Review

Cell PRESS

Genome sequencing and population genomics in non-model organisms

Hans Ellegren

Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

High-throughput sequencing technologies are revolutionizing the life sciences. The past 12 months have seen a burst of genome sequences from non-model organisms, in each case representing a fundamental source of data of significant importance to biological research. This has bearing on several aspects of evolutionary biology, and we are now beginning to see patterns emerging from these studies. These include significant heterogeneity in the rate of recombination that affects adaptive evolution and base composition, the role of population size in adaptive evolution, and the importance of expansion of gene families in lineage-specific adaptation. Moreover, resequencing of population samples (population genomics) has enabled the identification of the genetic basis of critical phenotypes and cast light on the landscape of genomic divergence during speciation.

The omics era of biology

One way of boldly characterizing some significant achievements in biology during the past century is to recognize three major developments: the modern synthesis, the emergence of molecular biology, and the 'omics' era, appearing with approximately 30-yr intervals. Preceded by one or two decades of active struggle to access the genome seriously, via markers and sequence tags, the ability to unravel the complete genetic code of organisms demarked a start of genomics during the 1990s. Soon thereafter, access to sequenced genomes laid the ground for further characterization of molecular and phenotypic features related to the genome, leading in turn to the coining of the phrase 'omics'. Today, there are numerous derivatives of the basic concept of large-scale biological analyses, with the common denominator of aiming to study the complete repertoire of particular molecules (e.g., transcriptome and peptidome; see Glossary), modifications (e.g., degradome and methylome) or traits (e.g., behaviourome and phenome). Clearly, biology is getting increasingly large scale, quantitative, and integrative; in fact, future generations of biologists will perhaps come to see the integration of the abovementioned achievements as an outstanding achievement in itself.

When new concepts and approaches enter the scene in science, there is usually an initial period of hype and high expectations for what is to come. Genomics was no exception. In the field of evolutionary biology, there were good reasons to expect important breakthroughs. Given that genome sequences of non-model organisms are accumulating at an unprecedented pace [1–14], it is time to evaluate the outcome of genome-sequencing projects and what one can learn about the evolution of natural populations from such endeavors. In this review, I begin by describing the current status of genome sequencing in non-model organisms of animals, plants, and other eukaryotes, and what the sequences can inform about evolution. I then discuss the state of the art in the field of population genomics in which whole-genome sequencing of population samples offers an exciting reverse genetic venue towards, for example, the study of adaptation, trait evolution, and species divergence.

Genome sequences of non-model organisms: an overview

Status of genome sequences

Acknowledging that there is no clear definition of how large a proportion of a genome should have been sequenced to merit being referred to as a genome sequence (Box 1), an indication of the progress in the field can be obtained by noting that the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) currently (April, 2013) lists publically available information on genome sequence assemblies from 644 eukaryotes (Table 1). Although the first reported eukaryotic genome sequences were mainly classical, experimental models, such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and mice, the mere number on the list indicates that most species presently sequenced represent non-model organisms. However, the list is biased in favor of certain taxonomic groups. More than 0.1% of all vertebrate genomes have been sequenced, with mammals representing the so far best-characterized class, with >1% of all species sequenced. For plants and fungi, the proportion of species sequenced is on the order of 0.01%, whereas for insects, only some 0.001% of the species have been subject to genome sequencing. The list is also biased towards domesticated species of horticultural (e.g., orange, pear, and peach) or agricultural (major crops and farm animals) interests where genome sequencing has been motivated by a facilitated improvement in breeding.

Corresponding author: Ellegren, H. (Hans.Ellegren@ebc.uu.se).

Keywords: ecological genomics; evolutionary genomics; genome sequencing; molecular evolution; adaptive evolution; positive selection; population genomics; speciation genetics.

0169-5347/$ – see front matter

 http://dx.doi.org/10.1016/j.tree.2013.09.008

Glossary

Allele frequency spectra: the distribution of allele frequencies among a large set of polymorphic sites. An unfolded spectrum uses information on the ancestral state in that the frequency of derived alleles is depicted. If such information is not available, a folded spectrum simply depicts the frequency of the minor allele (and, hence, has 0.5 as its maximum frequency).
Behaviourome ('mental map): a term mainly used in human biology that refers to the diversity of ideas an individual makes in any given situation or dilemma.
Chimera: incorrectly merged contigs (reads) that form a chimeric scaffold (contig). In the absence of independent means for the validation of scaffold structures, it might be that a certain fraction of chimeric scaffolds is unavoidable in assembly projects.
Degradome: the complete repertoire of proteases, involved in proteolytic degradation, present in a cell.
Depth of coverage: the number of sequence reads covering a nucleotide site, often expressed as the mean across all sites in the genome. Depth of coverage is a critical parameter in population genomic analysis because the probability of obtaining reads from both alleles at a heterozygous site (i.e., to call a SNP) increases with number of reads covering that site.
Effective population size (N_e): a measure of the size of an idealized population in which the effect of genetic drift on allele frequencies is similar to the population under consideration.
Genetic architecture: the genetic background to phenotypic traits, including their number, effect sizes, and dominance.
Genomic landscape: a metaphor for the spatial distribution (along chromosomes) of parameter values of a genomic feature, such as the abundance of genes and repeats, or measures of diversity and divergence.
Genome-wide association studies (GWAS): studies based on the use of large numbers of SNP markers genotyped in a group showing a particular trait, and in a control group, with the aim of finding association between trait and markers.
Hill–Robertson interference: the effect that natural selection has on linked sites. For example, the spread of an advantageous mutation in the population can be hindered by linkage to a disadvantageous mutation on the same background. Interference decreases with increasing genetic distance to selected loci.
Linkage disequilibrium: when the association between alleles at two or more loci is not random.
Methylome: the genomic distribution of nucleotide sites modified by the addition of methyl groups by methyltransferases. Methylation of cytosines preceding guanine is the most common form of methylation in many vertebrate genomes. Cytosines can also be methylated in other sequence contexts and, in plants, targets for methylation are more promiscuous. Methylation affects transcription and, thus, is implicated in several processes of gene regulation.
N50: the length of the scaffold in a genome assembly that, when scaffolds are sorted by size, all scaffolds larger than this size contain 50% of all assembled DNA.
Nucleotide diversity (π): the average pairwise heterozygosity between two randomly drawn chromosomes from the population. At equilibrium and in the absence of selection, π should be the same as the expected value of the population genetic parameter theta estimated from the number of segregating sites.
Peptidome: the complete repertoire of translated peptides (small proteins, such as hormones) in the genome.
Phenome: the complete repertoire of the phenotypes of an organism.
Positive selection: natural selection for an advantageous allele, giving it an increased fixation probability.
RAD-tags: restriction site-associated DNA markers obtained by digesting genomic DNA with specific restriction enzymes, ligation of adaptors, amplification, and sequencing. This can reduce the complexity of genomic samples and enable sequencing of the same, targeted regions of the genome in multiple individuals. As a result, genotypes at specific SNPs can be obtained by sequencing, hence the term 'genotyping by sequencing' (GBS).
Reverse genetics: an approach that uses signals in genetic data, such as locally reduced genetic diversity arising from a selective sweeps, to elucidate the phenotypic effects of the gene or genomic region in question. This is in contrast to forward genetics, in which the starting point is a phenotype and where one seeks to track its genetic basis.
Scaled selection coefficient (γ): the selection coefficient (s, the relative fitness dis-/advantage of a derived allele) multiplied with N_e, to take into account the fact that the efficiency of selection is directly proportional to population size.
Selective sweep: natural selection for an advantageous allele that brings with it linked diversity at the haplotype background in which the advantageous allele resides (the region 'hitch-hikes' through the population).
Soft sweep: natural selection for advantageous alleles that are part of the standing genetic variation in a population (in practise, existing on different genetic background, due to recombination events). Under this scenario, the rate of adaptive evolution is not limited by the rate of supply of new mutations.
Standing genetic variation: polymorphism already existing in the population, in contrast to the appearance of new variants by mutation. Selection on standing genetic variation may, for instance, occur if environmental changes make a previously neutral variant non-neutral.
Transcriptome: the complete repertoire of transcribed sequences in the genome, including expression both from protein-coding genes and from noncoding RNAs.

More recently, there has been a rapid accumulation of genome sequences of wild species with a more or less pronounced goal of integrating genome information into studies of ecology and evolution. Some of these represent 'ecological models', for example, *Arabidopsis lyrata* (the outcrossing close relative to *A. thaliana*) [15], three-spine stickleback [7], *Heliconius* butterflies [3], and collared flycatcher [5].

Example of progress: avian genome sequences

Developments in genome sequencing of birds provide an illustrative example of how the field has progressed. Chicken, a major model organism and one that is key to global food production, was sequenced in 2004 [16], one of the first vertebrate genomes to be sequenced. In 2010, the next two avian genomes were reported; zebra finch [17], a model for studies of ethology and neurobiology, and turkey [18], a species of agricultural relevance. Subsequently, in 2012–2013, another ten avian genomes have been published, with bearing on studies of speciation and adaptation (collared flycatcher [5], rock pigeon [14], large [19] and medium ground finch [20], and ground tit [21]), conservation (Puerto Rico amazon [22], peregrine, and saker falcons [23]), or learning (budgerigar [24]). More avian genomes are in the pipeline. Moreover, the progress is representative of the technological achievements in genome sciences. Chicken and zebra finch were sequenced with Sanger technology, turkey with a combination of Sanger and next-generation technology, and the more recently derived avian genomes with high-coverage parallel sequencing platforms alone (Box 2). Given that these platforms have now been state of the art for some years and currently

Box 1. What is a genome sequence?

Most eukaryotic genomes are characterized by complex repetitive structures that are difficult, if not presently impossible, to assemble. These include interspersed repeats (transposable elements) as well as tandem arrays of similar sequence, such as in centromeres and telomeres. To this should be added the existence of sequences with unusual base composition, or other deviant structures, which tend to remain resistant to sequencing. Therefore, it is necessary to make a distinction between the DNA sequence of a genome and the DNA sequence (currently) obtainable, or obtained, by efforts toward genome sequencing. Notably, the fraction of the genome that is amendable to sequencing and assembly varies considerably among organisms. Broadly speaking, the more repetitive a genome is, the more difficult it is to assemble and this is clearly evident in the case of large plant genomes where repeats (such as long terminal repeat retrotransposons) might constitute >60% of the genome [86,87]. On top of that, for genomes that are the result of recent polyploidization events, as is the case for many plants, assembly is hampered by the existence of two or more similar copies of a significant proportion of the genome. It follows that a 'genome sequence' as it is used in the literature is not an absolute notion and, even for the most well-characterized genomes, significant parts might have yet to be included. One practical consequence of this is that the failure of finding an expected signal in a genome scan can simply be because the region in question is not included in the assembly.

Table 1. Number of sequenced eukaryotic genomes[a]

Kingdom	Phylum	Class	Number of genomes
Animalia	Annelida	Clitellata	1
		Polychaeta	1
	Arthropoda	Arachnida	5
		Branchiopoda	1
		Chilopoda	1
		Insecta	69
		Maxillopoda	1
	Chordata	Actinopterygii	1
		Amphibia	1
		Aves	11
		Mammalia	73
		Reptilia	6
		Leptocardii	1
	Tunicata	Appendicularia	2
		Ascidiacea	1
	Cnidaria	Anthozoa	2
		Cubozoa	1
		Hydrozoa	1
	Echinodermata	Asteroidea	1
		Echinoidea	2
	Hemichordata		1
	Mollusca	Bivalvia	1
		Gastropoda	2
	Placozoa		1
	Porifera	Demospongiae	1
	Platyhelminthes	Trematoda	3
		Turbellaria	1
	Nematoda	Secernentea	21
		Chromadorea	2
Fungi	Ascomycota		178
	Basidiomycota		48
	Other fungi		22
Rhizaria	Cercozoa	Chlorarachnea	1
Archaeplastida	Rhodophyta	Florideophyceae	1
		Cyanidiophyceae	2
Chromalveolata	Cryptophyta	Cryptophyceae	1
	Heterokontophyta	Bacillariophyceae	1
		Coscinodiscophyceae	2
		Eustigmatophyceae	2
		Oomycetes	12
Alveolata	Apicomplexa		20
	Ciliophora	Ciliatea	1
		Spirotrichea	6
		Oligohymenophorea	1
	Perkinsozoa	Perkinsea	1
Excavata	Euglenozoa	Kinetoplastea	13
	Percolozoa	Heterolobosea	1
Choanoflagellatea			2
Unikonta	Amoebozoa	Mycetozoa	2
	Metamonada	Parabasalia	1
Plantae	Chlorophyta	Chlorophyceae	2
		Trebouxiophyceae	1
		Trebouxiophyceae	1
		Prasinophyceae	4
	Metaphyta		62

[a]Information from National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov), April 2013.

Box 2. Genome sequencing in brief

With the extraordinary throughput provided by current technology, the generation of sequence data is no longer a bottleneck in genome sequencing. However, repetitive DNA constitutes an obstacle for approaching complete genome coverage and also affects another key aspect of genome sequencing: sequence continuity. The assembly pipeline using data generated from, for example, Illumina (HiSeq or MiSeq), Roche (454), and Life Technologies (Ion Proton/Torrent and SOLiD) platforms is typically a two-step process with algorithms for the construction of contigs and scaffolds. Contig building is at the core of shotgun sequencing and involves tiling overlapping reads from unique sequence (Figure I). During the era of Sanger-based genome sequencing, contigs were typically connected by the aid of end sequencing of large insert size clones [bacterial artificial chromosomes (BACs), cosmids, and fosmids], augmented with physical mapping of such clones (i.e., BAC fingerprinting). Merging contigs into scaffolds in high-throughput sequencing typically relies on using information from read pairs (i.e., reads from both ends of genomic fragments used for library construction; Figure I). However, when one or both reads of a fragment correspond to repetitive DNA, scaffolding is problematic and, therefore, repeat regions tend to hinder construction of continuous sequence assembly. In general, the larger the insert size of sequencing libraries, the higher the probability that unique sequence flanking repeats can be bridged. With insert sizes up to 20–40 kb, assemblies of Gb-sized vertebrate genomes currently reach a scaffold N50 of at least several Mb, sometimes more. However, there is a trade-off between tweaking the parameter settings of assembly algorithms to maximize scaffold length and to minimize the incidence of chimeric scaffolds.

Regardless of the efficiency of the scaffolding process, genome assemblies based on high-throughput sequencing data will comprise a long list of sequence segments of unknown location in the genome. Thus, the ease by which genomes can now be accessed comes with the price that assignment of contigs and scaffolds to chromosomes cannot be made without complementary information. Knowing the genomic location of sequences is essential for many applications of genomic data. There are several means for merging scaffolds. The use of optimal mapping [88] and sequencing platforms offering long reads [24] is still in its infancy, but might soon represent standard methodology in genome assembly projects. A traditional approach is to integrate assembly data with genetic linkage maps. Linkage mapping requires pedigrees, which might be difficult to establish in some non-models (but is all the more easier in others). Even a modest linkage map can anchor most scaffolds if they are large [5]. If a high-density linkage map is available, the need for complementary physical mapping approaches is essentially circumvented and allows for the amalgamation of scaffolds into close-to full chromosome sequences. Alternatively, reference-based assembly using information from related species [89] will become increasingly useful as more species are sequenced.

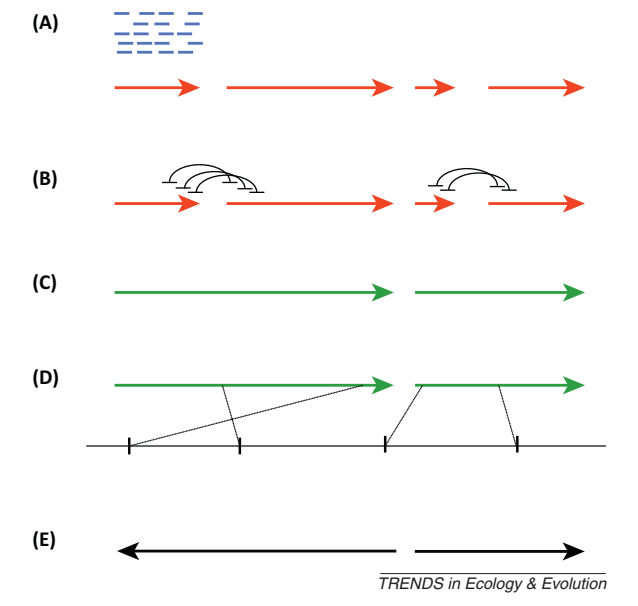

Figure I. Schematic illustration of different steps in the genome assembly process. **(A)** Overlapping short reads (blue) are merged to form contigs (red). **(B)** Read pairs (i.e., short reads from the ends of a genomic fragment) that map to two different contigs act as anchors to join the contigs into **(C)** scaffolds (green).

represent the standard approach to genome sequencing, the term 'next-generation sequencing' is becoming increasingly misplaced and will not be used herein.

Genome sequences and evolutionary genetics

Genome sequences, in contrast to sequence data from individual loci, reveal the biology of the genome and how the genetic material is organized. They show the types and abundance of transposable elements, how densely the genome is packaged with genes, and the genomic landscape of many other features, such as base composition, noncoding RNAs, chromatin marks, and nucleotide modifications. One such example is also the rate of recombination, a critical parameter in evolutionary and population genetic studies. Although recombination fractions from linkage maps have been available for many species for some time, it is not until there was to access assembled chromosome sequences that it was possible to estimate recombination rates (amount of recombination per physical unit DNA) with some accuracy and resolution [25]. A major conclusion from such studies is that the recombination landscape is often quite heterogeneous, more so than was previously thought, including small but ephemeral hot-spot regions of recombination as well as general trends of higher recombination toward chromosome ends [26]. Recombination affects the efficacy of selection by a phenomenon known as Hill–Robertson interference, which implies that selection at linked sites interferes with selection at a focal site. For example, linkage between an advantageous allele and deleterious alleles in neighboring regions hinders the spread of the favorable variant. When the recombination rate is high, genetic linkage will extend over shorter physical distances in the genome and make focal loci less vulnerable to opposing forces at other loci. An interesting consequence of this is that adaptive evolution should be more common in those regions of the genome experiencing high rates of recombination (and vice versa for regions with low rates, *cf.* nonrecombining Y and W chromosomes). Does this mean that selection for increased recombination in regions containing genes for which the encoded proteins are exposed to variable environments, such as immune defence genes? Or, does selection for rearrangements move such genes to high-recombination environments? These questions should be possible to address with data now becoming available.

The ability to obtain recombination rate estimates by combining linkage analysis and genome sequences has also provided new insight into the evolution of nucleotide composition and its links to life history. GC-biased gene

conversion is a process in which C and G nucleotides have a higher probability of being the donor during meiotic conversion events at heterozygous GC/AT sites. As a consequence, genomic regions with high recombination (gene conversion) rates should evolve towards a high GC content [27] and this has been suggested to explain the heterogeneous landscape of base composition seen in many organisms [28,29]. If the recombination landscape remains stable over evolutionary timescales, which can be expected if the karyotype is evolutionarily stable, such as in birds [30], the build up of a heterogeneous landscape of base composition should be particularly pronounced [31]. Moreover, the effect of GC-biased gene conversion should be stronger in large populations because, although being a neutral process, it behaves similarly to selection, in the sense that nonrandom fixation probabilities increasingly override the effects of genetic drift as the effective population size (N_e) grows large [28,29]. Another interesting but not yet fully explored consequence of biased gene conversation is that it has the potential to reduce the efficacy of purifying selection in high recombination regions by aggravating the removal of deleterious AT/CG mutations [32].

Comparative genomics and molecular evolution

The access to genome sequences from multiple species has brought the field of molecular evolution to a level where inferring the evolutionary processes affecting sequence evolution is increasingly done from a whole-genome perspective, rather than from the pattern seen in a random sample of loci. Besides providing a more complete picture of sequence evolution, this has had the advantage of enabling studies of the genomic variation in patterns generated by relevant processes, such as the intensity and character of selection. Undoubtedly, one of the most important findings made possible by access to genome sequences from more than a limited number of model organisms relates to the quantification of the proportion of the genome evolving under purifying selection. This can be revealed by the identification of sequences conserved beyond neutral expectations for the accumulation of mutations with no effect of fitness in alignments of multiple species. A study analyzing the sequence of 29 mammalian genomes concluded that approximately 5% of the human genome is constrained with respect to sequence evolution, with a key finding that approximately 70% of the constrained sequence is not associated with protein-coding transcripts, but is instead located in introns and intergenic DNA. Such noncoding sequences include specific chromatin regulators, RNA species, and other regulatory motifs. However, a significant proportion of conserved mammalian sequence still remains to be annotated.

Recently, a heated debate has arisen over the fraction of functional sequence in the human genome because estimates based on annotation are higher (70–80% of the genome) than estimates based on evolutionarily conserved sequence (<10%). The ENCODE project [33], representing the former camp, uses the concept of function and suggests that features such as transcribed sequence (with the majority representing nonprotein-coding sequence), histone modification, and open chromatin are generally indicative of function. However, this has been criticized, most strongly articulated by Graur and colleagues [34], with the argument that unless the action of purifying selection can be demonstrated, and power to detect selection is not a limiting factor, then functionality must be questioned. For an evolutionary biologist, this criticism makes sense; quoting [35]: 'Unless a genomic functionality is actively protected by selection, it will accumulate deleterious mutations and will cease to be functional.' Although taxon specific, the human data are likely to have general relevance to evolutionary biology and similar studies are starting to appear for other groups of organisms, such as plants [36]. However, for weakly selected mutations, it must be kept in mind that they might be effectively neutral in populations with small N_e, as in humans, and hence do not contribute to sequence conservation. When scaled selection coefficients (γ) are higher, one might expect to see larger fractions of the genome conserved, which is indeed the case in the *Drosophila* genus [35].

Genomes reveal lineage-specific adaptations

Learning about the biology of genomes also has a more direct bearing on evolution. Many genome-sequencing projects of non-models have led to the identification of lineage-specific genes likely underlying evolutionary novelty at the phenotypic level. As exemplified for the recently sequenced animal and plant genomes listed in Table 2, a common denominator of several of these studies has been the observation of expanded gene families implicated in lineage-specific adaptations. Some prominent examples include expansion of tomato genes involved in the modification of cell wall architecture and thereby fruit development and ripening [4], expansion of domains related to hypoxic stress in yak as an adaptation to high altitudes [9], and expansion of bile salt-stimulated lipases, capable of hydrolyzing triglycerides into monoglycerides and subsequently releasing digestible free fatty acids, in hibernating bats [12]. That gene duplication and neofunctionalization can be key steps in functional divergence is an old idea [37]. However, it is not until the access to complete genomes that one can now conclude that they represent common mechanisms for the formation of new phenotypes and thereby a seedbed for adaptation. Indeed, using an existing gene as template for a new gene, by duplication followed by modification, is likely to be a faster route towards a new functional protein than the evolution of a new gene from random, noncoding sequence.

Genomic support for lineage-specific adaptations is also interpreted from genome-wide scans for positively selected genes (Box 3). In bats, for instance, positive selection of genes involved in repairing lesions in DNA is seen as a response to the negative consequences of radical oxygen species, which have a damaging effect on DNA, generated by high metabolism during flight [12]. In yaks, adaptive evolution of several genes involved in the cellular response to hypoxia has been documented [9]. These two examples illustrate that adaptation at the genomic level is probably a combination of the acquisition of new genes (see above) and modification of existing genes driven by positive selection. The rapidly accumulating number of genomes sequenced will be important for the detection of adaptive evolution

Box 3. Inferring adaptive evolution from genomic data

Positive selection implies an increased fixation probability for advantageous alleles. Therefore, the rate of evolution at functionally important sites under positive selection should be higher compared with the situation for neutral nucleotide sites. By contrast, when evolving under the influence of purifying selection, the rate should be lower than at neutrality. If synonymous sites are taken as a neutral reference, an indication of the strength of selection at nonsynonymous and potentially functional sites can be obtained by the ratio of the substitution rates at the two categories of sites (d_N/d_S). If this ratio, often referred to as ω, is larger than 1, positive selection is inferred; a popular software, PAML, uses a likelihood ratio test approach for making statistical inference [90]. Given that the test can have limited power when averaging rates over sites in a protein (because all sites in a protein might be unlikely to evolve in an adaptive manner), so-called 'branch-site' models can be applied that only consider a subset of codons.

Additional power to detect positive selection can be obtained by combining data on substitutions and polymorphisms. If the ratio of the number of nonsynonymous substitutions to the number of nonsynonymous polymorphisms is higher than the corresponding ratio at synonymous sites, then positive selection can be inferred. The rationale behind this test is that advantageous mutations can quickly sweep through the population to reach fixation so that observed nonsynonymous polymorphisms should mainly reflect largely neutral variants. Developments of this well-known McDonald–Kreitman (MK) test [91] include calculations of the neutrality index (the odds ratio from the MK table [92]) and the recently presented measure 'direction-of-selection' (DoS) [93]. Moreover, derivates of the MK table [40,41] applied to sequence data from large number of genes enable one to address quantitatively the overall extent of adaptive evolution in coding sequences, expressed as the proportion of nonsynonymous substitutions driven to fixation by positive selection (α).

because the power to make inference on positive selection typically increases with increasing number of sequences aligned.

Evolvability and the rate of adaptation are key concepts in evolutionary biology. Genome sequencing also offers a route towards quantifying the overall role of adaptive evolution, and how this varies among lineages and is related to life history. By contrasting the rate at which nucleotide substitutions that are likely to have functional consequences accumulate with the rate of presumably neutral substitutions (Box 3), genome-wide estimates of the incidence of adaptive evolution have recently been made. Expressed as the proportion of functionally relevant substitutions driven to fixation by positive selection, estimates vary from close to zero (human [38] and selfing plants [39]) to approximately 50% (e.g., *Drosophila* [40]). Intuitively, it should be possible to explain this variation by an expected positive correlation between the incidence of adaptive evolution and N_e, under a scenario of adaptive evolution being limited by the supply of new mutations (rather than mainly acting on standing genetic variation). In large populations, selection for advantageous mutations is more efficient both because γ is higher for any given value of the selection coefficient (s) and because, for any given value of γ, lower values of the selection coefficient (s) are effectively selected [41]. However, is it realistic that few functional variants evolve adaptively in populations with small N_e? The answer probably lies partly in that adaptive evolution is difficult to estimate in small populations because slightly deleterious mutations are more likely to be effectively neutral. They will thereby contribute to divergence such that estimates of adaptive evolution are impaired. Moreover, it has been suggested that N_e is not necessarily a strong predictor of adaptive evolution in a model of recurrent environmental changes and alternating periods of adaptive walks and stasis with purifying selection dominating [42]. Furthermore, as pointed out in [41], a higher incidence of adaptive evolution does not necessarily translate into faster adaptation if the effect size of substitutions in large populations tends to be smaller in magnitude.

Population genomics

The term 'population genomics' started to appear in the literature from the late 1990s, mainly in the context of large-scale polymorphism analyses in humans. Approximately 10 years ago, biologists began to foresee that large-scale population genetic approaches would be both feasible and important for studies of natural populations [43,44]. Since then, the use of sequence or genotype data from multiple, although individually analyzed, loci spread across the genome has often been referred to as population genomic analyses. With the generation of sequence data no longer representing a bottleneck in genome analyses, a logical step following from the access to genome assemblies is whole-genome resequencing of population samples from species with an assembled genome. This provides the necessary platform for analyses of genome-wide polymorphism data, that is, population genomics in its true sense [45]. Importantly, population genomics is not only a matter of scaling up to increase power for making inference about population processes, but also offers a means to study the genomic landscape and variance of allelic diversity within and between populations.

For the rest of this review, I concentrate on the outcome and potential of population genomic analyses based on whole-genome resequencing data. Genome-wide yet intermediate-scale approaches to population genomics have been covered elsewhere, including genotyping-by-sequencing (RAD-tag sequencing [46,47]), exome sequencing [48], and transcriptome sequencing [49]. As an introductory, cautionary note to the work to be presented, the '$n = 1$ constraint' in population genomics should be kept in mind [50]; most studies concern a single instance of the outcome of evolution.

Methodological aspects.

A typical pipeline for a population genomic study has a few critical steps: (i) design of sequencing strategy; (ii) generation of sequence data; (iii) mapping of sequence reads to the assembly; (iv) variant calling (genotyping); and (v) downstream population genetic or molecular evolutionary analyses. Sequencing strategy includes aspects such as the depth of coverage and whether individually tagged samples or pools of individuals have been used, as well as issues common to any population genetic study (for instance, number of individuals per population and number of populations, gender, and the need for outgroup species). Data from pools sequenced at high depth can be used to estimate directly population allele frequencies based on the relative abundance of reads with alternative alleles [51,52]. However, because it can be difficult to obtain equimolar amounts of DNA from all individuals in a pool

Table 2. Examples of key findings from recently derived genome sequences from animals and plants

Common name	Latin name	Finding	Refs
Yak	*Bos grunniens*	Adaptation to life at high altitudes has been accompanied by expansion of gene families related to hypoxic stress	[9]
Tree shrew	*Tupaia belangeri*	Loss of the gene encoding prostate-specific transglutaminase 4, which is involved in the formation or dissolution of seminal coagulum, might be related to low levels of sperm competition in this group	[94]
Bears	*Ursus* sp.	Revealed largely independent evolutionary histories of an enigmatic species trio (black, brown, and polar bear), but with admixture giving footprints of alternative histories in different parts of genome	[77]
Anole lizard	*Anolis carolinensis*	Homogeneous genomic landscape of base composition, unlike the 'isochore' structure of other amniote genomes	[95]
Peregrine and saker falcon	*Falco peregrinus* and *Falco cherrug*	Bone morphogenetic protein 4 (Bmp4) exonization and duplication of two genes implicated in avian beak morphology might explain adaptation to a predatory life style	[23]
Ground tit	*Pseudopodoces humilis*	Expansion of gene families implicated in energy metabolism and potentially related to a high-altitude life style of this species	[21]
Turkey	*Meleagris gallopavo*	Expansion of keratin gene family, which is a major component of avian feather and claws	[18]
Green sea and soft-shell turtle	*Chelonia mydas* and *Pelodiscus sinensis*	The enigmatic position of turtles within aminotes seems resolved with genome-wide phylogenomic analysis: turtles are a sister group to birds and crocodilians, with an estimated divergence time of 257 million years ago. Lizards are an outgroup to these lineages	[96,97]
African coelacanth	*Latimeria chalumnae*	Gene losses associated with vertebrate transition from water to land, including loss of genes involved with, for example, fin, otolith, and ear development. Lack of immunoglobulin M (IgM) indicates an immune system operating differently from other vertebrates	[98]
Sea lamprey	*Petromycon marinus*	Identification of lamprey genes only shared with gnathostomes reveals genetic innovations that emerged at base of vertebrate evolution; includes functions related to myelination and neuropeptide and neurohormone signaling that are characteristic to vertebrate central nervous system	[99]
Three-spine stickleback	*Gasterosteus aculeatus*	Inversions distinguish marine and freshwater ecotypes	[7]
Platyfish	*Xiphophorus maculatus*	High retention of genes implicated in cognition after teleost genome duplication might explain behavioral complexity in fishes	[100]
Moth	*Plutella xylostella*	Expansion of gene families used in detoxification of plant defense compounds	[11]
Postman butterfly	*Heliconius melpomene*	Visual complexity facilitated by expression of a duplicate ultraviolet opsin. Extensive expansion of chemosensory genes	[3]
Monarch butterfly	*Danaus plexippus*	Changes in gene repertoire behind formation and function of visual input into sun compass system	[101]
Pacific oyster	*Crassostrea gigas*	Expansion of genes encoding inhibitors of apoptosis and heat shock protein 70, involved in protection of cells against heat and other stresses, might be central for ability of oysters to tolerate prolonged air exposure	[13]
Tapeworms	Several genus	Lack of ability to synthesize fatty acids and cholesterol *de novo* is compensated by ability to scavenge essential fats from novel fatty acid proteins	[102]
Owl limpet, a polychaete annelid, and a leech	*Lottia gigantea*, *Capitella teleta*, and *Helobdella robusta*	Phylogenomic analysis supports tripartite view of bilaterians and the monophyly of annelids, molluscs, and platyhelminthes	[10]
Foxtail millet	*Setaria italica*	Identification of pathways for photoperiod-induced flowering time	[1]
Bread wheat	*Triticum aestivum*	Insight into origin of hexaploid bread wheat genome from diploid ancestors. Expansion of gene families associated with defense, nutritional content, energy metabolism, and growth might be the result of domestication	[2]
Potato	*Solanum tuberosum*	Expansion of Kunitz protease inhibitor gene family potentially involved with resistance to biotic stress in root tubers	[103]
Tomato	*Solanum lycopersicum*	Two genomic triplications have set the stage for evolutionary novelties by neofunctionalization. Expansion of gene families involved in modification of cell wall architecture and thereby fruit development and ripening, provides an example	[4]
Lyrate rockcress	*Arabidopsis lyrata*	The larger genome of this outcrossing species compared with its close selfing relative *Arabidopsis thaliana* suggests pervasive selection for genome shrinking during transition to selfing	[15]
Cotton	*Gossypium raimondii*	Extreme genetic complexity resulting from five- to sixfold ploidy increase followed by allopolyploidization. A derived ability to produce defense terpenoids, such as gossypol, by the evolution of a new family of cadinene synthases	[104,105]

Table 2 (*Continued*)

Common name	Latin name	Finding	Refs
Sweet orange	*Citrus sinensis*	Genome sequence comparisons suggest that sweet orange originated from a backcross hybrid between pummelo and mandarin	[106]
Peach	*Prunus persica*	Expansion of gene families involved with sorbitol metabolism (sorbitol transporters and dehydrogenases) has contributed to the sweet taste	[107]
N/A	*Thellungiella salsuginea*	Compared with *A. thaliana* (divergence time 7–12 million years ago), the evolution of new genes in functional categories, such as 'response to salt stress', 'osmotic stress', and 'water deprivation', is likely related to the high salinity- and drought-tolerant phenotype of this species	[108]
Norway spruce	*Picea abies*	Despite the >100 times larger genome size of spruce than the main plant model organism *A. thaliana*, the number of genes in the two genomes is about the same. The large genome size (20 Gb) of spruce, and other conifers, seems to be the result of an accumulation of transposable elements	[87]

and due to stochastic variation in the amplification efficiency of individual DNAs, two features that can bias the occurrence of different alleles, there can be low confidence in such allele frequency estimates. For this reason and, importantly, because of the benefits in downstream analyses of obtaining genotypes, most studies use individually tagged samples.

Deciding on the depth of coverage is partly a matter of how large a proportion of the genome one aims to obtain data from. For example, for a 1-Gb genome resequenced at 1× coverage (i.e., each site covered on average by one read), approximately 70% of an individual genome is expected to be covered by at least one read. With ten individuals sequenced to the same depth and with the likelihood for sites to be sequenced approximately constant across the genome, data from <3% of the genome might be expected to be obtained from all ten individuals. To obtain higher genomic representation, genomes have to be sequenced to higher depth.

Depth of coverage is also relevant for the scoring of polymorphism and, intuitively, the confidence by which this can be done increases with sequencing depth. Much methodological work has recently been devoted to this critical step in population genomics [53–56], with likelihood approaches finding increased use [57,58]. Methods that bypass genotyping and use site-frequency spectra show great promise for downstream analyses for which retention of rare alleles are critical; the software ANGSD (http://popgen.dk/wiki/index.php/ANGSD) might be a particularly useful tool. As a trade off between confidence in genotyping and cost reduction, and between calling too many single nucleotide polymorphism (SNPs) and undercalling heterozygotes, several recent studies of 'large' genomes have used 5–10 × coverage in resequencing efforts [5,59,60].

Reverse genetics by selective sweep mapping.

As soon as a reasonably large number of genetic markers became available (e.g., microsatellites), population geneticists started to make genome scans for local reduction in genetic diversity as a means to detect selection based on the concept of genetic hitchhiking (selective sweep mapping) [61]. However, because the signal of selection can be restricted to small regions of the genome, it has not been possible to exploit the full potential of sweep mapping until recently. Genome scans have the potential to uncover those regions of the genome that have been subject to recent selection, as manifested in, for example, reduced nucleotide diversity, extended linkage disequilibrium, or runs of homozygosity. The beauty of this approach lies in that traits that are important to adaptation need not be defined *a priori*, providing an unbiased reverse genetic approach to understanding the genetic architecture behind adaptive evolution.

Sweep mapping based on resequencing has been successfully applied in studies of many domestic animals and plants [6,60,62–68]. This is probably because strong artificial selection leaves distinct footprints in the genome when alleles are rapidly brought to fixation, more so than results from selection in natural populations. However, even in the case of soft sweeps, reduced local levels of diversity should be detectable given the unprecedented resolution provided by whole-genome resequencing, especially when augmented by other population genetic analyses. Inferring selection from levels of nucleotide diversity requires formulation of a null model that takes into account how much variation in diversity is expected by neutral processes alone (e.g., stochastic variation in coalescence and mutation rate variation) and by the extent of Hill–Robertson interference due to recombination rate variation. Composite likelihood methods are useful and robust, and can provide estimates of the strength of selection [69,70].

Trait mapping

Genome resequencing has the potential to replace marker-based approaches in genome-wide association studies (GWAS) aimed at finding loci underlying phenotypic traits. Using cohorts of extreme phenotypes, the whole genome can be scanned for enrichment of certain alleles at causative loci [71]. This approach was recently used in connection with sequencing of the pigeon genome [14]. Resequencing of several breeds with and without the head crest phenotype, a tuft of elongated feathers atop the head that is a common ornament in many birds, led to the identification of a particular allelic variant in the gene encoding the ephrin receptor B2 present in all crested breeds, but absent in all others. Sample sizes here were smaller than necessary in GWAS, a consequence of the fact that the whole genome was targeted instead of a select set of SNPs included in genotyping arrays.

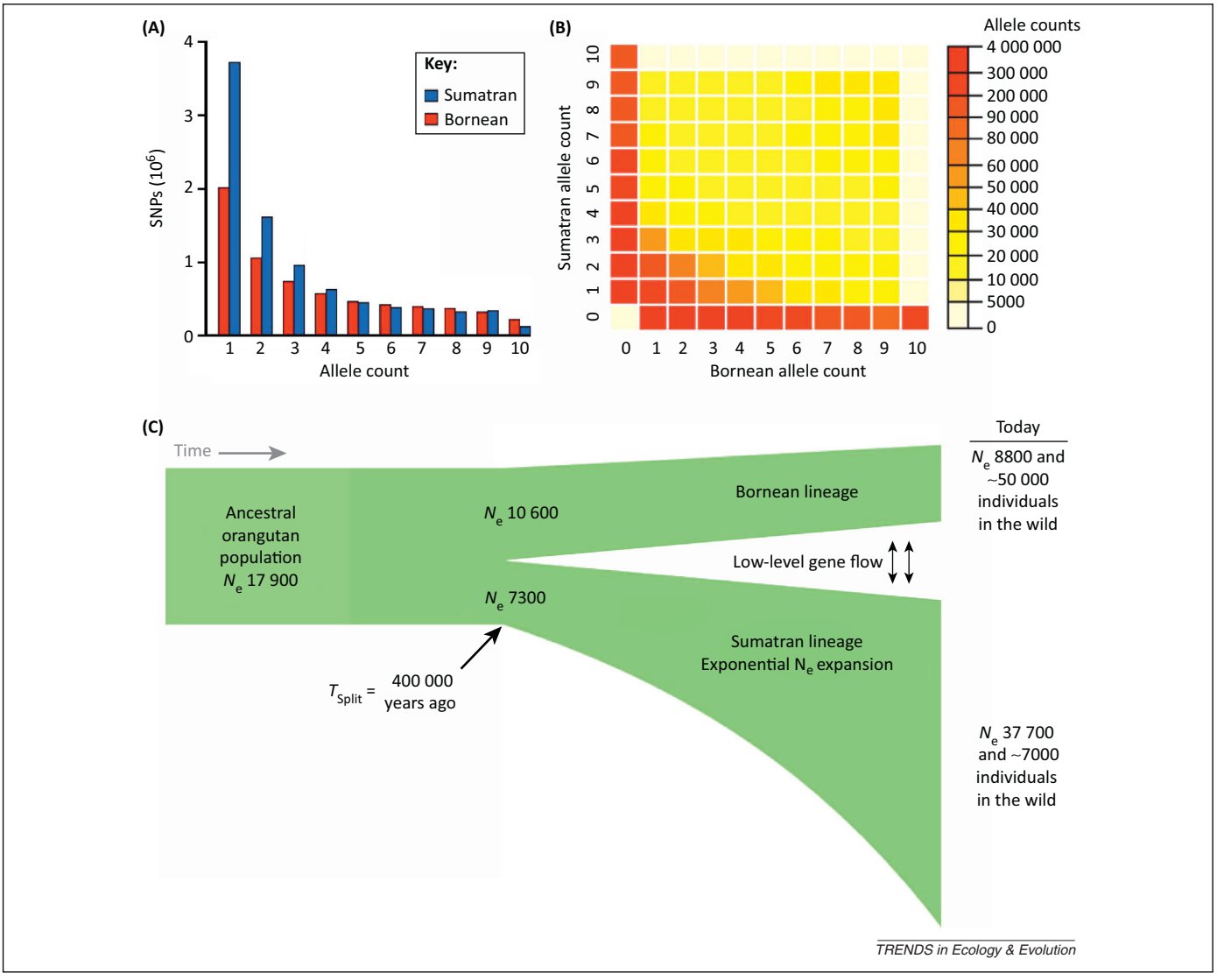

Figure 1. Demographic inference of orangutan populations based on whole-genome resequencing data [75]. **(A)** Unfolded allele frequency spectra based on the number of allele counts among ten chromosomes of each population. The Sumatran population has a higher proportion of rare alleles, a pattern expected under recent population expansion. **(B)** The typical heat map used for likelihood approaches applied to 2D allele frequency spectra for diverging populations. The concentration of observations in cells along each axis coupled with only some observations along the diagonal in the lower left corner indicate largely isolated populations of Sumatran and Bornean orangutans, with limited gene flow. **(C)** A summary demographic model that depicts relatively recent (400 000 years ago) divergence and a significant expansion in Sumatra following the split. However, despite this expansion and a more than fourfold higher effective population size (N_e) in Sumatra than in Borneo, the census size of Bornean orangutans is nearly tenfold higher than that of the Sumatran population. Reproduced, with permission, from [75].

Related to trait mapping is the analysis of groups of independent populations living in similar environments. If adaptation to these environments evolved in parallel, with the same genes or genomic regions involved in independent populations, then populations sharing habitat might show higher genetic similarity in those regions than in the rest of the genome. This approach was taken in studies of multiple populations of marine and freshwater three-spine sticklebacks, using whole-genome resequencing data [7]. The approach was able to recover successfully the *EDA* locus known to be associated with repeated armor evolution as well as several other regions potentially involved in ecotypic differentiation.

Demography, population divergence, and speciation

Patterns of genetic diversity within and between populations are shaped by demography, differentiation, and the extent of reproductive incompatibility. Whole-genome polymorphism data offer the promise of revealing complex demographic scenarios and assessing to what extent gene flow and introgression affect the character of genetic diversity [72]. The perhaps most important aspect here is that with data available from across the genome, it is possible analyze whether certain genomic regions have been less prone (or particularly prone) to gene flow than others, and then being able to ask why this has been so. One likelihood model-based approach for demographic inference (e.g., as integrated in the program δaδi) uses diffusion approximation to the allele frequency spectra of diverging populations [73]. A suite of other approaches (reviewed in [72]) is based on sampling genealogies and calculation of the likelihood for different models in a coalescence framework. Approximate Bayesian computation (ABC) has become increasingly popular in this context: it bypasses exact likelihood calculation by using summary statistics to characterize patterns of variation observed in

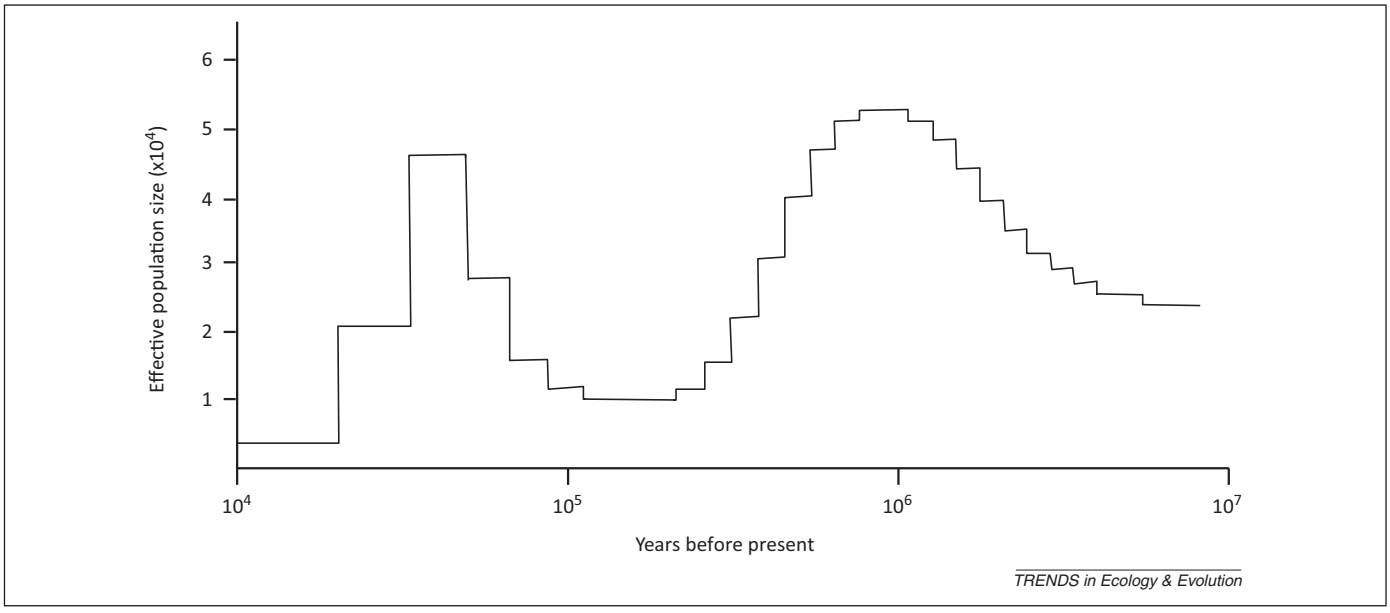

Figure 2. Changes in effective population size (N_e) of the giant panda and its ancestors according to estimates based on the pairwise sequentially Markovian coalescent model. For most of the time period inferred, fluctuations in population size are consistent with changes in climate. Accordingly, population declines coincide with the last and the penultimate glacial maxima. However, in the very recent past, severe population contraction is the result of negative effects of anthropogenic activities. Note that the model has no resolution for <20 000 years ago. Redrawn from [59].

the data [74]. Inference of the population history of Bornean and Sumatran orangutans is shown in Figure 1 as an example of the use of allele frequency spectra for demographic analyses [75].

An exciting recent development is a coalescent-based hidden Markov model (pairwise sequentially Markovian coalescent; PSMC [76]) that uses the local density of heterozygous sites in diploid data for inferring past demography. Thus, with deep coverage, genome sequence data from a single individual are sufficient to make demographic inference [23,77]. PSMC has been applied to genome sequence data from falcons [23], bears [77], flycatchers [78], and pandas [59], with informative results (Figure 2). The PSMC approach has poor resolution for the more recent past because the number of coalescence events is then limited. However, when extended to include data from multiple genomes, more recent demography can also be inferred [79].

Just as genome scans for regions with low diversity within populations are indicative of recent selective sweeps, scans for regions with high divergence between populations point to diversifying selection and limited gene exchange. With increasing power and resolution provided by developments in genomic technology, the field of speciation genetics is witnessing a formidable explosion. As recently reviewed elsewhere [80,81], there is an emerging picture of reproductive isolation under a model of ecological speciation beginning at loci involved in local adaptation. Divergence hitchhiking around such regions can build up 'islands' of the genome in which lineage sorting is becoming increasingly complete by being protected from gene flow at hybridization. Most genomic data illuminating these processes have so far been based on RAD-tag sequencing [3,82,83]; however, one should expect to see an increasing number of studies using whole-genome resequencing because, in principle, this is technically more straightforward and can provide more complete pictures. In a recent study of *Ficedula* flycatchers based on the resequencing of population samples of collared flycatcher and pied flycatcher (Figure 3), 'genome islands of divergence' were detected on essentially all chromosomes and coincided with the location of centromeres and telomeres, potentially pointing to a role of meiotic drive processes in species divergence [5].

Prospects

Progress in genome analyses of non-model organisms has benefitted from developments made in the research tools applied to model organisms, including humans. Just as this has been the case when it comes to genome sequencing, one should in coming years expect to see other advancements in the footsteps of current trends. Large-scale analyses of the proteome would be one such example, studies of nucleotide modifications and their role in, for example, nongenetic inheritance, would be another. Moreover, new ways of approaching questions in ecology and evolution could be anticipated from the ability to generate vast amounts of sequence data. For instance, phylogenomics is still in its infancy when it comes to using whole-genome sequence data [84] and will benefit from targeted genome-sequencing efforts of critical taxa. Of general relevance in this context is the characterization of incongruences between phylogenetic trees constructed from sequences in different genomic regions (gene tree–species tree conflicts). This should enable quantitative analyses of the role of incomplete lineage sorting and hybridization in ancient speciation events. The observations that parts of the human genome are more similar to those of gorilla than to chimpanzees [85] and that parts of the chimpanzee genome is more similar to those of humans than to bonobo [8], provide inspiring examples of unexpected findings.

Another approach that specifically can benefit studies of ecology and evolution is the use of multiple genomes from related species for tracing the origin of adaptive traits. By

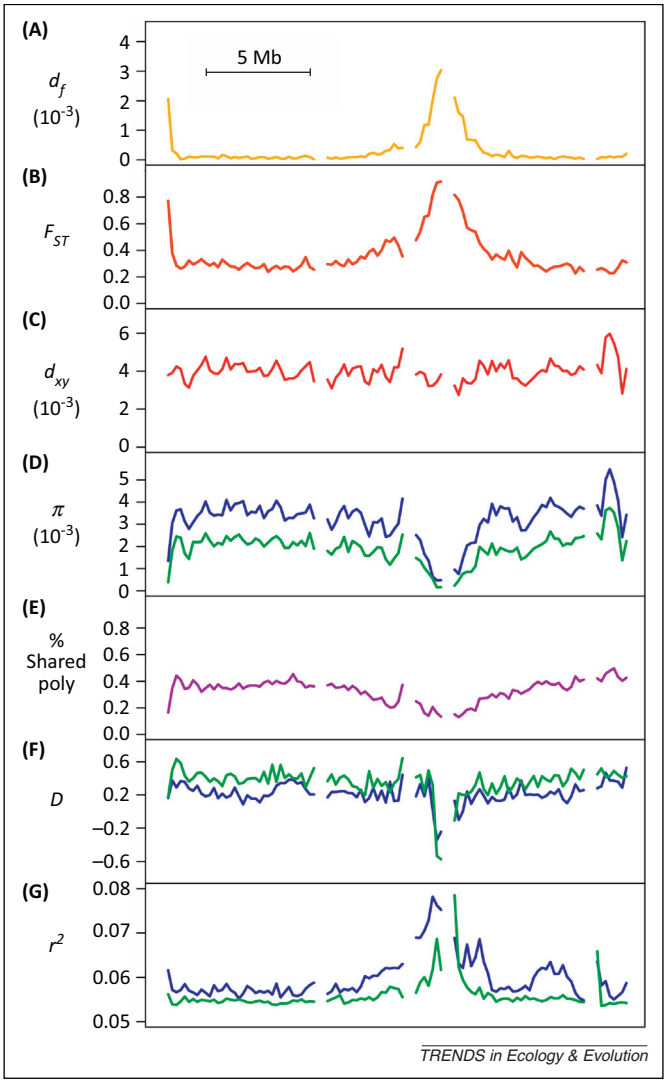

Figure 3. Population genomic analyses of collared flycatcher (blue) and pied flycatcher (green) chromosome 4A based on whole-genome resequencing data (200 kb windows). **(A)** and **(B)** (yellow and red) are between-species divergences estimated by the density of fixed differences (d_f) and the fixation index (F_{ST}), respectively. **(C)** d_{xy}, the total pairwise divergence between chromosomes from the two species. **(D)** nucleotide diversity (π) of each species. **(E)** The proportion of shared polymorphisms among all polymorphic sites. **(F)** and **(G)** show Tajima's D and linkage disequilibrium as estimated by r^2, respectively. Together, these results point to two 'divergence peaks' in this chromosome, one at the left terminal end and one at a position at approximately 12 Mb. Divergence peaks are characterized by reduced nucleotide diversity and shared polymorphism, negative Tajima's D, and extended linkage disequilibria. The fact that both species show reduced diversity in peak regions is unexpected under a scenario of divergent selection by local adaptation in one of the populations. Note that the total pairwise divergence between species is not elevated in divergence peaks, a consequence of the fact that increased levels of fixed differences between species are balanced by lowered levels of diversity within species. Adapted, with permission, from [5].

mapping substitutions onto a phylogeny of species, one can pinpoint in which (internal or terminal) node adaptive evolution has taken place, which is critical for understanding the connection between evolution at genetic and phenotypic level. As an example, genome sequencing in zebra finch revealed a highly enriched fraction of positively selected ion channel genes (e.g., glutamate receptors) that respond to song exposure in the auditory forebrain and might explain the derived trait of vocal learning in songbirds [17]. However, this inference was made from alignments of zebra finch with chicken as the only other bird species (plus several nonavian outgroup species). Given that the Neoaves lineage leading to zebra finch contains some 20 avian orders, of which most are not capable of vocal learning, finding that a critical substitution occurred in the early songbird lineage would strengthen the connection between positive selection of genes involved in neural processes and this trait.

Concluding remarks

The number of sequenced genomes is accumulating at a faster than ever rate, with no signs of deceleration. It would not come as a surprise if most ecologists and evolutionary biologists were to have access to the genome sequence of their study organisms in the not too distant future. There are several take-home messages from this review of what genome sequences of non-model organisms have so far informed about evolution. For example, a picture of heterogeneously distributed recombination events across the genome has emerged and this, in turn, may generate heterogeneous landscapes of base composition and adaptive evolution. Up to 50% or more of substitutions changing the amino acid sequence of proteins is estimated to have been driven to fixation by positive selection in large populations, giving a quantitative measure of adaptive evolution at the protein level. However, adaptive evolution also seems to be due frequently to the expansion of gene families, coupled with acquisition of new function in new copies. For purifying selection, comparisons of genome sequences from multiple species have revealed that a larger proportion of the genome than was previously thought evolves under constraint (although the precise amount is debated). By large-scale resequencing of assembled genomes in population samples, the field of population genomics is becoming an exciting venue for the identification of genes and genomic regions involved in, for example, fitness-related traits and speciation. This is nicely illustrated by successful sequencing-based rather than marker-based GWAS mapping and accumulating evidence for distinct divergence islands within a background environment of low genomic differentiation during the speciation process.

Acknowledgments

I am grateful to members of my lab group and Jochen Wolf's lab group for helpful discussions and to Christen Bossu for comments on the manuscript. This work was supported by an Advanced Investigator Grant (NEXTGENMOLECOL) from the European Research Council, a Wallenberg Scholar Award from the Knut and Alice Wallenberg Foundation and from the Swedish Research Council (2007-8731 and 2010-5650).

References

1 Bennetzen, J.L. *et al.* (2012) Reference genome sequence of the model plant *Setaria*. *Nat. Biotechnol.* 30, 555–561
2 Brenchley, R. *et al.* (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491, 705–710
3 Heliconius Genome Sequencing Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487, 94–98
4 Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641
5 Ellegren, H. *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* 491, 756–760
6 Guo, S. *et al.* (2013) The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nat. Genet.* 45, 51–58

7 Jones, F.C. *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, 55–61
8 Prufer, K. *et al.* (2012) The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486, 527–531
9 Qiu, Q. *et al.* (2012) The yak genome and adaptation to life at high altitude. *Nat. Genet.* 44, 946–949
10 Simakov, O. *et al.* (2013) Insights into bilaterian evolution from three spiralian genomes. *Nature* 493, 526–531
11 You, M. *et al.* (2013) A heterozygous moth genome provides insights into herbivory and detoxification. *Nat. Genet.* 45, 220–225
12 Zhang, G. *et al.* (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science* 339, 456–460
13 Zhang, G. *et al.* (2012) The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490, 49–54
14 Shapiro, M.D. *et al.* (2013) Genomic diversity and evolution of the head crest in the rock pigeon. *Science* 339, 1063–1067
15 Hu, T.T. *et al.* (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* 43, 476–481
16 ICGSC (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432, 695–716
17 Warren, W.C. *et al.* (2010) The genome of a songbird. *Nature* 464, 757–762
18 Dalloul, R.A. *et al.* (2010) Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol.* 8, e1000475
19 Rands, C. *et al.* (2013) Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics* 14, 95
20 Zhang, G. *et al.* (2012) The genome of Darwin's finch (*Geospiza fortis*). *GigaDB* http://dx.doi.org/10.5524/100040
21 Cai, Q. *et al.* (2013) The genome sequence of the ground tit *Pseudopodoces humilis* provides insights into its adaptation to high altitude. *Genome Biol.* 14, R29
22 Oleksyk, T. *et al.* (2012) A locally funded Puerto Rican parrot (*Amazona vittata*) genome sequencing project increases avian data and advances young researcher education. *Gigascience* 1, 14
23 Zhan, X. *et al.* (2013) Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat. Genet.* 45, 563–566
24 Koren, S. *et al.* (2012) Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* 30, 693–700
25 Tortereau, F. *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics* 13, 586
26 Smukowski, C.S. and Noor, M.A.F. (2011) Recombination rate variation in closely related species. *Heredity* 107, 496–508
27 Duret, L. and Galtier, N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* 10, 285–311
28 Lartillot, N. (2013) Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol. Biol. Evol.* 30, 489–502
29 Romiguier, J. *et al.* (2010) Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res.* 20, 1001–1009
30 Ellegren, H. (2010) Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol. Evol.* 25, 283–291
31 Mugal, C.F. *et al.* (2013) Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Mol. Biol. Evol.* 30, 1700–1712
32 Webster, M.T. and Hurst, L.D. (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet.* 28, 101–109
33 Dunham, I. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74
34 Graur, D. *et al.* (2013) On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol.* 5, 578–590
35 Bergman, C.M. and Kreitman, M. (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11, 1335–1345
36 Hupalo, D. and Kern, A.D. (2013) Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol. Biol. Evol.* 30, 1729–1744
37 Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag
38 Zhang, L. and Li, W-H. (2005) Human SNPs reveal no evidence of frequent positive selection. *Mol. Biol. Evol.* 22, 2504–2507
39 Gossmann, T.I. *et al.* (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. *Mol. Biol. Evol.* 27, 1822–1832
40 Smith, N.G.C. and Eyre-Walker, A. (2002) Adaptive protein evolution in *Drosophila*. *Nature* 415, 1022–1024
41 Gossmann, T.I. *et al.* (2012) The effect of variation in the effective population size on the rate of adaptive molecular evolution in eukaryotes. *Genome Biol. Evol.* 4, 658–667
42 Lourenco, J.o.M. *et al.* (2013) The rate of molecular adaptation in a changing environment. *Mol. Biol. Evol.* 30, 1292–1301
43 Black, W.C., IV *et al.* (2001) Population genomics: genome-wide sampling of insect populations. *Ann. Rev. Entomol.* 46, 441–469
44 Luikart, G. *et al.* (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat. Rev. Genet.* 4, 981–994
45 Begun, D.J. *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5, e310
46 Davey, J.W. *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510
47 Narum, S.R. *et al.* (2013) Genotyping-by-sequencing in ecological and conservation genomics. *Mol. Ecol.* 22, 2841–2847
48 Li, Y. *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* 42, 969–972
49 Nuzhdin, S.V. *et al.* (2012) Genotype-phenotype mapping in a post-GWAS world. *Trends Genet.* 28, 421–426
50 Buerkle, C.A. *et al.* (2011) The n = 1 constraint in population genomics. *Mol. Ecol.* 20, 1575–1581
51 Boitard, S. *et al.* (2012) Detecting selective sweeps from pooled next-generation sequencing samples. *Mol. Biol. Evol.* 29, 2177–2186
52 Cheng, C.D. *et al.* (2012) Ecological genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* 190, 1417–1432
53 DePristo, M.A. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43, 491–498
54 Le, S.Q. and Durbin, R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* 21, 952–960
55 Altman, A. *et al.* (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. *Hum. Genet.* 131, 1541–1554
56 Nielsen, R. *et al.* (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* 12, 443–451
57 Nielsen, R. *et al.* (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE* 7, e37558
58 Wang, Y. *et al.* (2013) An integrative variant analysis pipeline for accurate genotype/haplotype inference in population NGS data. *Genome Res.* 23, 833–842
59 Zhao, S. *et al.* (2013) Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nat. Genet.* 45, 67–71
60 Varshney, R.K. *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* 31, 240–246
61 Schlötterer, C. (2002) A microsatellite-based multilocus screen for the identification of local selective sweeps. *Genetics* 160, 753–763
62 Groenen, M.A.M. *et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature* 491, 393–398
63 Rubin, C-J. *et al.* (2012) Strong signatures of selection in the domestic pig genome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19529–19536
64 Axelsson, E. *et al.* (2013) The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* 495, 360–364

65 Larkin, D.M. *et al.* (2012) Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7693–7698
66 Jiao, Y. *et al.* (2012) Genome-wide genetic changes during modern breeding of maize. *Nat. Genet.* 44, 812–815
67 Hufford, M.B. *et al.* (2012) Comparative population genomics of maize domestication and improvement. *Nat. Genet.* 44, 808–811
68 Huang, X. *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501
69 Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160, 765–777
70 Nielsen, R. *et al.* (2005) Genomic scans for selective sweeps using SNP data. *Genome Res.* 15, 1566–1575
71 Emond, M.J. *et al.* (2012) Exome sequencing of extreme phenotypes identifies *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nat. Genet.* 44, 886–889
72 Sousa, V. and Hey, J. (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nat. Rev. Genet.* 14, 404–414
73 Gutenkunst, R.N. *et al.* (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5, e1000695
74 Beaumont, M.A. (2010) Approximate Bayesian Computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* 41, 379–406
75 Locke, D.P. *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature* 469, 529–533
76 Li, H. and Durbin, R. (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475, 493–496
77 Miller, W. *et al.* (2012) Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proc. Natl. Acad. Sci. U.S.A.* 109, 382–390
78 Nadachowska-Brzyska, K. *et al.* (2013) Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genet.* (in press)
79 Sheehan, S. *et al.* (2013) Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* 194, 647–662
80 Feder, J.L. *et al.* (2012) The genomics of speciation-with-gene-flow. *Trends Genet.* 28, 342–350
81 Nosil, P. and Feder, J.L. (2012) Genomic divergence during speciation: causes and consequences. *Phil. Trans. R. Soc. B* 367, 332–342
82 Keller, I. *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Mol. Ecol.* 22, 2848–2863
83 Parchman, T.L. *et al.* (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Mol. Ecol.* 22, 3304–3317
84 Floudas, D. *et al.* (2012) The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science* 336, 1715–1719
85 Scally, A. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature* 483, 169–175
86 Jia, J. *et al.* (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496, 91–95
87 Nystedt, B. *et al.* (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584
88 Dong, Y. *et al.* (2013) Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* 31, 135–141
89 Kim, J. *et al.* (2013) Reference-assisted chromosome assembly. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1785–1790
90 Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591
91 McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652–654
92 Rand, D.M. and Kann, L.M. (1996) Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans. *Mol. Biol. Evol.* 13, 735–748
93 Stoletzki, N. and Eyre-Walker, A. (2011) Estimation of the neutrality index. *Mol. Biol. Evol.* 28, 63–70
94 Fan, Y. *et al.* (2013) Genome of the Chinese tree shrew. *Nat. Commun.* 4, 1426
95 Alfoldi, J. *et al.* (2011) The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477, 587–591
96 Wang, Z. *et al.* (2013) The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat. Genet.* 45, 701–706
97 Chiari, Y. *et al.* (2012) Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* 10, 65
98 Amemiya, C.T. *et al.* (2013) The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496, 311–316
99 Smith, J.J. *et al.* (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.* 45, 415–421
100 Schartl, M. *et al.* (2013) The genome of the platyfish, *Xiphophorus maculatus*, provides insights into evolutionary adaptation and several complex traits. *Nat. Genet.* 45, 567–572
101 Zhan, S. *et al.* (2011) The Monarch butterfly genome yields insights into long-distance migration. *Cell* 147, 1171–1185
102 Tsai, I.J. *et al.* (2013) The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496, 57–63
103 Consortium, P.G.S. (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475, 189–195
104 Paterson, A.H. *et al.* (2012) Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* 492, 423–427
105 Wang, K. *et al.* (2012) The draft genome of a diploid cotton *Gossypium raimondii*. *Nat. Genet.* 44, 1098–1103
106 Xu, Q. *et al.* (2013) The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* 45, 59–66
107 Verde, I. *et al.* (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45, 487–494
108 Wu, H-J. *et al.* (2012) Insights into salt tolerance from the genome of *Thellungiella salsuginea*. *Proc. Natl. Acad. Sci. U.S.A.* 109, 12219–12224