sweeps' for species with larger N_e may alleviate the extremity of its influence. Theory that can quantify the genome-wide influence of recurrent soft sweeps in adaptation will prove valuable in connecting population size and diversity to selection and linkage.

© 2013 Blackwell Publishing Ltd

Notions of adaptive molecular evolution

Mustonen & Lassig (2007, 2009, 2010) formalized the notion of adaptation as a nonequilibrium stationary population state with a surfeit of beneficial relative to deleterious substitutions. A population experiences such a positive 'fitness flux' as a byproduct of fluctuations in selection (presumably resulting from environmental or ecological fluctuations) with a timescale on the order of the reciprocal of the mutation rate. This contrasts with a static fitness landscape in which positive selection might also fix beneficial mutations, but this generally would reflect the balance between deleterious and compensatory beneficial mutations; it also differs from a frenetically changing fitness landscape, in which substitutions tend towards quasi-neutrality (Mustonen & Lassig 2009). Consequently, the time-dependent fitness flux concept of adaptation is more restrictive than simple fixation by positive selection, and perhaps better reflects ecologically motivated notions of adaptive evolution (Mustonen & Lassig 2007, 2009, 2010; Mustonen *et al.* 2008). Because their approach explicitly implements finite-sites models of evolution, it may be particularly valuable for analysis of hyperdiverse organisms.

For recurring protein adaptation, as might be expected for organisms with positive fitness flux, tests of positive selection using variants of the McDonald–Kreitman test (MK test; McDonald & Kreitman 1991) may prove especially valuable for hyperdiverse species. Indeed, MK tests detect positive fitness flux (Mustonen & Lassig 2007). Special care must be taken with synonymous sites in MK tests to appropriately incorporate finite-sites mutational models, ancestral state inference and translational selection (McDonald & Kreitman 1991; Eyre-Walker 2002; Charlesworth *et al.* 2005; Desai & Plotkin 2008; Eyre-Walker & Keightley 2009; Mustonen & Lassig 2009). Interestingly, the fitness flux model of adaptation predicts that long-term fluctuating population sizes will yield a signal of positive fitness flux across the genome: net adaptive evolution (Mustonen & Lassig 2010). Perhaps lineages that are undergoing speciation, or that only recently became reproductively isolated (e.g. within *Drosophila*), might also tend to be in a state of positive fitness flux—which could account for positive associations between rates of speciation and molecular divergence (Webster *et al.* 2003; Pagel *et al.* 2006; Venditti & Pagel 2010).

Fisher (1958) argued that the rate of adaptive evolution will be most efficient and rapid in large populations [e.g. scaling linearly or logarithmically with N_e (Neher *et al.* 2010; Weissman & Barton 2012)]. Ohta (1972) subsequently predicted an inverse relationship between population size and the rate of adaptive substitution, based on the logic that large populations have bigger ranges that span more heterogeneous environmental conditions that, in turn, will retard the origin and fixation of globally advantageous mutations. And then, Gillespie (2001) predicted that the rate of adaptive evolution will be essentially uncorrelated with population size, owing to the interaction between selection and genetic linkage resulting in genetic draft (see also Neher & Shraiman 2011; Weissman & Barton 2012). Both interspecific and intra-genomic studies are consistent with higher N_e being associated with positive selection causing faster substitution rates (α, the fraction of amino-acid-changing mutations that are fixed by positive selection, or ω_a, the rate of positively selected substitution), in-line with Fisher's prediction (Presgraves 2005; Petit & Barbadilla 2009; Gossmann *et al.* 2010, 2012; Halligan *et al.* 2010; Siol *et al.* 2010; Andolfatto *et al.* 2011; Strasburg *et al.* 2011). However, it is unresolved empirically whether α might saturate with N_e, as would be predicted by Gillespie. Adding estimates of α for hyperdiverse eukaryotes would help distinguish these alternatives. A rough and simplistic estimate of α for the hyperdiverse *C.* sp. 5 [i.e., $\alpha \cong 1 - (\bar{d}_S \cdot \bar{\theta}_A)/(\bar{d}_N \cdot \bar{\theta}_{\text{neu}})$ (Smith & Eyre-Walker 2002)], gives $\alpha \cong 74\%$, suggesting that increasing N_e might continue to yield increasing α estimates in eukaryotes. Mustonen & Lassig's (2009) view of adaptation, for example, suggests that small populations may more likely fall in a range of positive fitness flux (i.e. $\kappa \gg \kappa^*$, the rate of environmental change being less than the average lifetime of polymorphisms), consistent with Ohta's prediction. By estimating rates of adaptive evolution in eukaryote taxa that span an extreme of population sizes, including hyperdiverse species, we may be able to more confidently evaluate these alternative predictions.

Balancing selection and the maintenance of hyperdiversity

Balancing selection can be a powerful force to maintain genetic variability, whether through heterozygote advantage, frequency dependence, or some types of spatial or temporal fluctuations in selection pressures (Charlesworth 2006; Mitchell-Olds *et al.* 2007). However, linkage of functionally important polymorphisms to nearby neutral variants will constantly be eroded by recombination, provided the alleles are not contained within inversions. As a result, over the long term, balancing selection leaves only a narrow footprint on the genome (Charlesworth 2006). Even in humans, with its long haplotype blocks, the narrow window of elevated linked variation around loci subject to long-term balancing selection makes it difficult to find examples of such selection in the genome (Bubb *et al.* 2006; Andres *et al.* 2009), and other factors like population admixture,

© 2013 Blackwell Publishing Ltd

pooled sampling of subpopulations and violation of the infinite sites mutation model can all generate skews in the site frequency spectrum in the same direction as will long-term balancing selection (Cutter *et al.* 2012). Input of duplication mutations, which would occur regularly in species with very large populations given estimated duplication rates (Lipinski *et al.* 2011), can provide a means of generating 'fixed heterozygosity' of beneficial allele combinations that does not require balancing selection (Mitchell-Olds *et al.* 2007). Also, recent work proposes that new beneficial mutations might commonly be subjected to short-term heterozygote advantage, which could leave genomic signatures of partial hard selective sweeps (Sellis *et al.* 2011). However, theory on frequent partial sweeps, which might also approximate the fate of loci subject to frequency-dependent selection with allele turnover, generally acts to reduce the amount of linked neutral polymorphism, albeit to a much lesser degree than do recurrent hard selective sweeps favouring a single positively selected allele (Coop & Ralph 2012). Consequently, even if widespread, balancing selection seems unlikely to provide a plausible explanation for genome-wide persistence of neutral hyperdiversity.

Constraint and the limits of natural selection

Purifying selection against detrimental alleles should be more efficacious in larger populations (Wright & Andolfatto 2008; Ellegren 2009). This greater efficacy of selection with increasing N_e is most important for weakly selected sites, however, because the probability of extinction of strongly deleterious mutations is essentially independent of N_e. Stated another way, species with larger N_e will have a smaller fraction of neutrally evolving sites, following from the nearly-neutral amendment to the neutral theory of molecular evolution (Ohta 1973). This effect should be evident in nucleotide sequences for hyperdiverse populations in several well-defined ways.

Constraint on protein sequence

If hyperdiversity is caused by large N_e rather than high mutation rates, then we expect polymorphism at nonsynonymous sites to be disproportionately small (i.e. low θ_A/θ_{neu}): even mildly deleterious changes to the amino acids encoded in genes will be eliminated by purifying selection (low θ_A), and most mutations with fitness effects are indeed deleterious (Eyre-Walker & Keightley 2007). This effect is clear at the scale of divergence amongst species (d_N/d_S) that differ in population size, with lineages having larger N_e tending to exhibit lower d_N/d_S (i.e. more effective purifying selection) (Wright & Andolfatto 2008; Ellegren 2009). For similar reasons, we also expect this stronger purifying selection to skew more strongly the variant frequency spectrum towards rare variants at such amino-acid changing sites (Tajima 1989). Nevertheless, hyperdiverse species might still segregate a high absolute number of nonsynonymous polymorphisms.

Constraint on synonymous sites

Weak selection on synonymous sites provides an ideal test case for exploring the limits to natural selection introduced by effective population size. In many species, it is reasonable to assume that synonymous site evolution is unconstrained, with any codon skews being caused by mutational or gene conversion biases (Duret 2002; Hershberg & Petrov 2008). However, in species that have sufficiently large N_e, selection operates effectively even for weak benefits associated with translational efficiency, accuracy and/or mRNA stability in highly expressed genes, causing biased complements of alternative synonymous codons (Duret 2002; Chamary *et al.* 2006; Hershberg & Petrov 2008). Consequently, in such species, it is inappropriate to assume selective neutrality for synonymous sites in highly expressed genes. Indeed, genes with more strongly skewed codon compositions, indicative of selection on codon usage, exhibit depressed levels of synonymous-site diversity in *C. remanei* (Dey *et al.* 2012) (Fig. 2). A more appropriate estimate of neutral diversity corrects for the direct effects of selection on synonymous sites and, in the case of *C. remanei*, such correction yields a nearly 50% increase (Dey *et al.* 2012). While this adjustment accounts for direct selection on synonymous sites, it does not incorporate any effects of linked selection on diversity nor does it correct the variant frequency spectrum, which will tend to be skewed towards rare variants. Synonymous sites in lowly-expressed genes, with minimal effects of translational selection, may represent a tractable set of the 'most unconstrained' sites in the genome (Cutter & Moses 2011).

Several methods use polymorphism data to quantify the intensity of selection favouring 'preferred' codons (e.g., Akashi 1995; McVean & Vieira 1999; Cutter & Charlesworth 2006; Zeng 2010; Zeng & Charlesworth 2010). Notably, species with larger population size show more pronounced selection for codon bias (Petit & Barbadilla 2009). If a hyperdiverse species lacks evidence for effective selection on synonymous sites, then this would imply that N_e probably does not drive its exceptional levels of polymorphism; instead, a high mutation rate, recent admixture of differentiated subpopulations, or an artefact of polyploidy might be the causes. Conversely, evidence of selection operating

© 2013 Blackwell Publishing Ltd

effectively on weakly selected sites should provide a good heuristic for deducing that the species has a very large long-term effective population size (e.g. free-living vs. parasitic nematodes, Cutter *et al.* 2006b). Further studies of the evolution of codon usage in hyperdiverse species will help dissect this interplay between mutation, genetic drift and natural selection (Box 4).

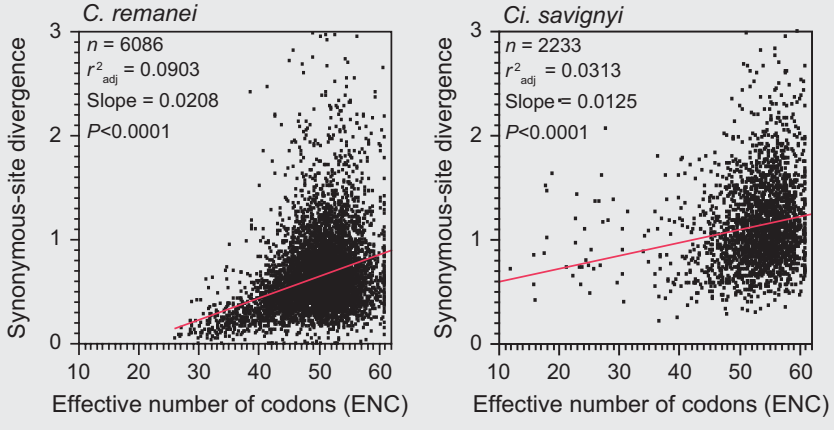
Functional inference of noncoding regions

Unveiling the genomic regulatory code is critical for understanding developmental evolution and the genesis of many diseases (Wray 2007; Wittkopp & Kalay 2012). Despite rapid progress in mapping regulatory interactions between TFs and their targets (Gerstein *et al.* 2010; Negre *et al.* 2011; Niu *et al.* 2011; Shen *et al.* 2012), the identification of *cis*-regulatory response elements (CRE) within genomes still presents a major challenge. Comparative genomics is a powerful approach for identifying functional elements via phylogenetic footprinting, relying on the assumption that similar sequences between multiple species are conserved because of purifying selection to maintain similar functions (Fig. 3). Although functional conservation is not always coupled with sequence conservation (Ludwig *et al.* 2005; Borneman *et al.* 2007), phylogenetic footprinting has been applied successfully to the human genome (Frazer *et al.* 2001, 2004; Dermitzakis *et al.* 2002) and to the identification of CREs in model organisms (Kellis *et al.* 2003; Marri & Gupta 2009).

Box 4 Insights on Ciona from synonymous sites

A brief case study of weak selection on synonymous sites permits some interesting insight into the potential causes of hyperdiversity in *Ci. savignyi*. A classic signature of translational selection is that rates of inter-species divergence will be depressed at synonymous sites (d_S) in genes with stronger codon bias (i.e. lower d_S for genes with lower values of the effective number of codons metric, ENC) (Sharp & Li 1987; Wright 1990). The more effective that selection is for biased codon usage over the long term, the clearer this pattern should be. Curiously, however, the $d_S \times$ ENC association is rather weak in *Ci. savignyi* (Fig. A), albeit significant even after incorporating $GC3_s$ as a covariate (not shown). Despite the nematode *C. remanei* having a much lower estimate of θ_{neu}, divergence of synonymous sites in the nematode are much more sensitive to codon usage (Fig. A). This suggests that the hyperdiversity observed in *Ci. savignyi* might not result entirely from exceptionally large long-term N_e, but perhaps instead results from a high rate of mutation or possibly recent admixture amongst previously separated populations following its invasive spread across marine habitats. The related *Ci. intestinalis* is hypothesized to have a higher mutation rate than vertebrates, and also is comprised of two cryptic species (Tsagkogeorga *et al.* 2012). Alternatively, translational selection might have unusually low selection coefficients in *Ciona*.

Figure A. Weak association between codon bias and synonymous-site divergence in *Ciona* implies little effective long-term selection on synonymous sites. Synonymous-site divergence for *C. remanei* reflects the lineage-specific values from (Cutter 2008a), based on a codon model of molecular evolution whereas *Ci. savignyi* values are half the Nei-Gojobori distance with orthologs of *Ci. intestinalis* (Small *et al.* 2007; Wang *et al.* 2010a); genes with values >3 excluded. effective number of codon (ENC) calculated with codonW (http://codonw.sourceforge.net); genes with values <10 excluded. See also Fig. 2 for the relation between polymorphism and codon bias in these species



© 2013 Blackwell Publishing Ltd

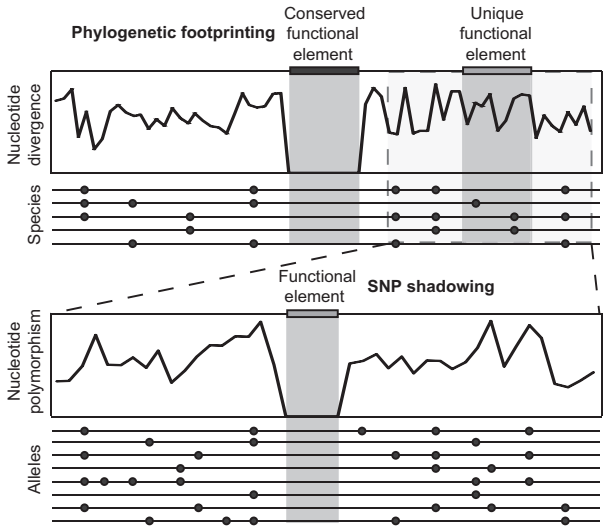


Fig. 3 Functional genome annotation of *cis*-regulation response elements (CRE) based on the pattern of selective constraints amongst genomes. Amongst-species comparative genomics is a powerful approach to identify conserved CREs by revealing phylogenetic footprints in the pattern of nucleotide divergence, but by nature fails to identify divergent and species-specific CREs (top). SNP shadowing, based on the pattern of allelic variation in hyperdiverse species, can circumvent some of the limitations of phylogenetic footprinting and helps refine *cis*-regulatory maps (bottom). Skewed measures of the variant frequency spectrum (e.g. Tajima's D) can be applied in a similar manner. Functional elements could be unique to a lineage owing to novel regulatory effects or to positional turnover of element motifs.

A drawback of phylogenetic footprinting is its sensitivity to species divergence: if the species are too distant, then important elements will be missed, if the species are too close, then functional conservation cannot be distinguished from shared ancestry. A similar strategy, termed phylogenetic shadowing, aims to improve resolution by exploiting species that occupy key phylogenetic positions (Boffelli *et al.* 2003). With the high density of polymorphisms within hyperdiverse populations, fine-scale genome annotation may be derived in an analogous way by using population genetic criteria to identify functional elements. We term this approach 'SNP shadowing,' which uses the same logic that purifying selection will constrain allelic variation for functional elements below neutral expectations, and also shift the site frequency spectrum towards rare variants. Boffelli *et al*. (2004) elegantly demonstrated the validity of the approach with a small number of individuals of the species *Ci. intestinalis*, by identifying conserved noncoding sequences that regulate in *cis* the expression of two nearby early-development TFs. SNP shadowing can identify species-specific CREs that would be inaccessible from comparisons with distantly related species, will assist in locating CREs with rearranged positions because of turnover (Doniger *et al.* 2005; Doniger & Fay 2007) and also may be particularly valuable in annotating the genomes of nonmodel hyperdiverse species. Although perturbations of patterns of polymorphism by selective sweeps, background selection or mutational cold-spots could complicate the detection of functional elements, applying the criterion of selective constraints in combination with other approaches such as the clustering of TF binding sites and assays of epigenetic features is likely to provide high-resolution *cis*-regulatory maps of the genome (Hardison & Taylor 2012).

Species and speciation

Standard coalescent theory tells us that the genealogical history of a given neutrally evolving locus traces to a single common ancestor $4N_e$ generations ago, on average, and that we expect neutral alleles for >95% of loci to have single common ancestors within the last $6N_e$ generations (Hein *et al.* 2004, p. 24). A pair of species will be reciprocally monophyletic at 95% of loci typically only after $>9N_e$ generations (Hudson & Coyne 2002). For hyperdiverse species with large N_e, such time depths will be very long ago. Consequently, hyperdiverse sister species will either (i) share very ancient most recent common ancestors, if most loci are reciprocally monophyletic, implying low speciation rates, or (ii) share with each other a great deal of ancestral polymorphism. Ancestral polymorphism amongst closely related species will create extensive gene tree discordance, making it essential to apply isolation-with-migration types of approaches to study differentiation (Hey 2006; but see Becquet & Przeworski 2009; Strasburg & Rieseberg 2010; Hey & Pinho 2012) and providing an interesting perspective to the divergence with gene flow debate (Porter & Johnson 2002; Nosil 2008; Pinho & Hey 2010).

It is challenging to discern whether hyperdiversity might hinder or facilitate speciation. On the one hand, hyperdiverse species may tolerate more easily the introgression of divergent genetic material through hybridization between populations or incipient species. Ecological and life-history characteristics that promote hyperdiversity—such as having large ranges with high dispersal—also may counteract extrinsic barriers to reproductive isolation. On the other hand, haplotypes with extremely dense differences might hinder recombination and DNA repair, which could counteract species cohesion (Harfe & Jinks-Robertson 2000). When coupled with local adaptation, population structure across a broad geographic range can accelerate speciation (Gavrilets 2003). Hyperdiverse taxa will harbour an abundance of raw material for divergent local adaptation, the crux of ecological speciation scenarios (Rundle & Nosil 2005), but divergent selection may be impotent to yield complete reproductive isolation unless it is

© 2013 Blackwell Publishing Ltd

sufficiently strong to offset gene flow. These features likely contribute to the observed high overall polymorphism within *Mimulus* (Sweigart & Willis 2003), *Paramecium* (Catania *et al.* 2009) and other species complexes.

Does high N_e or high μ affect species diversification rates? Using Hey's (2009) heuristic population genetic criteria for identifying species (separated for at least $1N_e$ generations with migration rate $N_em < 1$), it is clear that the maximum rate of speciation must inevitably be slower for lineages with larger population sizes (see also Nei *et al.* 1983; Gavrilets 2003). However, this genealogical species concept differs from the more broadly useful biological species concept in genetic work on reproductive isolation (Coyne & Orr 2004). 'Ubiquitous dispersal' of superabundant, very small organisms—which likely are predisposed to hyperdiversity—also is proposed to limit both species diversification and extinction (Fenchel 1993; Finlay 2002). Based on different logic (Gavrilets 2003; Pagel *et al.* 2006; Venditti & Pagel 2010)—namely, bottlenecks or extreme subdivision causing accelerated fixation of genetic differences and that induce reproductive isolation—a similar negative effect of population size on net diversification rates has been predicted. Note that we presume that time is measured in units of generations; as we expect hyperdiverse taxa typically to have short generation times (Box 2), this must be appropriately accounted in contrasts of net speciation rates. Because large N_e also can retard extinction (Wright 1983; Gaston 2000), it is a challenge to disentangle speciation from extinction when evaluating net diversification rates (Coyne & Orr 2004).

Lanfear *et al.* (2010) recently concluded that bird taxa with higher mutation rates tend to have higher net speciation rates. In a related example, Venditti & Pagel (2010) similarly reported a positive correlation between rate of diversification and synonymous-site substitution. If positive associations between speciation and mutation rates prove general, and provided that taxa with higher mutation rates will exhibit higher estimates of θ_{neu}, then we might expect taxa with more polymorphism to show greater net diversification rates—contrary to the above predictions based on N_e. Sexual selection also can potentially drive faster speciation rates when population sizes are larger (Gavrilets 2000). Further work is required to understand the balance of factors important for diversification in the context of hyperdiversity, including a species' propensity for local adaptation, mutation and extinction, as well as its susceptibility to geographic or other barriers to gene flow.

Linkage disequilibrium and recombination

Provided that a population's hyperdiversity is a byproduct of large N_e, rather than high μ, we also expect the effective rate of recombination to be high, as measured by the population recombination parameter, $\rho = 4N_ec$ (where c is the per-generation recombination rate between a pair of loci or polymorphic variants). It then follows that LD in the population should be low, overall, and will decay rapidly within the genome as a function of the distance between loci or polymorphic sites. LD does decay rapidly in hyperdiverse *Caenorhabditis* (Cutter *et al.* 2006a; Wang *et al.* 2010b). A beneficial consequence of this is that population genetic models that assume free recombination will provide a very good approximation for many applications, even for SNPs that are physically quite close together. However, high diversity means that finite-sites models of mutation should be implemented in the estimation of LD (Fearnhead & Donnelly 2001). A drawback to such rapid breakdown of LD is that population association mapping techniques for identifying the genetic basis of phenotypic variation will only find successfully the small subset of quantitative trait loci (QTL) that are extremely close to the markers used for genotyping (Ioannidis *et al.* 2009). Given a QTL under such circumstances, however, identifying causal nucleotide changes would then be relatively easy. As genome-wide re-sequencing of population samples becomes mainstream (Pool *et al.* 2010), this concern for association mapping will largely be obviated, particularly for compact genomes like those of *Caenorhabditis* (100–150 Mb) and *Drosophila* (150–200 Mb).

Hyperdiverse populations also offer an opportunity to study the pattern and process of noncrossover recombination (gene conversion) on a finer scale than in traditional systems. The high density of SNPs means that gene conversion tract lengths can be defined with greater precision (Sawyer 1989). Exploiting genomic heterogeneity in nucleotide composition and SNP density should facilitate analyses of G/C-nucleotide biased conversion (Marais 2003) and heterozygosity-dependence (Hilliker *et al.* 1991; Stephan & Langley 1992). High heterozygosity might inhibit the gene conversion repair of DNA double-strand breaks, with important evolutionary consequences (Stephan & Langley 1992; Harfe & Jinks-Robertson 2000). Incorporating gene conversion into estimation of ρ also is critical in hyperdiverse taxa (Andolfatto & Nordborg 1998; Cutter 2008b). Thus, study of hyperdiverse organisms should provide new insights into the mechanisms and evolutionary consequences of noncrossover recombination, particularly when coupled with the molecular tools developed for functional interrogation of related model organisms (Robert *et al.* 2008).

Just as the middle of the last century was plagued by the question of how could there be so much allozyme diversity in populations (Lewontin 1974), we now face the opposite challenge of explaining why neutral sequence diversity is not even greater than it is in

© 2013 Blackwell Publishing Ltd

species with large populations (Frankham 2012; Leffler *et al.* 2012). Recurring hard selective sweeps provide one plausible answer (Maynard Smith & Haigh 1974; Gillespie 2001), owing to linkage of selected targets to flanking loci causing reductions in neutral genetic variation through genetic hitchhiking. But, as discussed above, much selection is purifying [yielding reductions in diversity through background selection (Charlesworth *et al.* 1993; Charlesworth 2012)], and much of positive selection might not involve hard sweeps (Pritchard *et al.* 2010). By assessing the factors that control the interplay between selection and recombination in species with differing population size, with hyperdiverse species as a valuable extreme, we may better understand the confluence of factors that regulate natural genetic variation (Cutter & Payseur 2013).

Genome complexity

Effective population size is proposed as a critical factor in the dynamics of a variety of genomic features, including gene families (Lynch & Conery 2003a), the origin and loss of introns (Lynch 2002), operons (Lynch 2007) and interactome complexity (Fernandez & Lynch 2011). A difficulty with N_e-based explanations for genome complexity across the tree-of-life is that contrasts often are most stark between high-N_e prokaryotes and lower-N_e eukaryotes, thus conflating complexity and phylogenetic history (Lynch & Conery 2003b; Lynch 2006, 2007; Whitney & Garland 2010). Detailed studies of hyperdiverse eukaryotic lineages should provide a bridge to better test the generality of effective population size as a dominant factor in the evolution of genome architecture.

Studying population processes in hyperdiverse eukaryotes, particularly sexually reproducing metazoans, provides an exceptional counterpoint to taxa with small population sizes. For example, this can involve comparisons of selection on segregating variation for the presence/absence of introns within genes (Li *et al.* 2009; Schrider *et al.* 2011b) or of operon gene composition (Cutter & Agrawal 2010). Operons are genomic structures in which multiple genes are transcribed in a single RNA transcript, which is subsequently processed for translation of the separate genes (Blumenthal 2004). Lynch (2007) proposed that operons originated in the common ancestors of eukaryotes and prokaryotes, but that these structures have only been retained in eukaryote lineages that typically experience large N_e. Indeed, operons are common in the genomes of *Caenorhabditis* and *Ciona*, encompassing 15–20% of all encoded genes (Blumenthal *et al.* 2002; Satou *et al.* 2008; Cutter & Agrawal 2010). The evolution of different modes of regulatory control over gene expression also are hypothesized to be sensitive to population size, with enhancers favoured over suppressors in large populations (Gerland & Hwa 2009). Horizontal transfer of genetic material is common in bacteria, but its potential role in eukaryote evolution is poorly known, except that it seems to most often involve transmission of transposable elements (Schaack *et al.* 2010). Given that species with higher diversity tend to have fewer transposable elements and smaller genomes (Lynch & Conery 2003b), and that theoretical models implicate stronger selection for element elimination in large populations (Brookfield & Badge 1997), perhaps horizontal transfer of transposable elements would play a smaller role in genome architecture in hyperdiverse lineages. Hyperdiverse species may therefore provide a unique opportunity to study the ongoing molecular evolution and function of these intriguing genomic features.

The large pool of neutral mutations in hyperdiverse populations also can play a creative role in the origin of evolutionary novelties by defining the substrate upon which subsequent evolution is contingent (Stoltzfus 1999; Lynch 2007). One way that neutral mutations can influence the trajectory of evolution is through epistatic interactions between the fitness effect of a mutation and its genetic background (Kimura 1985). It is through these epistatic interactions with neutral mutations that alleles that otherwise would be deleterious can alter gene function and spread in populations (Zhang & Rosenberg 2002; Ortlund *et al.* 2007).

More drastic features of genome complexity—such as segmental and whole-genome duplications, karyotype changes, copy number variants, large inversions and other rearrangements—also might be sensitive to large effective population size. Specifically, if small N_e and genetic drift facilitate the fixation of such dramatic mutational events in many organisms, then we might expect large N_e taxa to have them only rarely as polymorphisms within species (Jakobsson *et al.* 2008) or as differences between species, because drift is a proportionately weaker force. Moreover, when large-scale differences do become fixed in hyperdiverse taxa, we could more confidently conjecture that positive selection drove the fixation and rapid accumulation of such large-scale changes to genome architecture.

Concluding remarks

Hyperdiversity in eukaryotes opens up exciting opportunities to enlighten our understanding of the evolutionary process. Such organisms represent a valuable bridge between bacteria and typical multicellular eukaryotic study systems, with key attributes of each: high polymorphism and enormous population size coupled with complex development and sexual reproduction. These features make hyperdiverse species novel

© 2013 Blackwell Publishing Ltd

models for dissecting a variety of fundamental issues, from genome complexity to modes of adaptation, mutational dynamics and fine-scale inference of sequence function. While we anticipate that the near future will reveal many eukaryote taxa harbouring extreme polymorphism levels, through application of emerging high-throughput sequencing technologies to molecular ecology, the existence of known hyperdiverse relatives of the model organisms *C. elegans* and *D. melanogaster* are particularly promising for detailed laboratory analyses.

Acknowledgements

We are grateful to Michael Brudno for providing sequence alignments for *Ciona*, to Josh Shapiro for providing diversity values for *C. elegans*, and to Heather Maughan for providing diversity values for *Haemophilus influenzae*. We thank three anonymous reviewers for their comments on the manuscript and Stephen Wright for insightful discussions. ADC is supported by the Natural Sciences and Engineering Research Council of Canada, grants from the National Institutes of Health and by a Canada Research Chair. RJ is supported by a fellowship from the Ontario Ministry of Research and Innovation.

References

Abouheif E, Wray GA (2002) Evolution of the gene network underlying wing polyphenism in ants. *Science*, **297**, 249–252.

Abouheif E, Akam M, Dickinson WJ *et al.* (1997) Homology and developmental genes. *Trends in Genetics*, **13**, 432–433.

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. *Genetics*, **139**, 1067–1076.

Amos W (2010) Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? *Proceedings of the Royal Society B: Biological Sciences*, **277**, 1443–1449.

Andersen EC, Gerke JP, Shapiro JA *et al.* (2012) Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. *Nature Genetics*, **44**, 285–290.

Andolfatto P, Nordborg M (1998) The effect of gene conversion on intralocus associations. *Genetics*, **148**, 1397–1399.

Andolfatto P, Wong KM, Bachtrog D (2011) Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species. *Genome Biology and Evolution*, **3**, 114–128.

Andres AM, Hubisz MJ, Indap A *et al.* (2009) Targets of balancing selection in the human genome. *Molecular Biology and Evolution*, **26**, 2755–2764.

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics*, **33**, 707–740.

Bachtrog D, Andolfatto P (2006) Selection, recombination and demographic history in *Drosophila miranda*. *Genetics*, **174**, 2045–2059.

Baker CR, Tuch BB, Johnson AD (2011) Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proceedings of the National Academy of Sciences, USA*, **108**, 7493–7498.

Barriere A, Yang SP, Pekarek E *et al.* (2009) Detecting heterozygosity in shotgun genome assemblies: lessons from obligately outcrossing nematodes. *Genome Research*, **19**, 470–480.

Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **355**, 1553–1562.

Barton N (2010) Understanding adaptation in large populations. *PLoS Genetics*, **6**, e1000987.

Baudry E, Depaulis F (2003) Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, **165**, 1619–1622.

Becquet C, Przeworski M (2009) Learning about modes of speciation by computational approaches. *Evolution*, **63**, 2547–2562.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, e310.

Blueweiss L, Fox H, Kudzma V *et al.* (1978) Relationships between body size and some life history parameters. *Oecologia*, **37**, 257–272.

Blumenthal T (2004) Operons in eukaryotes. *Briefings in Functional Genomics and Proteomics*, **3**, 199–211.

Blumenthal T, Evans D, Link CD *et al.* (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature*, **417**, 851–854.

Boffelli D, McAuliffe J, Ovcharenko D *et al.* (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, **299**, 1391–1394.

Boffelli D, Weer CV, Weng L *et al.* (2004) Intraspecies sequence comparisons for annotating genomes. *Genome Research*, **14**, 2406–2411.

Borneman AR, Gianoulis TA, Zhang ZD *et al.* (2007) Divergence of transcription factor binding sites across related yeast species. *Science*, **317**, 815–819.

Boudry P, Collet B, Cornette F, Hervouet V, Bonhomme F (2002) High variance in reproductive success of the Pacific oyster (*Crassostrea gigas*, Thunberg) revealed by microsatellite-based parentage analysis of multifactorial crosses. *Aquaculture*, **204**, 283–296.

Boyko AR, Williamson SH, Indap AR *et al.* (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, **4**, e1000083.

Brookfield JFY, Badge RM (1997) Population genetics models of transposable elements. *Genetica*, **100**, 281–294.

Bubb KL, Bovee D, Buckley D *et al.* (2006) Scan of human genome reveals no new loci under ancient balancing selection. *Genetics*, **173**, 2165–2177.

Callahan B, Neher RA, Bachtrog D, Andolfatto P, Shraiman BI (2011) Correlated evolution of nearby residues in Drosophilid proteins. *PLoS Genetics*, **7**, e1001315.

Casteleyn G, Leliaert F, Backeljau T *et al.* (2010) Limits to gene flow in a cosmopolitan marine planktonic diatom. *Proceedings of the National Academy of Sciences, USA*, **107**, 12952–12957.

Catania F, Wurmser F, Potekhin AA, Przybos E, Lynch M (2009) Genetic diversity in the *Paramecium aurelia* species complex. *Molecular Biology and Evolution*, **26**, 421–431.

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics*, **7**, 98–108.

Chao L, Carr DE (1993) The molecular clock and the relationship between population size and generation time. *Evolution*, **47**, 688–690.

Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, **2**, 379–384.

© 2013 Blackwell Publishing Ltd

Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. *Genetics*, **190**, 5–22.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth B, Bartolome C, Noel V (2005) The detection of shared and ancestral polymorphisms. *Genetics Research*, **86**, 149–157.

Chenuil A, Hoareau T, Egea E *et al.* (2010) An efficient method to find potentially universal population genetic markers, applied to metazoans. *BMC Evolutionary Biology*, **10**, 276.

Chevin LM, Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics*, **180**, 1645–1660.

Clark AG (1997) Neutral behavior of shared polymorphism. *Proceedings of the National Academy of Sciences, USA*, **94**, 7730–7734.

Coop G, Ralph P (2012) Patterns of neutral diversity under general models of selective sweeps. *Genetics*, **192**, 205–224.

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

Cutter AD (2008a) Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Molecular Biology and Evolution*, **25**, 778–786.

Cutter AD (2008b) Multilocus patterns of polymorphism and selection across the X-chromosome of *Caenorhabditis remanei*. *Genetics*, **178**, 1661–1672.

Cutter AD, Agrawal AF (2010) The evolutionary dynamics of operon distributions in eukaryote genomes. *Genetics*, **185**, 685–693.

Cutter AD, Charlesworth B (2006) Selection intensity on preferred codons correlates with overall codon usage bias in *Caenorhabditis remanei*. *Current Biology*, **16**, 2053–2057.

Cutter AD, Moses AM (2011) Polymorphism, divergence, and the role of recombination in *Saccharomyces cerevisiae* genome evolution. *Molecular Biology and Evolution*, **28**, 1745–1754.

Cutter AD, Payseur BA (2013) Genomic signatures of selection at linked sites: unifying the disparity among species. Nature Reviews Genetics. In press.

Cutter AD, Baird SE, Charlesworth D (2006a) High nucleotide polymorphism and rapid decay of linkage disequilibrium in wild populations of *Caenorhabditis remanei*. *Genetics*, **174**, 901–913.

Cutter AD, Wasmuth J, Blaxter ML (2006b) The evolution of biased codon and amino acid usage in nematode genomes. *Molecular Biology and Evolution*, **23**, 2303–2315.

Cutter AD, Dey A, Murray RL (2009) Evolution of the *Caenorhabditis elegans* genome. *Molecular Biology and Evolution*, **26**, 1199–1234.

Cutter AD, Wang G-X, Ai H, Peng Y (2012) Influence of finite-sites mutation, population subdivision and sampling schemes on patterns of nucleotide polymorphism for species with molecular hyperdiversity. *Molecular Ecology*, **21**, 1345–1359.

De S, Babu MM (2010) A time-invariant principle of genome evolution. *Proceedings of the National Academy of Sciences, USA*, **107**, 13004–13009.

Dehal P, Satou Y, Campbell RK *et al.* (2002) The draft genome of *Ciona intestinalis*: insights into chordate and vertebrate origins. *Science*, **298**, 2157–2167.

Denver DR, Howe DK, Wilhelm LJ *et al.* (2010) Selective sweeps and parallel mutation in the adaptive recovery from deleterious mutation in *Caenorhabditis elegans*. *Genome Research*, **20**, 1663–1671.

Der R, Epstein C, Plotkin JB (2012) The dynamics of neutral and selected alleles when the offspring distribution is skewed. *Genetics*, **191**, 1331–1344.

Dermitzakis ET, Reymond A, Lyle R *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, **420**, 578–582.

Desai MM, Plotkin JB (2008) The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, **180**, 2175–2191.

Dey A, Jeon Y, Wang G-X, Cutter AD (2012) Global population genetic structure of *Caenorhabditis remanei* reveals incipient speciation. *Genetics*, **191**, 1257–1269.

Dickinson WJ (1995) Molecules and morphology: where's the homology. *Trends in Genetics*, **11**, 119–121.

Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Computational Biology*, **3**, e99.

Doniger SW, Huh J, Fay JC (2005) Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Research*, **15**, 701–709.

Donmez N, Brudno M (2011) Hapsembler: an assembler for highly polymorphic genomes. *Research in Computational Molecular Biology*, **6577**, 38–52.

Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

Duret L (2002) Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, **12**, 640–649.

Dyer KA, Jaenike J (2004) Evolutionarily stable infection by a male-killing endosymbiont in *Drosophila innubila*: molecular evidence from the host and parasite genomes. *Genetics*, **168**, 1443–1455.

Dyer KA, Charlesworth B, Jaenike J (2007) Chromosome-wide linkage disequilibrium as a consequence of meiotic drive. *Proceedings of the National Academy of Sciences, USA*, **104**, 1587–1592.

Edwards SV (2009) Is a new and general theory of molecular systematics emerging? *Evolution*, **63**, 1–19.

Eldon B, Wakeley J (2006) Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, **172**, 2621–2633.

Ellegren H (2009) A selection model of molecular evolution incorporating the effective population size. *Evolution*, **63**, 301–305.

Estes S, Lynch M (2003) Rapid fitness recovery in mutationally degraded lines of *Caenorhabditis elegans*. *Evolution*, **57**, 1022–1030.

Estes S, Phillips PC, Denver DR, Thomas WK, Lynch M (2004) Mutation accumulation in populations of varying size: the distribution of mutational effects for fitness correlates in *Caenorhabditis elegans*. *Genetics*, **166**, 1269–1279.

Eyre-Walker A (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics*, **162**, 2017–2024.

Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, **8**, 610–618.

Eyre-Walker A, Keightley PD (2009) Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular Biology and Evolution*, **26**, 2097–2108.

© 2013 Blackwell Publishing Ltd

Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.
Fenchel T (1993) There are more small than large species? *Oikos*, **68**, 375–378.
Fernandez A, Lynch M (2011) Non-adaptive origins of interactome complexity. *Nature*, **474**, 502–505.
Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. *Science*, **296**, 1061–1063.
Fisher RA (1958) *The Genetical Theory of Natural Selection*. Dover Publications, New York.
Frankham R (1995a) Conservation genetics. *Annual Review of Genetics*, **29**, 305–327.
Frankham R (1995b) Effective population-size / adult-population size ratios in wildlife: a review. *Genetical Research*, **66**, 95–107.
Frankham R (2012) How closely does genetic diversity in finite populations conform to predictions of neutral theory? Large deficits in regions of low recombination. *Heredity*, **108**, 167–178.
Frazer KA, Sheehan JB, Stokowski RP *et al.* (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Research*, **11**, 1651–1659.
Frazer KA, Tao H, Osoegawa K *et al.* (2004) Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Research*, **14**, 367–372.
Fujita MK, Leache AD, Burbrink FT, McGuire JA, Moritz C (2012) Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution*, **27**, 480–488.
Gaston KJ (2000) Global patterns in biodiversity. *Nature*, **405**, 220–227.
Gavrilets S (2000) Rapid evolution of reproductive barriers driven by sexual conflict. *Nature*, **403**, 886–889.
Gavrilets S (2003) Models of speciation: what have we learned in 40 years? *Evolution*, **57**, 2197–2215.
Gerland U, Hwa T (2009) Evolutionary selection between alternative modes of gene regulation. *Proceedings of the National Academy of Sciences, USA*, **106**, 8841–8846.
Gerstein MB, Lu ZJ, Van Nostrand EL *et al.* (2010) Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, **330**, 1775–1787.
Gillespie JH (2000a) Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics*, **155**, 909–919.
Gillespie JH (2000b) The neutral theory in an infinite population. *Gene*, **261**, 11–18.
Gillespie JH (2001) Is the population size of a species relevant to its evolution? *Evolution*, **55**, 2161–2169.
Gossmann TI, Song B-H, Windsor AJ *et al.* (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Molecular Biology and Evolution*, **27**, 1822–1832.
Gossmann TI, Keightley PD, Eyre-Walker A (2012) The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biology and Evolution*, **4**, 658–667.
Goto H, Inomata N, Szmidt AE, Yamazaki T (2004) Nucleotide variation of seven genes in *Drosophila kikkawai*. *Genes & Genetic Systems*, **79**, 19–26.
Goyal S, Balick DJ, Jerison ER *et al.* (2012) Dynamic mutation-selection balance as an evolutionary attractor. *Genetics*, **191**, 1309–1319.
Haasl RJ, Payseur BA (2011) Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, **106**, 158–171.
Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PDN (2006) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences, USA*, **103**, 968–971.
Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD (2010) Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genetics*, **6**, e1000825.
Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, **13**, 469–483.
Hardison RC, Roskin KM, Yang S *et al.* (2003) Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research*, **13**, 13–26.
Harfe BD, Jinks-Robertson S (2000) DNA mismatch repair and genetic instability. *Annual Review of Genetics*, **34**, 359–399.
Hedgecock D (1994) Does variance in reproductive success limit effective population sizes of marine organisms? In: *Genetics and Evolution of Aquatic Organisms* (ed. Beaumont AR), pp. 122–134. Chapman & Hall, London.
Hedrick P (2005) Large variance in reproductive success and the Ne/N ratio. *Evolution*, **59**, 1596–1599.
Hein J, Schierup MH, Wiuf C (2004) *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.
Hermisson J, Pennings PS (2005) Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, **169**, 2335–2352.
Hernandez RD, Williamson SH, Bustamante CD (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Molecular Biology and Evolution*, **24**, 1792–1800.
Hershberg R, Petrov DA (2008) Selection on codon bias. *Annual Review of Genetics*, **42**, 287–299.
Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development*, **16**, 592–596.
Hey J (2009) On the arbitrary identification of real species. In: *Speciation and Patterns of Diversity*(eds Butlin R, Bridle J & Schluter D), pp. 15–28. Cambridge University Press, Cambridge.
Hey J, Pinho C (2012) Population genetics and objectivity in species diagnosis. *Evolution*, **66**, 1413–1429.
Hilliker AJ, Clark SH, Chovnick A (1991) The effect of DNA sequence polymorphisms on intragenic recombination in the *rosy* locus of *Drosophila melanogaster*. *Genetics*, **129**, 779–781.
Hodgkinson A, Eyre-Walker A (2010) Human triallelic sites: evidence for a new mutational mechanism? *Genetics*, **184**, 233–241.
Hohenlohe PA, Bassham S, Etter PD, *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.
Hollister JD, Ross-Ibarra J, Gaut BS (2010) Indel-associated mutation rate varies with mating system in flowering plants. *Molecular Biology and Evolution*, **27**, 409–416.
Hopf TA, Colwell LJ, Sheridan R *et al.* (2012) Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, **149**, 1607–1621.
Huang SF, Chen ZL, Huang GR *et al.* (2012) HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Research*, **22**, 1581–1588.
Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

© 2013 Blackwell Publishing Ltd

Hudson RR, Coyne JA (2002) Mathematical consequences of the genealogical species concept. *Evolution*, **56**, 1557–1565.

Hughes AL, Friedman R, Rivailler P, French JO (2008) Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Molecular Biology and Evolution*, **25**, 2199–2209.

Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proceedings of the National Academy of Sciences, USA*, **101**, 10667–10672.

Ioannidis JPA, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. *Nature Reviews Genetics*, **10**, 318–329.

Jakobsson M, Scholz SW, Scheet P *et al.* (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, **451**, 998–1003.

Jenkins PA, Song YS (2011) The effect of recurrent mutation on the frequency spectrum of a segregating site and the age of an allele. *Theoretical Population Biology*, **80**, 158–173.

Jiang D, Smith WC (2005) Self- and cross-fertilization in the solitary ascidian *Ciona savignyi*. *Biological Bulletin*, **209**, 107–112.

Johnson NA, Porter AH (2007) Evolution of branched regulatory genetic pathways: directional selection on pleiotropic loci accelerates developmental system drift. *Genetica*, **129**, 57–70.

Jovelin R (2009) Rapid sequence evolution of transcription factors controlling neuron differentiation in Caenorhabditis. *Molecular Biology and Evolution*, **26**, 2373–2386.

Jovelin R, Ajie BC, Phillips PC (2003) Molecular evolution and quantitative variation for chemosensory behaviour in the nematode genus Caenorhabditis. *Molecular Ecology*, **12**, 1325–1337.

Jovelin R, Dunham JP, Sung FS, Phillips PC (2009) High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in Caenorhabditis. *Genetics*, **181**, 1387–1397.

Karasov T, Messer PW, Petrov DA (2010) Evidence that adaptation in Drosophila is not limited by mutation at single sites. *PLoS Genetics*, **6**, e1000924.

Kassen R, Bataillon T (2006) Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nature Genetics*, **38**, 484–488.

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**, 740–743.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kelly JK (2006) Geographical variation in selection, from phenotypes to molecules. *American Naturalist*, **167**, 481–495.

Kim JH, Waterman MS, Li LM (2007) Diploid genome reconstruction of *Ciona intestinalis* and comparative analysis with *Ciona savignyi*. *Genome Research*, **17**, 1101–1110.

Kimura M (1968) Evolutionary rate at molecular level. *Nature*, **217**, 624–626.

Kimura M (1969) The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, **61**, 893–903.

Kimura M (1985) The role of compensatory neutral mutations in molecular evolution. *Journal of Genetics*, **64**, 7–19.

Kimura M, Crow JF (1963) The measurement of effective population number. *Evolution*, **17**, 279–288.

Kiontke K, Barriere A, Kolotuev I *et al.* (2007) Trends, stasis, and drift in the evolution of nematode vulva development. *Current Biology*, **17**, 1925–1937.

Kiontke K, Felix M-A, Ailion M *et al.* (2011) A phylogeny and molecular barcodes for Caenorhabditis, with numerous new species from rotting fruits. *BMC Evolutionary Biology*, **11**, 339.

Knowles LL, Maddison WP (2002) Statistical phylogeography. *Molecular Ecology*, **11**, 2623–2635.

Kreitman M (2000) Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics*, **1**, 539–559.

Lanfear R, Ho SYW, Love D, Bromham L (2010) Mutation rate is linked to diversification in birds. *Proceedings of the National Academy of Sciences, USA*, **107**, 20423–20428.

Langley CH, Crepeau M, Cardeno C, Corbett-Detig R, Stevens K (2011) Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics*, **188**, 239–246.

Langley CH, Stevens K, Cardeno C *et al.* (2012) Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics*, **192**, 533–598.

Latta RG (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *American Naturalist*, **151**, 283–292.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait loci and trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Lee H, Choi M-K, Lee D *et al.* (2012) Nictation, a dispersal behavior of the nematode *Caenorhabditis elegans*, is regulated by IL2 neurons. *Nature Neuroscience*, **15**, 107–112.

Leffler EM, Bullaughey K, Matute DR *et al.* (2012) Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biology*, **10**, e1001388.

Lemaire P (2011) Evolutionary crossroads in developmental biology: the tunicates. *Development*, **138**, 2143–2152.

Lewontin RC (1974) *The Genetic Basis to Evolutionary Change*. Columbia University Press, New York.

Li W, Tucker AE, Sung W, Thomas WK, Lynch M (2009) Extensive, recent intron gains in Daphnia populations. *Science*, **326**, 1260–1262.

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, **25**, 1451–1452.

Lipinski KJ, Farslow JC, Fitzpatrick KA *et al.* (2011) High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Current Biology*, **21**, 306–310.

Longman-Jacobsen N, Williamson JF, Dawkins RL, Gaudieri S (2003) In polymorphic genomic regions indels cluster with nucleotide polymorphism: Quantum Genomics. *Gene*, **312**, 257–261.

Ludwig MZ, Palsson A, Alekseeva E *et al.* (2005) Functional evolution of a cis-regulatory module. *PLoS Biology*, **3**, e93.

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, **21**, 936–939.

Lynch M (2002) Intron evolution as a population-genetic process. *Proceedings of the National Academy of Sciences, USA*, **99**, 6118–6123.

Lynch M (2006) The origins of eukaryotic gene structure. *Molecular Biology and Evolution*, **23**, 450–468.

Lynch M (2007) *The Origins of Genome Architecture*. Sinauer, Sunderland, MA.

© 2013 Blackwell Publishing Ltd

Lynch M, Conery JS (2003a) The evolutionary demography of duplicate genes. *Journal of Structural and Functional Genomics*, **3**, 35–44.

Lynch M, Conery JS (2003b) The origins of genome complexity. *Science*, **302**, 1401–1404.

Macpherson JM, Sella G, Davis JC, Petrov DA (2007) Genome-wide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in Drosophila. *Genetics*, **177**, 2083–2099.

Maiden MCJ, Bygraves JA, Feil E *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences, USA*, **95**, 3140–3145.

Marais G (2003) Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics*, **19**, 330–338.

Marks DS, Colwell LJ, Sheridan R *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE*, **6**, e28766.

Marquet PA, Quinones RA, Abades S *et al.* (2005) Scaling and power-laws in ecological systems. *Journal of Experimental Biology*, **208**, 1749–1769.

Marri S, Gupta BP (2009) Dissection of *lin-11* enhancer regions in *Caenorhabditis elegans* and other nematodes. *Developmental Biology*, **325**, 402–411.

Maughan H, Redfield RJ (2009) Extensive variation in natural competence in *Haemophilus Influenzae*. *Evolution*, **63**, 1852–1866.

Maynard Smith J, Haigh J (1974) Hitch-hiking effect of a favorable gene. *Genetical Research*, **23**, 23–35.

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature*, **351**, 652–654.

McDonald MJ, Wang WC, Huang HD, Leu JY (2011) Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biology*, **9**, e1000622.

McVean GAT, Vieira J (1999) The evolution of codon preferences in Drosophila: a maximum-likelihood approach to parameter estimation and hypothesis testing. *Journal of Molecular Evolution*, **49**, 63–75.

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

Milgroom MG, Fry WE (1997) Contributions of population genetics to plant disease epidemiology and management. *Advances in Botanical Research*, **24**, 1–30.

Mitchell-Olds T, Willis JH, Goldstein DB (2007) Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics*, **8**, 845–856.

Mustonen V, Lassig M (2007) Adaptations to fluctuating selection in Drosophila. *Proceedings of the National Academy of Sciences, USA*, **104**, 2277–2282.

Mustonen V, Lassig M (2009) From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in Genetics*, **25**, 111–119.

Mustonen V, Lassig M (2010) Fitness flux and ubiquity of adaptive evolution. *Proceedings of the National Academy of Sciences, USA*, **107**, 4248–4251.

Mustonen V, Kinney J, Callan CG, Lassig M (2008) Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences, USA*, **105**, 12376–12381.

Negre N, Brown CD, Ma LJ *et al.* (2011) A cis-regulatory map of the Drosophila genome. *Nature*, **471**, 527–531.

Neher RA, Shraiman BI (2011) Genetic draft and quasi-neutrality in large facultatively sexual populations. *Genetics*, **188**, 975–996.

Neher RA, Shraiman BI, Fisher DS (2010) Rate of adaptation in large sexual populations. *Genetics*, **184**, 467–481.

Nei M, Maruyama T, Wu C-I (1983) Models of evolution of reproductive isolation. *Genetics*, **103**, 557–579.

Nielsen R (2005) Molecular signatures of natural selection. *Annual Review of Genetics*, **39**, 197–218.

Ninio J (2010) Frail hypotheses in evolutionary biology. *PLoS Genetics*, **6**, e1001067.

Niu W, Lu ZJ, Zhong M *et al.* (2011) Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. *Genome Research*, **21**, 245–254.

Nordborg M, Hu TT, Ishino Y *et al.* (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology*, **3**, e196.

Nosil P (2008) Speciation with gene flow could be common. *Molecular Ecology*, **17**, 2103–2106.

Nunney L (1996) The influence of variation in female fecundity on effective population size. *Biological Journal of the Linnean Society of London*, **59**, 411–425.

Nydam ML, Harrison RG (2010) Polymorphism and divergence within the ascidian genus Ciona. *Molecular Phylogenetics and Evolution*, **56**, 718–726.

Ohta T (1972) Population size and rate of evolution. *Journal of Molecular Evolution*, **1**, 305–314.

Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature*, **246**, 96–98.

Olson-Manning CF, Wagner MR, Mitchell-Olds T (2012) Adaptive evolution: evaluating empirical support for theoretical predictions. *Nature Reviews Genetics*, **13**, 867–877.

Orr HA (2003) The distribution of fitness effects among beneficial mutations. *Genetics*, **163**, 1519–1526.

Orr HA (2010) The population genetics of beneficial mutations. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **365**, 1195–1201.

Orr HA, Betancourt AJ (2001) Haldane's sieve and adaptation from the standing genetic variation. *Genetics*, **157**, 875–884.

Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW (2007) Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, **317**, 1544–1548.

Otto SP, Whitlock MC (2009) The impact of epistatic selection on the genomic traces of selection. *Molecular Ecology*, **18**, 4985–4987.

Pagel M, Venditti C, Meade A (2006) Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science*, **314**, 119–121.

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. *Molecular Biology and Evolution*, **5**, 568–583.

Pannell JR, Charlesworth B (1999) Neutral genetic diversity in a metapopulation with recurrent local extinction and recolonization. *Evolution*, **53**, 664–676.

Pavlidis P, Hutter S, Stephan W (2008) A population genomic approach to map recent positive selection in model species. *Molecular Ecology*, **17**, 3585–3598.

Payseur BA, Cutter AD (2006) Integrating patterns of polymorphism at SNPs and STRs. *Trends in Genetics*, **22**, 424–429.

Perez-Losada M, Browne EB, Madsen A *et al.* (2006) Population genetics of microbial pathogens estimated from multilocus

© 2013 Blackwell Publishing Ltd

sequence typing (MLST) data. *Infection Genetics and Evolution*, **6**, 97–112.

Perry GH, Melsted P, Marioni JC *et al.* (2012) Comparative RNA sequencing reveals substantial genetic variation in endangered primates. *Genome Research*, **22**, 602–610.

Peters RH, Wassenberg K (1983) The effect of body size on animal abundance. *Oecologia*, **60**, 89–96.

Petit N, Barbadilla A (2009) Selection efficiency and effective population size in Drosophila species. *Journal of Evolutionary Biology*, **22**, 515–526.

Pinho C, Hey J (2010) Divergence with gene flow: models and data. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 215–230.

Piskol R, Stephan W (2011) The role of the effective population size in compensatory evolution. *Genome Biology and Evolution*, **3**, 528–538.

Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Research*, **20**, 291–300.

Porter AH, Johnson NA (2002) Speciation despite gene flow when developmental pathways evolve. *Evolution*, **56**, 2103–2111.

Pray LA, Goodnight CJ, Stevens L, Schwartz JM, Yan G (1996) The effect of population size on effective population size: an empirical study in the red flour beetle *Tribolium castaneum*. *Genetics Research*, **68**, 151–155.

Presgraves DC (2005) Recombination enhances protein adaptation in *Drosophila melanogaster*. *Current Biology*, **15**, 1651–1656.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.

Przeworski M, Coop G, Wall JD (2005) The signature of positive selection on standing genetic variation. *Evolution*, **59**, 2312–2323.

Ralph P, Coop G (2010) Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*, **186**, 647–668.

Rannala B, Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, **164**, 1645–1656.

Robert VJ, Davis MW, Jorgensen EM, Bessereau JL (2008) Gene conversion and end-joining-repair double-strand breaks in the *Caenorhabditis elegans* germline. *Genetics*, **180**, 673–679.

Ross-Ibarra J, Wright SI, Foxe JP *et al.* (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. *PLoS ONE*, **3**, e2411.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Satou Y, Mineta K, Ogasawara M *et al.* (2008) Improved genome assembly and evidence-based global gene model set for the chordate *Ciona intestinalis*: new insight into intron and operon populations. *Genome Biology*, **9**, R152.

Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G (2011) Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. *PLoS Genetics*, **7**, e1001302.

Sauvage C, Bierne N, Lapegue S, Boudry P (2007) Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. *Gene*, **406**, 13–22.

Sawyer S (1989) Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, **6**, 526–538.

Schaack S, Gilbert C, Feschotte C (2010) Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, **25**, 537–546.

Schrider DR, Hourmozdi JN, Hahn MW (2011a) Pervasive multinucleotide mutational events in eukaryotes. *Current Biology*, **21**, 1051–1054.

Schrider DR, Stevens K, Cardeno CM, Langley CH, Hahn MW (2011b) Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Research*, **21**, 2087–2095.

Sella G, Petrov DA, Przeworski M, Andolfatto P (2009) Pervasive natural delection in the Drosophila genome? *PLoS Genetics*, **5**, e1000495.

Sellis D, Callahan BJ, Petrov DA, Messer PW (2011) Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences, USA*, **108**, 20666–20671.

Sharp PM, Li WH (1987) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Molecular Biology and Evolution*, **4**, 222–230.

Shen Y, Yue F, McCleary DF *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

Shoemaker DD, Dyer KA, Ahrens M, McAbee K, Jaenike J (2004) Decreased diversity but increased substitution rate in host mtDNA as a consequence of Wolbachia endosymbiont infection. *Genetics*, **168**, 2049–2058.

Silander OK, Tenaillon O, Chao L (2007) Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS Biology*, **5**, e94.

Siol M, Wright SI, Barrett SCH (2010) The population genomics of plant adaptation. *New Phytologist*, **188**, 313–332.

Small KS, Brudno M, Hill MM, Sidow A (2007) Extreme genomic variation in a natural population. *Proceedings of the National Academy of Sciences, USA*, **104**, 5698–5703.

Smith NGC, Eyre-Walker A (2002) Adaptive protein evolution in Drosophila. *Nature*, **415**, 1022–1024.

Stephan W (2010) Genetic hitchhiking versus background selection: the controversy and its implications. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **365**, 1245–1253.

Stephan W, Langley CH (1992) Evolutionary consequences of DNA mismatch inhibited repair opportunity. *Genetics*, **132**, 567–574.

Stevens MH, Sanchez M, Lee J, Finkel SE (2007) Diversification rates increase with population size and resource concentration in an unstructured habitat. *Genetics*, **177**, 2243–2250.

Stolfi A, Christiaen L (2012) Genetic and genomic toolbox of the chordate *Ciona intestinalis*. *Genetics*, **192**, 55–66.

Stoltzfus A (1999) On the possibility of constructive neutral evolution. *Journal of Molecular Evolution*, **49**, 169–181.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Storz JF, Wheat CW (2010) Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution*, **64**, 2489–2509.

Strasburg JL, Rieseberg LH (2010) How robust are "Isolation with Migration" analyses to violations of the IM model? A simulation study. *Molecular Biology and Evolution*, **27**, 297–310.

© 2013 Blackwell Publishing Ltd

Strasburg JL, Kane NC, Raduski AR *et al.* (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology and Evolution*, **28**, 1569–1580.

Sweigart AL, Willis JH (2003) Patterns of nucleotide diversity in two species of Mimulus are affected by mating system and asymmetric introgression. *Evolution*, **57**, 2490–2506.

Tajima F (1983) Evolutionary relationship of DNA-sequences in finite populations. *Genetics*, **105**, 437–460.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.

Tajima F (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics*, **143**, 1457–1465.

Takahasi KR (2009) Coalescent under the evolution of coadaptation. *Molecular Ecology*, **18**, 5018–5029.

Teshima KM, Przeworski M (2006) Directional positive selection on an allele of arbitrary dominance. *Genetics*, **172**, 713–718.

Therriault TW, Herborg L-M (2008) Predicting the potential distribution of the vase tunicate *Ciona intestinalis* in Canadian waters: informing a risk assessment. *ICES Journal of Marine Science*, **65**, 788–794.

Tian D, Wang Q, Zhang P *et al.* (2008) Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, **455**, 105–108.

True JR, Haag ES (2001) Developmental system drift and flexibility in evolutionary trajectories. *Evolution and Development*, **3**, 109–119.

Tsagkogeorga G, Cahais V, Galtier N (2012) The population genomics of a fast evolver: high levels of diversity, functional constraint, and molecular adaptation in the tunicate *Ciona intestinalis*. *Genome Biology and Evolution*, **4**, 852–861.

Tsompana M, Abad J, Purugganan M, Moyer JW (2005) The molecular population genetics of the Tomato spotted wilt virus (TSWV) genome. *Molecular Ecology*, **14**, 53–66.

Vences M, Thomas M, Bonett RM, Vieites DR (2005) Deciphering amphibian diversity through DNA barcoding: chances and challenges. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1859–1868.

Venditti C, Pagel M (2010) Speciation as an active force in promoting genetic evolution. *Trends in Ecology & Evolution*, **25**, 14–20.

Vinson JP, Jaffe DB, O'Neill K *et al.* (2005) Assembly of polymorphic genomes: algorithms and application to *Ciona savignyi*. *Genome Research*, **15**, 1127–1135.

Wang J (2005) Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**, 1395–1409.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J (2010a) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*, **8**, 77–80.

Wang G-X, Ren S, Ren Y, Ai H, Cutter AD (2010b) Extremely high molecular diversity within the East Asian nematode *Caenorhabditis* sp. 5. *Molecular Ecology*, **19**, 5022–5029.

Wares JP, Pringle JM (2008) Drift by drift: effective population size is limited by advection. *BMC Evolutionary Biology*, **8**, 235.

Webster AJ, Payne RJH, Pagel M (2003) Molecular phylogenies link rates of evolution and speciation. *Science*, **301**, 478–478.

Weissman DB, Barton NH (2012) Limits to the rate of adaptive substitution in sexual populations. *PLoS Genetics*, **8**, e1002740.

Wetterbom A, Sevov M, Cavelier L, Bergstrom TF (2006) Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *Journal of Molecular Evolution*, **63**, 682–690.

Whitney KD, Garland T (2010) Did genetic drift drive increases in genome complexity? *PLoS Genetics*, **6**, e1001080.

Wiehe THE, Stephan W (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Molecular Biology and Evolution*, **10**, 842–854.

Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, **13**, 59–69.

Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, **8**, 206–216.

Wray GA, Abouheif E (1998) When is homology not homology? *Current Opinion in Genetics & Development*, **8**, 675–680.

Wright S (1933) Inbreeding and homozygosis. *Proceedings of the National Academy of Sciences, USA*, **19**, 411–420.

Wright DH (1983) Species-energy theory: an extension of species-area theory. *Oikos*, **41**, 496–506.

Wright F (1990) The effective number of codons used in a gene. *Gene*, **87**, 23–29.

Wright SI, Andolfatto P (2008) The impact of natural selection on the genome: emerging patterns in Drosophila and Arabidopsis. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 193–213.

Zeng K (2010) A simple multiallele model and its application to identifying preferred-unpreferred codons using polymorphism data. *Molecular Biology and Evolution*, **27**, 1327–1337.

Zeng K, Charlesworth B (2010) Studying patterns of recent evolution at synonymous sites and intronic sites in *Drosophila melanogaster*. *Journal of Molecular Evolution*, **70**, 116–128.

Zhang JZ, Rosenberg HF (2002) Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proceedings of the National Academy of Sciences, USA*, **99**, 5486–5491.

Zharkikh A, Troggio M, Pruss D *et al.* (2008) Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: problems and solutions. *Journal of Biotechnology*, **136**, 38–43.

Zhou R, Zeng K, Wu W *et al.* (2007) Population genetics of speciation in nonmodel organisms: I. ancestral polymorphism in mangroves. *Molecular Biology and Evolution*, **24**, 2746–2754.

Zhu L, Wang Q, Tang P, Araki H, Tian D (2009) Genomewide association between insertions/deletions and the nucleotide diversity in bacteria. *Molecular Biology and Evolution*, **26**, 2353–2361.

Research in A.D.C. laboratory addresses genome evolution, reproductive isolation, and adaptation, focusing on population genetic and molecular evolutionary approaches in nematodes. R.J. is interested in molecular population genetics, the evolution of gene regulation and the molecular evolution of gene networks and miRNAs. A.D. is studying molecular population genetics and reproductive isolation in Caenorhabditis nematodes.

© 2013 Blackwell Publishing Ltd

Data accessibility

Citations for data sources are given in figure legends.

Glossary

Hyperdiversity : Nucleotide diversity, defined as the average proportion of nucleotides that differ between a random pair of alleles, $\geq 5\%$.

N_e *:* The effective population size that corresponds to an idealized 'Wright-Fisher' population of size N, which describes the rate of genetic drift in the population.

θ *:* The population mutation rate parameter ($N_e\mu$, mutation rate μ), estimated from segregating polymorphisms within a population.

α *:* The proportion of nonsynonymous site changes that are fixed by positive selection.

d_N *:* The rate of accumulation of fixed changed at nonsynonymous sites in protein-coding sequences.

d_S *:* The rate of accumulation of fixed changes at synonymous sites in protein-coding sequences.

site frequency spectrum : The distribution of single nucleotide variant frequencies within a population sample.

codon bias : The unequal use of alternative synonymous codons caused by mutational biases, gene conversion biases or selection.

inbreeding depression : The reduction of fitness observed in the progeny of crosses from related genotypes because of the expression of recessive mutations that are masked in the parents.

ancestral polymorphism : Polymorphism present in an ancestral population that can lead to shared polymorphism in descendant species or discordance of gene trees and species trees.

reciprocal monophyly : Two lineage groups in which genes share a more recent common ancestor within each group than with any other lineages, leading to perfectly congruent gene trees and species trees.

linkage disequilibrium : The nonrandom association of variants within a population sample.

hard selective sweep : Fixation of a new beneficial allele by positive selection that induces linked variants to also become fixed by genetic hitchhiking.

soft selective sweep : Fixation of a beneficial allele by positive selection acting on pre-existing genetic variants or on a set of equivalent alleles that arose through multiple independent mutational events

fitness flux : A measure of adaptation quantified from an excess of beneficial over deleterious allele fixations across the genome

SNP : single nucleotide polymorphism

STR : short tandem repeat or microsatellite locus

© 2013 Blackwell Publishing Ltd