Dissecting evolution and disease using comparative vertebrate genomics

Jennifer R. S. Meadows[1] *and Kerstin Lindblad-Toh*[1,2]

Abstract | With the generation of more than 100 sequenced vertebrate genomes in less than 25 years, the key question arises of how these resources can be used to inform new or ongoing projects. In the past, this diverse collection of sequences from human as well as model and non-model organisms has been used to annotate the human genome and to increase the understanding of human disease. In the future, comparative vertebrate genomics in conjunction with additional genomic resources will yield insights into the processes of genome function, evolution, speciation, selection and adaptation, as well as the quantification of species diversity. In this Review, we discuss how the genomics of non-human organisms can provide insights into vertebrate biology and how this can contribute to the understanding of human physiology and health.

Reference genome
A high-quality species genome onto which other information is projected, such as genes, polymorphisms and elements of gene regulation.

Bacterial artificial chromosome
(BAC). Approximately 200,000 bp of sequence that has been cloned into a bacterial vector and can then be amplified and sequenced.

Whole-genome shotgun sequencing
The genome is shattered into smaller pieces and sequenced, originally with Sanger technology.

[1]*Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Box 582, Uppsala 75123, Sweden.*
[2]*Broad Institute of Massachusetts Institute of Technology and Harvard, 415 Main Street, Cambridge 02142, Massachusetts, USA.*
Correspondence to K.L.-T.
kersli@broadinstitute.org

doi:10.1038/nrg.2017.51
Published online 24 Jul 2017

Understanding that, on the whole, genomes are not unique is one of the key features exploited by comparative genomics to address important biological questions. Comparative genomics involves the evaluation of genome homology, composition, organization and function, both within and across the species border, and has been exploited to unravel a myriad of processes from genome evolution to regulation and disease predisposition. Since the foundation of the 13-year programme to generate the prototype human reference genome[1,2], population-scale data of tens to thousands of individuals from the diverse spectrum of vertebrates are now collected each year (FIG. 1).

In the late twentieth century, two complementary approaches ran in parallel to obtain the first sequence of a vertebrate species: human. The International Human Genome Sequencing Consortium (IHGSC) used a multistage strategy: first, mapping ~200 kb bacterial artificial chromosome (BAC) clones to human chromosomes, then sequencing each BAC individually to high coverage. This was followed by careful finishing, which increased continuity and the accuracy of each base[1]. In parallel, a second team used a novel whole-genome shotgun sequencing approach, end-sequencing and assembling blocks of 2–50 kb (REF. 2). To place the sheer size of this endeavour in context, the IHGSC represented the collaborative efforts of 20 separate institutions at a cost of US$2.7 billion dollars (see National Human Genome Research Institute 1991 financial year). Although the contribution to science was staggering, it was obvious that in order to understand the influence of genome variation in both evolution and disease, novel and cheaper technologies would be required to generate whole-genome data sets at both the population and species level.

Following this, Sanger sequencing was used to obtain high-quality draft genome assemblies for key models for human health, such as rat[3] and mouse[4]. For example, the mouse genome project sequenced the inbred laboratory strain C57BL/J6 and compared this with additional data generated from eight inbred strains and four wild mice species. The subsequent analysis provided a window into mouse genetic ancestry, identifying megabase-sized haplotype blocks of each wild mouse species interspersed in most laboratory strains[4]. As for all species, cataloguing and understanding genomic variation, in addition to genome structure and gene set, was key to future research endeavours.

Although it was still a costly exercise (~$1,000 per megabase of sequence; see National Human Genome Research Institute sequencing costs), the following decade saw Sanger sequencing used as the major methodology for the generation of high-quality genomes, including for dog[5], horse[6], cow[7], macaque[8] and opossum[9]. The value and power of comparative genomics was soon shown, when the analysis of human, rat, mouse and dog genomes facilitated the adjustment of the human and mammalian protein-coding gene count to around 20,000 genes[5]. This was in stark contrast to the 100,000 genes estimated before the existence of the human genome sequence and the tentative 40,000 genes extrapolated from the human genome sequence[1]. Although this comparison showed that the number of genes was far lower than originally projected, our understanding of genome regulation increased. Protein-coding genes

Sanger sequencing
An old standard type of sequencing in which the four bases are labelled with four fluorophores of different colours. It results in ~600 bp reads and was the methodology used for the human genome project.

were assessed to cover ~1.2% of a mammalian genome, but at least 4.2% of the genome was deemed functional and conserved between human, rat, mouse and dog[5]. This additional non-coding conserved sequence was no longer considered 'junk', rather it was assumed to contain regulatory elements such as promoters, enhancers and insulators, as well as non-coding RNAs[10]. These regulatory regions are key to understanding how complex organisms develop from embryogenesis through to their adult form, and how their organ systems can maintain homeostasis and adjust to stress.

In this Review, we discuss how to best use new and existing data to answer novel biological questions with comparative vertebrate genomics, including project design based on resource availability and scientific goals. In addition, we discuss the annotation of the human genome and the application of comparative genomics to aid the conservation of endangered species. Throughout

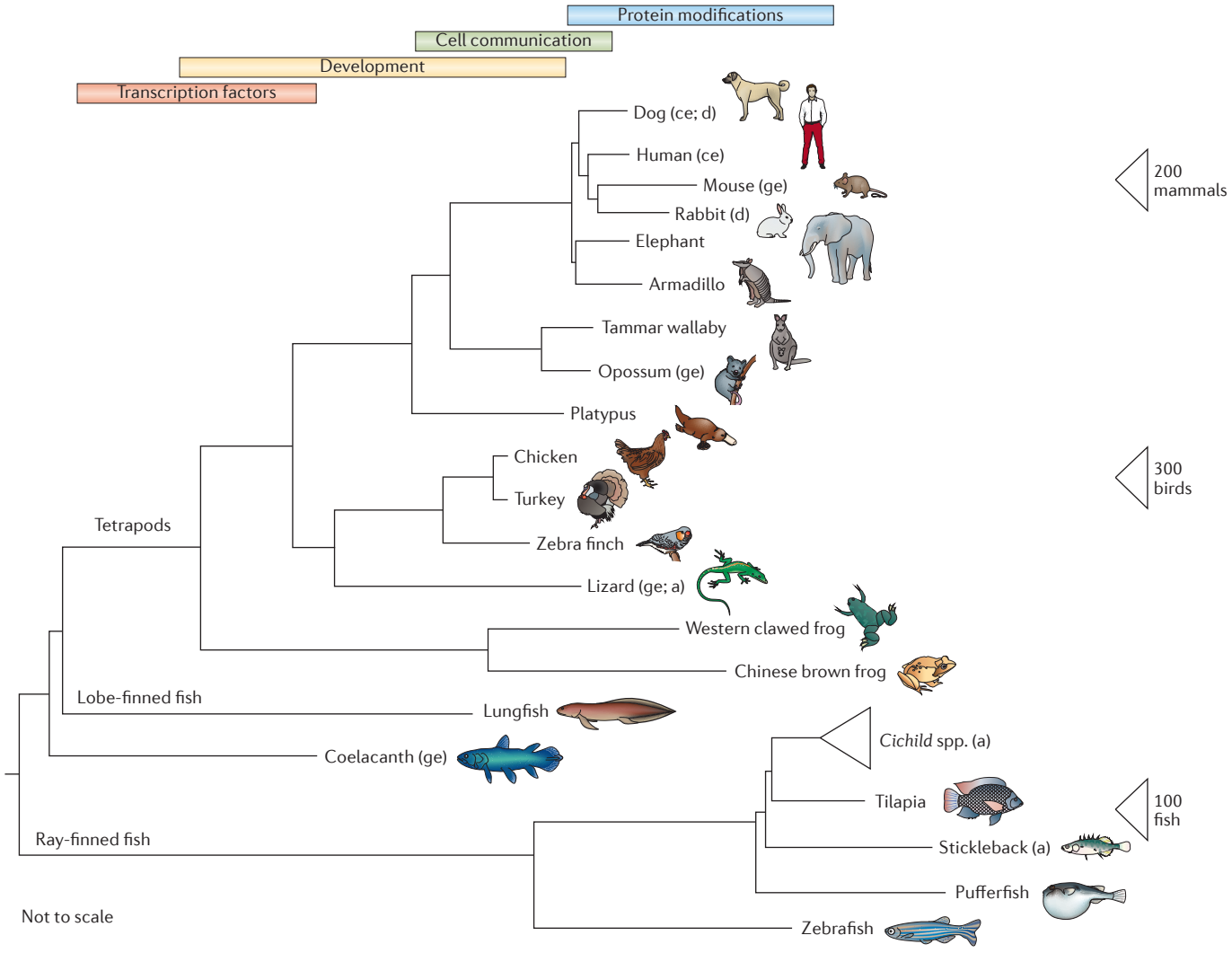

Figure 1 | **A snapshot of vertebrate genome sequencing projects.** Many of the first vertebrate whole-genome projects represented model species (for example, mouse and rat), but over time, additional resources representing natural model species have been added. Highlighted in this tree are some of the studies that have been undertaken, within and across lineages, to study the processes of natural adaptation (a; for example, stickleback adaptation to extreme aquatic environments[57]), domestication (d; for example, genetic signatures separating domestic dogs and wolves[69]), genome evolution (ge; for example, exaptation changes in regulatory sequence function between human and mouse[91]) and convergent evolution (ce; for example, metabolism and neurological development between human and dog[6,69,75]). As well as indicating the genetic distances between representative vertebrate species, this tree also illustrates the time periods when novel regulatory innovations arose and so can be studied using comparative genomics. In particular, regulatory elements near transcription factors (red) and developmental genes (yellow) evolved quickly in early vertebrate history, followed by cell communication (green) and protein modification (blue) in the more recent past. As whole-genome sequencing becomes more accessible, the expansion of each clade is set to increase, with the publication of 200 mammals, 300 birds and more than 100 fish expected by the close of 2017. Image adapted with permission from REF. 78, Macmillan Publishers Limited.

the text, we provide illustrative examples of sequenced vertebrate model species or key non-model organisms to be further studied.

The sequencing revolution

The desire to sequence a large number of whole genomes from both humans and many vertebrate species drove the development of new short-read sequencing (SRS) technologies, including ABI SOLiD[11], Roche 454 (REF. 12) and Illumina[13]. Each technology has its pros and cons in terms of cost, sample input, read-length and error rates; however, over the past 5 years, large amounts of data have been generated with the Illumina paired-end approach. This method can generate reads of up to 250 bp by sequencing each end of a larger fragment insert, which is subsequently used in the building of a reference-guided, or *de novo*-assembled, genome. Recognizing the need for long-range continuity, new technologies that facilitate genome assembly (for example, Dovetail's proximity ligation[14]) and long-read sequencing (LRS) (for example, PacBio's single-molecule real-time (SMRT)[15], Chromium 10x[16], Oxford Nanopore technologies[17] and Bionano genome mapping[18]) were developed.

Sequencing (SRS and LRS) and assembly tools are not used in isolation, and different methods can be combined in order to achieve higher quality genome builds. This includes improvements to existing gold-standard genomes, such as the current human build GRCh38, where optical mappers such as Bionano were used to confirm assemblies and refine haplotypes[19]. In addition, the high-coverage budgerigar (*Melopsittacus undulatus*) genome[20] was generated using a hybrid of sequencing technologies, including SRS (Roche 454 and Illumina) and LRS (PacBio) and — similarly — the *de novo* goat (*Capra hircus*) reference genome[21] combined PacBio, Illumina and Bionano methods to generate one of the most contiguous mammalian genomes so far. In that example, the additional chromosome interaction mapping in the form of chromosome conformation capture combined with deep sequencing (Hi-C)[22], facilitated the joining of scaffolds but also enabled the annotation of long-range chromosome interactions, which are key to deciphering *cis*- and *trans*-regulation.

The analyses carried out during the production of the goat reference genome can be used as guidelines to show the assembly gains that are obtainable when various LRS and SRS technologies are used in isolation or in various combinations[21]. This could be extremely useful in weighing project cost against contig and scaffold length (the strengths and weaknesses of different sequencing technologies have been previously reviewed in REF. 23).

We live in an era in which tens of thousands of human genomes are being sequenced, facilitating the interrogation of the selection pressures acting at each base in the human genome. For example, more than 4,000 human whole genomes are contained in the UK10K databases for rare genetic variants in health and disease, and 60,000 unrelated individuals are represented in the ExAC consortia database. Similarly, large collaborative projects, such as the 1000 bulls genome and the dog 10K genome project, are currently running within mammalian species. The Genome 10K consortium, which includes mammals but also birds and fish, has the long-term objective of obtaining high-contiguity genomes for 10,000 vertebrate species[24]. These projects have goals in common, mainly to catalogue, collate and apply genome-level variation to the understanding of genomic processes, be they disease mechanisms or evolution.

For any new study that generates a novel genome sequence, it is important to decide what additional resources would be valuable to make the most of this unique resource. For example, population genomics data would assist with the identification of single-nucleotide polymorphisms (SNPs), insertions and deletions (indels) and large-scale structural rearrangements. In addition, carrying out RNA sequencing (RNA-seq) can be valuable for the characterization of coding genes, non-coding RNAs and microRNAs, and methods to detect histone marks and to annotate 3D chromosomal interactions (for example, Hi-C) are required to unravel layers of genome regulation. In addition to the scientific goals, the special characteristics of the species being studied must be considered. This includes the availability of high-quality tissue for DNA or RNA extraction (which are not always easy to find), as well as potential restrictions based on membership of endangered species lists.

Design based on biology

Both within and across the species tree, comparative genomics has been used to explore a range of biological questions over diverse time periods. A broad range of sequenced vertebrate genome projects have been used for studies of natural adaptation, domestication, genome evolution, as well as genome conservation and convergent evolution. FIGURE 1 shows a snapshot of these projects on the vertebrate species tree, also highlighting four 'eras' of genome evolution based on the genes that are enriched close to novel regulatory innovations through vertebrate evolution. For example, genes encoding transcription factors and factors that regulate development evolved early in vertebrate history, followed by cell communication and protein modification genes, which evolved in the more recent past.

Successful project design depends on generating sufficient evolutionary power through comparison of species, populations or individuals. When studies use many individuals, success is not dependent on all samples having a high-quality genome sequence. Pooled genome sequence strategies (as for Darwin's finch[25]) and representative genome assembly approaches (which have reasonable N50 contig size per species but lack high long-range contiguity for all species; for example, the 29 mammals project[26]) are both examples of these principles in practice.

Chromosome organization, expansion and retraction of gene families or repeat content, and the innovation of regulatory elements, can all be studied with comparative genomics, but this requires higher contiguity for the genome assemblies under interrogation. Therefore, when designing comparative genomics projects it is important to carefully assemble the best genomic and phenotypic

High-quality draft genome assemblies
A form of genome assembly that has both long contigs (stretches of uninterrupted sequence, many kb in length) and supercontigs (structures of sequence hanging together but including smaller gaps) in the Mb range.

Haplotype blocks
Regions of the genome that are inherited together without recombination. These are characterized by high linkage disequilibrium.

Vertebrate model species
Vertebrate species that are studied to understand the biology or phenotype in another species.

Non-model organisms
Organisms that are examined to offer insight into themselves rather than principally studied to understand another trait, for example, human health.

Short-read sequencing
(SRS). Short-read technologies, such as Illumina, generate continuous sequence length of 100–250 bp.

Long-read sequencing
(LRS). Strategies such as PacBio's single-molecule real-time (SMRT) generate continuous sequence length in the order of many kilobases. Over time, these technologies will go down in price and will probably be the methods of choice.

Haplotypes
Versions of a gene or part of a gene, including several variants that are inherited together.

Chromosome interaction mapping
A methodology to analyse the 3D organization of chromatin. Looping can be functional, for example, bringing enhancers into contact with distal promoters.

Single-nucleotide polymorphisms
(SNPs). When a position in the genome can have two or more alleles. Biallelic markers are used to look for association of one allele (gene version) with disease.

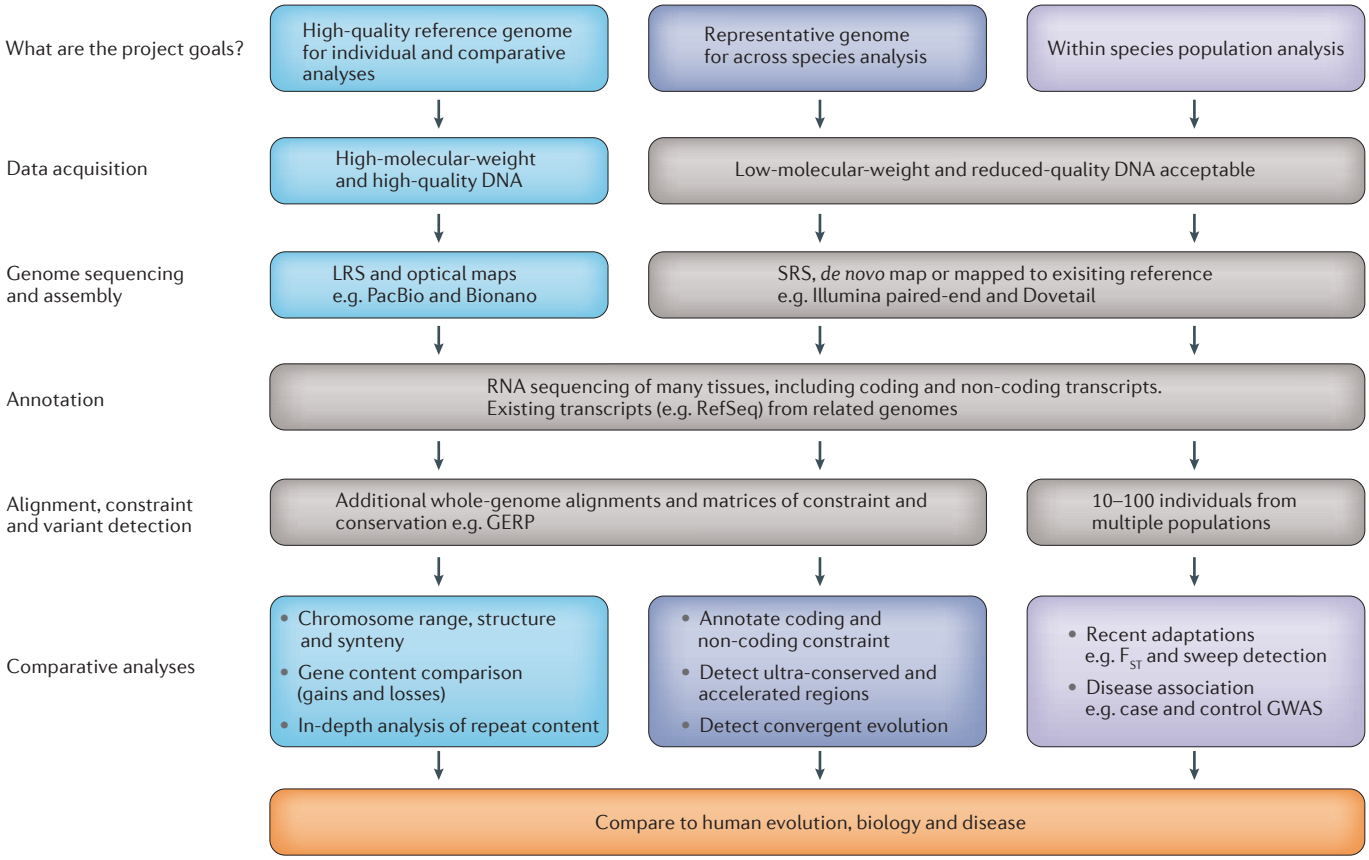

Figure 2 | **Designing a sequencing project.** The flowchart provides examples of the biological and technical issues that can affect the decision tree of project design: from the project aim, through to the choice of which sequencing and analysis solution may be best suited to enable success. For example, determining whether long-range contiguity is needed for the study of chromosome structure or gene-family expansion, or whether a majority of the genome sequence (albeit in smaller pieces) is sufficient for studying constraint and accelerated evolution. For another example, the study of recent adaptations requires that multiple populations are sequenced. For all genome projects, RNA sequencing of many tissues is key for annotation, but in lieu of this resource, comparative mapping of RefSeq annotations can provide insight into the transcriptome of the species under investigation. Each new genotype–phenotype correlation has the potential to increase our understanding of vertebrate physiology and homeostasis, and as such can be translated to dissect the genetics of human health and disease. F_{ST}, fixation index (F-statistics); GERP, genome evolutionary rate profiling; GWAS, genome-wide association studies; LRS, long-read sequencing; SRS, short-read sequencing.

information; in some cases this may include reassembly and re-annotation to integrate data sets and to minimize confounding effects[27]. FIGURE 2 provides an overview of some of the key biological and technical issues that need to be addressed when designing a sequencing experiment. These are described in more detail below.

The effect of genome content. A key consideration when designing projects is outcome: is the goal to capture as much genome sequence as possible or is it to study features relating to chromosomal structure or other long-range regulatory relationships? In a dream scenario, all genomes would be sequenced using methods that can achieve high coverage and long-range contiguity. This would ensure that these resources could be used for a multitude of purposes beyond their original design. However, a lack of high-quality DNA impairs the generation of a contiguous genome and similarly, long-range contiguity is contingent on the availability of affordable technology (previously, radiation hybrids and BAC maps, BACs and fosmids; now, PacBio, Dovetail and Bionano) (TABLE 1).

A single 30x coverage Illumina mammalian assembly (450-bp fragments, 2 × 250-bp reads, run on 1 lane on Illumina HiSeq 2500) might result in N50 contig and supercontig sizes of 30–200 kb and might cover the majority of genes. From here, conserved synteny to nearby species can be used to study the structural integrity of genomes with smaller supercontig N50 size. Genome build upgrades using newer and more expensive technologies, such as PacBio or Dovetail, may not greatly change the gene content or N50 contig size, however, the incorporation of these tools can increase the supercontig N50 size up to multiple megabases, resolve repeats and gene families, and ultimately enable structural analysis and sequence assignment of unmapped regions to chromosomes.

Vertebrate genomes can have unusual features, such as high repeat content (for example, *Nothobranchius furzeri*[28], a fish model for ageing), an elevated GC content (for example, rabbit[29] and dog[5], models for domestication, adaptation and disease predisposition), microchromosomes (for example,

RNA sequencing
(RNA-seq). The sequencing of all mRNA transcripts from a cell or tissues.

Histone marks
Histones are proteins that package DNA into units. Histone marks indicate where chromatin is open or closed and provide insights into genome regulation; for example, histone 3 lysine 27 acetylation (H3K27ac), is associated with active enhancers.

Adaptation
A trait that has changed to enable a species to function under certain circumstances or in a specific environment.

Table 1 | **Example resources for comparative genomics**

Purpose, task or tool	Description	Web address
Genome assemblies		
DISCOVAR *de novo* assembly	Assembler for Illumina reads	software.broadinstitute.org/software/discovar/
Draft level (old)	2x Sanger sequencing; 4 kb and 40 kb	genome.ucsc.edu/cgi-bin/hgGateway
Draft level (new)	Illumina paired-end reads (30–60x)	www.ncbi.nlm.nih.gov/genome/
High contiguity (old)	7x Sanger sequencing; multiple insert sizes	
High contiguity (new)	Dovetail or PacBio	dovetailgenomics.com
RACA	Reference-assisted chromosome assembly	bioen-compbio.bioen.illinois.edu/RACA/
Complicated genome regions resolved	PacBio	www.pacb.com
Population data		
ANGSD	Association mapping and population genetic analyses with next-generation sequencing data	www.popgen.dk/angsd/index.php/ANGSD
Codeml	Convergent analysis in software package PAML	abacus.gene.ucl.ac.uk/software.html
F-statistic	Population differentiation	–
MrBayes	Construct phylogenies using genes or genomes	mrbayes.sourceforge.net/
Saguaro	Detect signatures of selection within or across populations	saguarogw.sourceforge.net/
SNP discovery	Low coverage sequencing many individuals	software.broadinstitute.org/gatk/
Sweep discovery	Variants from multiple contrasting populations; window-based diversity	–
Alignment and visualization programmes		
BWA aligner	Read alignment to genome	bio-bwa.sourceforge.net
IGV	Display of read to genome alignments	software.broadinstitute.org/software/igv/
LASTZ	Pair-wise sequence aligner, suitable for whole chromosomes and comparative analyses	http://www.bx.psu.edu/~rsharris/lastz/
LiftOver	Tool for mapping from one species to another	genome.ucsc.edu/cgi-bin/hgLiftOver
SMALT	Align RNA sequencing data to reference	www.sanger.ac.uk/science/tools/smalt-0
SOAPdenovo	Short-read *de novo* aligner for Illumina data	soap.genomics.org.cn/soapdenovo.html
Gene annotation and expression		
Gene set Ensembl	For human and certain vertebrates	useast.ensembl.org/index.html
GTEx	Correlation between human RNA sequencing and variants; expression quantitative loci	www.gtexportal.org/home/
RefSeq NCBI	For human and certain vertebrates	www.ncbi.nlm.nih.gov/genome/annotation_euk/
Functional element and variant annotation		
29-mammals constraint	29 mammals used for determining constraints	genome.ucsc.edu/cgi-bin/hgGateway
ENCODE	Functional annotation on human genome	genome.ucsc.edu/cgi-bin/hgGateway
FAANG	Functional annotation on animal genomes	www.faang.org
PhastCons	Determining constraint across species	compgen.cshl.edu/phast/phastCons-HOWTO.html
SiPhy	Determining constraint across species	portals.broadinstitute.org/genome_bio/siphy/
SnpEff	Likelihood variant is functional (primarily coding)	snpeff.sourceforge.net
VEP	Likelihood variant is functional (coding or regulatory)	www.ensembl.org/vep

ANGSD, analysis of next-generation sequencing data; BWA, Burrows–Wheeler aligner; ENCODE, encyclopedia of DNA elements; FAANG, functional annotation of animal genomes; GTEx, genotype-tissue expression; IGV, integrative genomics viewer; Phast, phylogenetic analysis with space/time models; RACA, reference-assisted chromosome assembly; RefSeq NCBI, National center for biotechnology information reference sequence; SiPhy, site-specific phylogenetic analysis; SOAP, short oligonucleotide analysis package; SNP, single-nucleotide polymorphism; SnpEff, SNP effect; VEP, variant effect predictor.

Domestication
A complex process partly driven by human selection of standing natural variation.

chicken[30], a model for obesity and behaviour, or green anole lizard[31], a model for amniote evolution) or large genome size (for example, salamander[32], a model for muscle development). Each of these factors can complicate genome sequencing. In addition, it can be harder to sequence recent gene family expansions or specific gene families (for example, immune genes with similar functional alleles). All of these elements can impede assembly and may require high-coverage LRS technologies, such as PacBio, to resolve. Such

long reads, several kilobases in size, can span complex or repetitive regions with a single continuous read, thus eliminating ambiguity in the positions or size of genomic elements.

Standing variation, imputation and mapping. For variant detection, the relatively low-coverage sequencing of many pooled individuals from one or more study populations can be a cost-effective place to start. Variant detection can then be carried out using software packages such as CRISP, LoFreq and VarScan (cross-package use and accuracy are reviewed in REF. 33). It is important to carefully select the input population, species or individuals, so as to balance genetic diversity and the ability to detect variants, with the full representation of phenotypes being studied. For example, to look for genomic regions affecting the adaptation of Darwin's finches to their various island environments[25], the authors selected 5–10 individuals from each of the existing species, covering all feeding environments and key morphological traits (for example, beak size), plus a subset of species outgroups.

This framework enables signatures of selection, such as selective sweep signals, to be clearly distinguished and attributed to the phenotype segregating between the groups, as well as detection of sites of hybridization or introgression. Gene flow and sweep signals can be detected in different ways[34], including measures of genetic distance such as F-statistics[35], Tajimas' D[36] or methods using haplotypes in sliding windows, such as the integrated haplotype homozygosity score (iHS)[37].

For trait mapping within a species, it is important to generate a sufficient density of SNPs to be able to tag all haplotypes in the genome of interest. Haplotype structure is closely connected to the population history of the species. Using dogs as an example, within-breed haplotypes are long (megabase-sized) because of the relatively recent creation of different breeds of dog. Haplotype blocks are short in village dogs[38] or across breeds, as dog domestication happened in the distant past[5]. In many cases, genetic bottlenecks or species creation will have happened a long time ago, which brings the number of markers needed on SNP genotyping arrays into the millions (for example, domestic cattle[7]). However, as sequencing prices continue to decrease, it becomes feasible to generate whole-genome sequences for a subset of a population and use this as a reference panel to impute existing genome-wide SNP arrays to a higher density for trait mapping[1,39,40]. Not only does this increase experimental power through the addition of markers but also it covers parts of the genome missing from the original experiment. Popular tools for imputation include Beagle 4.1 (REF. 41), IMPUTE2 (REF. 42) or FImpute[43] (for large, across family data sets, such as those used in livestock genomics).

Complex mutation types: the good with the bad. We tend to think of mutations as only single base pair changes, but this is simply not the case. Structural variation such as deletions, duplications, inversions and translocations can all affect genome content and sequencability. Structurally complex regions can result in maps of high regional heterozygosity or genome fractionation because of the inability of SRS technology to rationalize or to span these regions. With careful phasing (identifying the chromosome that a particular SNP belongs to) and sequencing of key individuals, complex rearrangements can be dissected. Such was the case for the ruff bird species, *Philomachus pugnax*, in which a compound inversion was responsible for the three phenotypic forms of male birds in the same breeding population[44].

A technology shift is now being used to catalogue and dissect the biological relevance of these rearrangements. The first effort to apply PacBio SMRT to a diploid human genome showed gains compared with shotgun sequencing approaches and that it could resolve collocated structural variation, thus demonstrating that these large-scale elements are segregating in the human genome[45]. Similarly, when SMRT was applied to sequence a single gorilla genome[46], the results included the accurate assembly of complex regions such as the major histocompatibility locus, but also showed an enrichment of genome segmental duplication (segments of DNA with near-identical sequence). Comparing this *de novo* gorilla genome to human and chimp enabled new inferences over evolutionary timescales to be hypothesized, including the lineage-specific loss of genes and gene regulation[46].

Layering complexity: gene and transcript annotation. An in-depth annotation of genes, non-coding RNA and regulatory elements are the foundation building blocks for projects that aim to understand adaptation or disease predisposition (FIG. 2). A solid gene annotation can be achieved in multiple ways, including synteny mapping of transcripts from a well-annotated closely related species or through mapping RefSeq transcripts onto the genome in question. Alternatively, if the appropriate tissues can be acquired, RNA-seq and transcript assembly can be carried out. Mapping these transcripts to the genome gives the most species-specific annotation. The most recent annotation of the canine genome, CanFam3.1 (REF. 47), used all of these techniques to expand the catalogue of transcribed elements by a factor of four (~175,000 expressed loci). That set also included long non-coding RNAs (lncRNAs) lifted from human and was shown to have functional relevance in dog[47]. As RNAs are involved in processes from protein synthesis to post-transcriptional modification and gene regulation, enriching and sequencing these essential molecules in key tissues or single cells may reveal crucial biological mechanisms[2,48,49].

It is well established that the majority of genome-wide association study (GWAS) signals map outside genes, suggesting that the causative trait variant (or variants) may reside in regulatory elements[3,50,51]. This is probably true across vertebrates, making the detection and curation of regulatory regions important on a comparative genomics scale. Similarly to transcript annotations, LiftOver tools (which convert genome coordinates and genome annotation files between assemblies) can be used to map regulatory elements and data from the heavily studied human or mouse. For example, comparative Hi-C, a method to capture and sequence chromatin

Convergent evolution
The independent evolution of similar features in multiple species of different lineages.

Pooled genome sequence strategies
Several individuals and/or samples can be sequenced together as a group, either with or without barcode labelling to facilitate multiplexing.

Representative genome assembly approaches
When multiple individual genomes are sequenced from a species, the best is selected as a reference genome for that organism.

N50 contig
A statistic used to illustrate genome quality. Genomes are constructed of multiple contigs (a segment of the genome assembly that contains no gaps), each with different lengths. N50 size is the shortest sequence length containing half of the genome sequence.

Long-range contiguity
The linking of large, megabase-sized genomic regions in order to create large continuous lengths of sequence data.

Conserved synteny
The similarity of gene order in large regions of related (and distant) species.

Microchromosomes
Typical in some birds and lizards, these chromosomes are less than 20 Mb in size.

Selective sweep
A region of the genome where there is little to no population-level variation, as one haplotype with favourable alleles has become more common than other variants.

Hybridization
The mating of two different species or populations, resulting in equal proportions of genetic material from both parents.

Introgression
Gene flow from one species into the gene pool of another by the repeated backcrossing of a hybrid with one of its parental species.

conformation, recently revealed that topologically associating domains (TADs) were conserved across the species boundary[52]. This study used the diverse lineages of mouse, macaque, rabbit and dog, and so it is likely that, to at least some extent, this conservation will span Eutherian (placental mammals) genomes.

If obtaining tissues or cell lines is not a problem for the species under investigation, specific data sets can be derived. For example, chromatin immunoprecipitation with massively parallel DNA sequencing (ChIP-seq) for mapping transcription factor-binding sites, or variations of Hi-C that can identify where DNA loops connect enhancers with target genes, or assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) for identifying open chromatin. It would be desirable to obtain these data sets for every species; however, it can be both difficult and expensive to source tissues and therefore to complete the genome annotation for each cell type of every species. The ever-expanding Encyclopedia of DNA Elements (ENCODE)[53] (TABLE 1) is an excellent source for human-centric cell line data but also encompasses software tools and analysis pipelines that can be adapted across species. The scope of ENCODE is vast and the best way to gain familiarity with the project is through the special addition of *Nature*, the Nature Encode Explorer. The Functional Annotation of Animal Genomes consortium (FAANG) (TABLE 1) aims to facilitate a similar catalogue of elements for domestic and non-model organism species[54]. Although still in its pilot phase, it is an excellent resource for advice on a broad range of experimental issues from sample collection and analyses to result harmonization and meta-analysis[55].

In addition to species-specific experimental data, genome annotation can be achieved by identifying constrained elements across groups of species. For this, 20 low-coverage 2x Sanger mammalian genome sequences were generated and combined with existing high-coverage species to enable the analysis of 29 mammalian species[26]. This data set contained only ~85% of the genome for each new species but achieved a strong analysis power across species. Constrained elements, covering ~4.2% of the genome, were identified down to a resolution of 12 bp and candidate functions were assigned to ~60% of constrained bases[26].

Vertebrate comparative genomics

Species may be studied to gain insight into themselves (for example, barring feather patterns in chicken) or for the insight they provide for other processes or organisms (for example, using barring feathers in chickens as a model for melanocyte migration in vertebrates). In this way, all organisms can be both vertebrate model species and non-model organisms. A second level can be used to define model species: experimental or natural. Mice, rats, rhesus macaque and a few other species are clear experimental model species used for evolutionary and biomedical experiments in a laboratory setting. Natural models tend to be species that are studied in their natural habitat in order to understand the normal life or disease of one or more species, or the differences between individuals within a species. In this way, natural models can facilitate insight into biological homeostasis or health attributes. Examples of natural models include the study of behaviour in wild chimpanzees[56], adaptation to a freshwater environment by sticklebacks[57] (BOX 1) and the traits of growth[58], and parasite resistance[59] in agriculturally important species such as sheep. As genomics improves and enables the design of more targeted studies relating genotypes to phenotypes, it could be argued that almost all species can be model species at a natural and/or genomic level. Therefore, for the purpose of this Review, we consider the spectrum of vertebrate model and non-model organisms, providing examples of their key properties and questions that can be explored.

Natural disease models: domestic animals. Domestication is the complex process by which animals adapt to be in close proximity to man, often via selective breeding of natural variation. Whether for the farm or home, domestication, breed formation and subsequent generations of reproduction have resulted in ideal phenotypes based on coat colour, morphology, production values or certain types of behaviour. A side effect of this process has been the inadvertent enrichment for different types of diseases. This interplay between desired and deleterious traits, some of which are pleiotropic, has been particularly well used in the study of both canine (BOX 2) and feline genomics. For example, ~50% of pet dogs develop cancer in their lifetime[60], and some breeds show enrichment for heart, immunological and neuropsychiatric diseases[4,60,61]. Although most wild animals, including dogs, develop cancer, the rates are typically much lower than in domestic animals or humans[62]. In addition, domesticated cats have been characterized for more than 250 analogous human disease sets[63], including genetic diseases such as mammary tumours[64], diabetes mellitus[65] and eye disease[66], as well as viral diseases such as HIV[67] and viral induced leukaemia[68]. Although these animals, and other domestic animals, are not experimental models kept in a laboratory, they do receive regular veterinary care and so can be used for the study of many different traits and diseases. The Online Mendelian Inheritance in Animals (OMIA) is an excellent resource linking both associated and causal mutations to animal species and vertebrate physiology. Using this repository, it is possible to extrapolate many of these genomic changes to further understand human health and disease. However, when using these species as models it is also important to consider if the research solely aids humans or if the model organism itself is also a beneficiary.

To enable trait mapping, it is important to generate and annotate a good reference genome and to add diversity data across different populations. In particular, SNPs can be used for genome-wide association mapping using genotyping arrays, whereas additional genome-wide or targeted sequencing can be used to further characterize disease mutations. Genes and regulatory elements can be mapped onto the genome with RNA-, chromatin- or transcription factor-binding sequencing and/or by cross-species annotation using a more highly studied reference genome, such as human or mouse.

Integrated haplotype homozygosity score
(iHS). A method to calculate the amount of genetic similarity across regions in a species or population. High homozygosity suggests selection to be active on that region.

SNP genotyping arrays
A method to genotype predefined single-nucleotide variants distributed across the genome of the species under study. For humans, single-nucleotide polymorphism (SNP) arrays typically have millions of variants, whereas in dogs, hundreds of thousands of SNPs are used for genome-wide association mapping.

Topologically associating domains
(TADs). Regions of the genome packaged together in 3D space, most often containing one or a few genes and their regulatory signals. Genomic interactions within TADs are more frequent than those across TAD borders.

Box 1 | **Selection and adaptation to a freshwater environment**

Multiple fish species, including sticklebacks and herring, now live in both marine and freshwater environments. Marine stickleback colonized and adapted to innumerable new streams and lakes formed following the end of the last ice age, ~10,000 years ago. Similarly, herring is one of few marine fish that reproduce throughout the Baltic Sea. There, salinity drops from 35% in the Atlantic Ocean to 2–3% in the Bothnian Bay. The herring ecological adaptation must also be recent, as the Baltic Sea was formed following the last glaciation[83]. Both of these species therefore provide opportunities to study the genetic and molecular basis of adaptive evolution in a natural environment.

To understand the general mechanisms and specific gene loci involved in adaptation, both the stickleback and herring genomes were sequenced and analysed together with population-level genetic data. The questions were similar but the data sets varied based on the tools available at the time of sequencing. For both projects, there was an attractive hypothesis: regulatory variants may have had an important role in the evolution of adaptation in these naturally occurring species, as such mutations may have avoided the fitness costs associated with the pleiotropic consequences of protein-coding alterations.

A female freshwater threespine stickleback (*Gasterosteus aculeatus*) from Bear Paw Lake, Alaska, was sequenced to 9x coverage in Sanger sequence data in 2006 (REF. 57). The assembly, gasAcu1.0, has an N50 contig size of 83.2 kb, an N50 scaffold size of 10.8 Mb and a total gapped size of 463 Mb. The assembly was annotated using the Ensembl pipeline. This predicted 20,787 protein-coding genes, of which 7,614 showed one-to-one orthology with mammals and an additional 7,192 showed one-to-one orthology among fish.

To identify loci related to freshwater adaptation, 21 individuals spread across marine and freshwater ancestry were sequenced at low coverage with Illumina technology. Two different methods were used to identify 242 regions (0.5% of the genome) that had diverged between freshwater and marine individuals. The median size of recovered regions (<5 kb) approaches the size of individual genes. In addition to genic regions, the analyses highlighted many purely intergenic regions with potential adaptive functions.

The Atlantic herring is one of the most common fish in the world and has been a crucial food resource in northern Europe. Early studies limited to a small number of genes did not show differentiation between herring living in marine versus brackish waters. However, it would seem likely that the adaptations to freshwater are linked with genetic changes. The herring genome was Illumina-sequenced in 2012, resulting in an 808 Mb assembly (123x coverage with different insert sizes and scaffold N50 of 1.84 Mb) and contained 23,336 predicted coding gene models[84].

To study freshwater adaptation, 20 populations of herring from the Baltic Sea, Skagerrak, Kattegat, North Sea and Atlantic Ocean were sampled for sequencing pools. Each pool comprised 47–100 fish and was sequenced to ~30x coverage. Furthermore, 16 fish, 8 Baltic and 8 Atlantic herring, were sequenced individually to ~10x coverage. In-depth analysis of allele frequencies in Atlantic and Baltic population identified 122 regions with highly significant association to salinity. Intriguingly, 21 of the genes in these regions coincided with regions that have previously been associated with hypertension in human. In addition, 36 of these genes showed differential expression in fish kept in freshwater or seawater. For several genes, amplifications of promoter regions and other structural changes coincide with the adapted regions. These results show that similar methods of adaptation were used in both species.

Selective sweep mapping has been successfully used to look for traits that have been under intense selection, either in conjunction with domestication (for example, canine adaption to a starch diet[69]) or specialization (for example, horse gait[70]). In the case of horse, signatures of selection on equine chromosome 23 in all gaited breeds led to the identification of a shared 186 kb haplotype, including two genes, doublesex and mab3-related transcription factor 2 (*DMRT2*) and *DMRT3* (REF. 71). A *DMRT3* mutation within this haplotype, necessary for an alternative gait phenotype, provided further evidence for selection at this locus[70]. A follow-up study showed the mutation to be present in 68 out of 141 worldwide horse populations, suggesting that it is not a recent occurrence and therefore that alternative equine gaits were selected early by humans[72]. Notably, by introducing the horse gait mutation into murine models, it was shown that *DMRT3* was crucial to the normal development of a coordinated locomotor network[70]. As the gene is conserved across vertebrates, we posit that deregulation of this gene could have detrimental effects on the spinal circuits that control normal motion in humans.

Intraspecies comparison: a tool to study recent phenotypic adaptations. For species living in the wild, selection on natural phenotypic adaptations can be very strong, for example, fish species and salinity[73] or Caprinae species and altitude[74]. To enable the study of these types of adaptations (microevolution), a suitable reference genome with gene annotation is crucial. In addition, collating and sequencing 10–20 individuals from each population with known phenotypic status can enable the detection of selective sweeps when outward phenotypes appear the same. A crucial factor to consider is that selective sweeps often encompass multiple genes and a very large number of polymorphisms. It is likely that only one or a few of these variants are actually related to the trait under study.

Intriguingly, comparisons of marine herring populations (in the Atlantic Ocean) to brackish water herring populations (in the Baltic Sea) identified ~500 independent loci associated with the recent niche adaptation[73] (BOX 1). In addition, more than 100 independent loci showed genetic differentiation between spring- and autumn-spawning populations, irrespective of geographic origin. The locations of sweeps suggest that both coding and non-coding changes contribute to these adaptations. Frequently, large haplotype blocks, often spanning multiple genes and maintained by selection, are associated with the genetic differentiation required for each adaptation. This suggests that multiple variants within these blocks may work together to account for the differentiation between marine- or brackish-residing and spring- or autumn-spawning populations.

The adaptation of sheep to high altitude on the Tibetan plateau is another example of natural adaptation (FIG. 3). In this example, sequencing and contrasting seven populations of high- and low-altitude sheep revealed selective sweeps containing more than 200 genes enriched for functions related to angiogenesis, energy production and erythropoiesis[74]. Intriguingly, the gene endothelial PAS domain protein 1 (*EPAS1*) contained as many as 12 mutations differentiating between highland and lowland sheep, suggesting that strong selective pressure lead to the accumulation of additional mutations over time. Analyses of convergent evolution have also implicated *EPAS1* in the high-altitude adaptation of humans and dogs[75] (FIG. 3a). Using regulation track annotations for human (such as histone modifications), a comparative analysis suggests that it may be possible to infer that the variants within sweeps act in endothelial and lung cells, but not in embryonic stem cells (FIG. 3b). These data indicate which tissues or cell lines could be targets for the functional validation of selected variants. The comparative analysis also suggests that altitude

Microevolution
Changes in allele frequencies that happen within a population in a fairly short time span. This can be due to positive selection or drift.

tolerance attributed to *EPAS1* would not extend past mammalian lineages, as sequence conservation does not extend past elephant (FIG. 3c).

Interspecies genomic comparison. Although population-based sequencing is ideal for recent adaptations, selection events that happen over a long time period (macroevolution), often becoming fixed in a particular species, are better studied across species. This includes considerable phenotypic changes between vertebrate groups such as between Eutherian (placental) and Metatherian (marsupial) mammals, birds and fish. Vertebrate clades began emerging more than 450 million years ago, and each clade now contains marked phenotypic differences. This diversity is accompanied by underlying genetic changes, but a fraction of these will be silent. The large accumulation of silent changes in neutral sites over long time periods makes it challenging to distinguish which changes are functional and which are not.

To study these types of long-term changes, high-quality genomes are required, with high accuracy and contiguity. This enables the investigator to draw conclusions based on what is found and also what is missing from within a genome. Clades with multiple species representation facilitate enhanced differentiation of significant functional changes from the background consensus. These theories have been put into practice with the detection of convergent evolution between two distantly related species, the giant and red pandas[27] (FIG. 4). In this example, two members of the Carnivora order have evolved similar limb specializations to harvest their particular bamboo diets. Using a combination of whole-genome sequencing methods and comparative techniques, it was shown that selection on the limb development genes dynein cytoplasmic 2 heavy chain 1 (*DYNC2H1*) and pericentrin (*PCNT*) had taken place independently in each species, with the resultant shared phenotype.

The first genome comparison between Eutherian and Metatherian species, which diverged ~65 million years ago, was carried out when the opossum, *Monodelphis domesticata*, was compared with human, dog and mouse[9]. In this study, true innovation in the protein-coding genes of opossum seemed relatively rare, although some gene families involved in environmental interactions showed rapid turnover. Importantly, ~20% of Eutherian conserved non-coding elements (CNEs) were revealed to be recent inventions partly arising from transposable elements. This implicated transposons as a major creative force in the evolution of mammalian gene regulation. Lineage comparisons of gene categories can uncover the evolution required for the increased complexity of species and their specialized behaviours (FIG. 1).

The phylogeny of 49 avian species, representing all orders of Neoaves, was re-examined in 2014 using whole-genome data to contrast the results generated with mitochondrial or partial gene conservation models[4,76,77]. This new phylogenetic tree helped to determine early splits in the avian phylogeny and deciphered independent lineages of divergently and convergently evolving land and water bird species. Nonetheless, resolving some of the branches in Neoaves still proved challenging, which could be due to massive protein-coding sequence convergence and high levels of incomplete lineage sorting that occurred during a rapid species radiation after the Cretaceous–Paleogene mass extinction event about 66 million years ago.

Macroevolution
Changes in allele frequencies that happen between species over a longer time period. This can be due to positive selection or drift.

Silent changes
Changes in DNA sequence without biological consequence.

Neutral sites
Positions in DNA that are not functional and hence are free to mutate randomly.

Conserved non-coding elements
(CNEs). Regions of the genome that are not coding for proteins, but are similar in many species, suggesting a role for these elements in genome regulation.

Transposable elements
Mobile DNA sequences, similar to viruses, that can 'jump' around in the genome and integrate in new locations. These sequences can affect gene expression or give rise to novel regulatory elements.

Box 2 | **Comparative canine genetics informs genome biology and disease**

The striking similarities between domestic dogs and humans, in terms of disease predisposition, genome organization and living environment, means that dogs are good models for human health. Using a key example, this box shows how comparative genomics can inform both canine biology and human disease.

Steroid-responsive meningitis–arteritis (SRMA) in Nova Scotia duck-tolling retrievers (NSDTRs) shows a strong clinical similarity to human systemic lupus erythematosus (SLE). Human genome-wide association studies (GWAS) have reported disease association for more than 40 loci[85], but there is still missing heritability to be found. GWAS in NSDTRs has implicated 11 genes on 5 different chromosomes, linking multiple genes involved in T cell activation to canine disease[86]. Intriguingly, the NSDTRs went through two extreme bottlenecks in the early 1900s following outbreaks for the distemper virus[87]. The few individuals that survived may have been those dogs with the strong T cell activation required to overcome the virus. This selection for T cell activation then formed the foundation of the current breed population.

Delving into the NSDTR GWAS, one of the top signals pointed to a ~1.5 Mb region on canine chromosome 32 (CFA32) spanning multiple genes, including B cell scaffold protein with ankyrin repeats 1 (*BANK1*). This gene encodes a B cell-specific scaffold protein that is involved in the mobilization of calcium from intracellular stores following B cell receptor signalling and is one of the strongest risk loci in human SLE[88]. The expression of *BANK1* is altered in both human patients with SLE and dogs with SRMA. However, an in-depth study of the 1.5 Mb region in dogs showed that multiple risk haplotypes correlate with both the expression of multiple genes and disease sub-phenotypes in an intricate manner[89].

This canine study paved the way for re-sequencing studies in human cohorts, in which the results of comparative genomics (for example, dog, mouse and human) were collated and used to target important genes and immunological pathways. These sequencing studies capture both rare and common, coding and non-coding variants from thousands of genes. This method was used to detect an association between BTB domain and CNC homologue 1 (*BACH2*) and Addison disease[90]. Twenty-six disease-associated variants were found in a ~200 kb haplotype spanning the gene start. All but two overlapped enhancer blood cell histone marks (for B and T cells) and four were associated with *BACH2* expression quantitative trait loci (eQTLs)[90]. Taken together, these canine and human studies highlight the benefits of comparative diseases genetics, the importance of selection for genome evolution, as well as the extreme complexity of genome function and regulation.

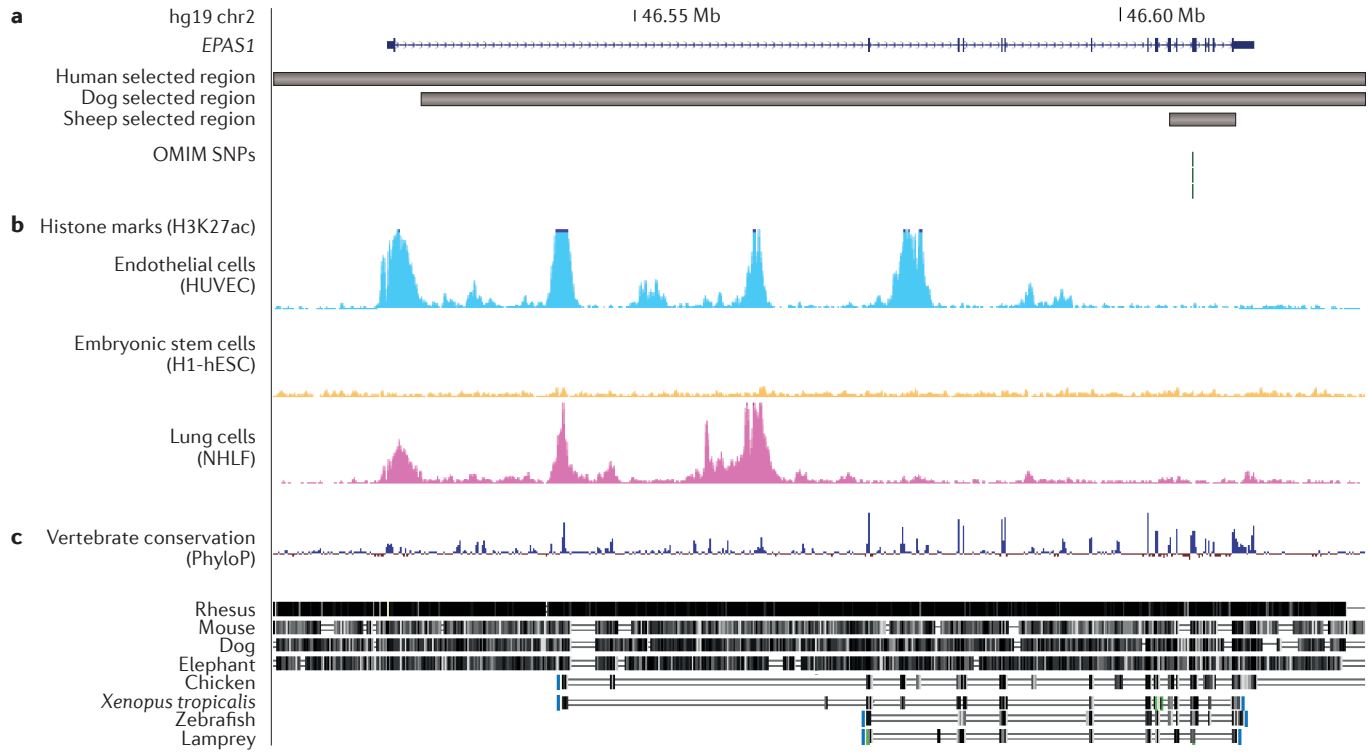

Figure 3 | **A case study in phenotypic adaptation: high altitude and *EPAS1*.** A key question in evolution regards which genes, or genomic regions, are involved in the successful colonization of one distinct environment from another. The comparison of populations adapted to low altitude versus those from the same species suited to high altitude has been investigated in numerous mammals, including sheep[74], dog and human[75]. The gene, endothelial PAS domain protein 1 (*EPAS1*), was implicated in all cases. **a** | In this figure, we use the LiftOver tool to place each result in the context of the human genome (hg19). Illustrated are the species-specific selective sweeps as well as coding and regulatory regions for human. In particular, single-nucleotide polymorphisms (SNPs) within the ovine sweep were shown to be significantly associated with corpuscular haemoglobin concentration[74], whereas for human, three functional missense mutations associated with erythrocytosis have been recorded in the Online Mendelian Inheritance in Man (OMIM). **b** | Regulatory marks, such as those indicating histone modifications (for example, histone 3 lysine 27 acetylation (H3K27ac)), can be used to infer in which tissue the gene is active. For example, peaks can be seen from endothelial (blue) and lung (pink) cells but not from embryonic stem cells (yellow). These marks can also indicate where regulation is acting in relation to the gene, for example, upstream of *EPAS1* and in the first intron of the gene. This information can be useful when interpreting the function of non-coding variants. **c** | High levels of conservation are maintained across mammalian lineages, suggesting that gene function and regulation may be maintained, and so what is learned from one mammal can be applied in a comparative analysis to another. Chr, chromosome; H1-hESC, normal human embryonic stem cells; HUVEC, human umbilical vein endothelial cell; NHLF, normal human lung fibroblasts; PhyloP, basewise conservation score calculated from Multiz alignment of 46 vertebrate species.

A similar themed analysis of 120 individuals, which represent all 22 of Darwin's finch species, found extensive evidence for interspecific gene flow throughout the species radiation. This also made the phylogenetic tree hard to determine. Although more recent hybridization events have given rise to species of mixed ancestry, the study was able to find a 240 kb haplotype that is strongly associated with beak shape diversity. This region encompassed ALX homeobox 1 (*ALX1*), a gene encoding a transcription factor that affects craniofacial development in humans[25].

Another interesting example of an interspecies genomic comparison involves the coelacanth, a lobe-finned fish thought to be lost to extinction 70 million years ago but rediscovered in 1938. This fish was originally thought to be the 'missing link' between fish and land-living tetrapods. Sequencing and analysis of the coelacanth genome and its relationship to mammals, birds and the lungfish led to the conclusion that the lungfish, and not the coelacanth, is the closest living relative of land-living tetrapods[78]. However, the coelacanth genome can still provide insight into vertebrate terrestrial adaptation. Intriguingly, coelacanth protein-coding genes are considerably more slowly evolving than those of tetrapods; however, analyses of changes in genes and regulatory elements during the vertebrate adaptation to land found an enrichment of genes involved in immunity, nitrogen excretion and the development of fins, tail, ears, eyes, brain and olfaction. Functional assays of enhancers involved in the fin-to-limb transition showed the importance of the coelacanth genome as an ancestral blueprint for understanding tetrapod evolution[78].

Positive selection
The force that makes certain genetic positions change in a certain favourable direction.

As shown above, large-scale comparative genomics between or within species has opened a Pandora's box for understanding genome evolution and genotype–phenotype correlations. There are good examples of these studies among mammals, birds and fish. As costs drop and long-read assemblies begin to reveal the more structurally complex regions of the genome, an even more detailed analysis will be possible.

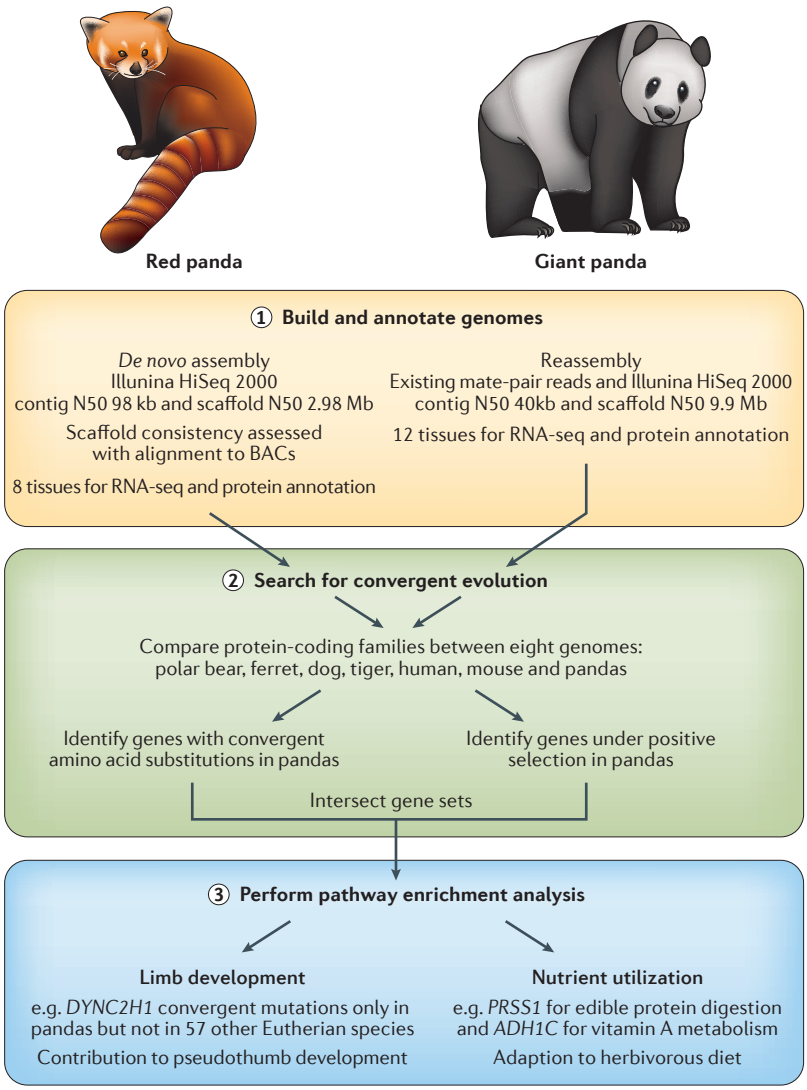

Figure 4 | **Comparative genomics for convergent evolution analyses.** In step 1, genomes from the red panda and the giant panda are built from scratch or reassembled and annotated using a combination of sequencing technologies and resources[27]. The quality of these assemblies is judged by comparison to other mammals in terms of length and gene content. Step 2 uses comparison to existing mammalian genomes to assess protein changes and positive selection in pandas. The goal is to find genes in which the two pandas match each other but are different from the six other Eutherian species studied. Convergent changes found with a small subset of mammals is validated by testing 59 Eutherian species and additional panda population members to ensure the result is not unique to one individual. In step 3, the final result is the identification of limb development and nutrient utilization genes that were independently mutated in each panda species but have resulted in the shared phenotypes required for a bamboo diet; namely, pseudothumb development for grasping stems and metabolic changes reflecting nutritional adaptation. *ADH1C*, alcohol dehydrogenase 1C; BAC, bacterial artificial chromosome; *DYNC2H1*, dynein cytoplasmic 2 heavy chain 1; N50, measure of contiguity; *PRSS1*, protease, serine 1; RNA-seq, RNA sequencing.

Annotating every base of mammalian genomes to unravel evolutionary mechanisms. The Genome 10K project was formed in 2010 to advocate for the sequencing of 10,000 vertebrate genomes[24]. Although the final goal remains elusive, if thousands of genomes are generated for this project, the power for interpreting genome structure and function will be immense, as the vast majority of physiological processes are shared across vertebrates. This means that any insight gained, be it from cats[79] or salamanders[32], can be compared with and contrasted to the human genome to inform physiology during states of health and disease.

The 29 mammals project[26] had the power to detect constraint at 12 bp resolution, identifying 10,000 genomic regions at which synonymous constraint overlapped with protein-coding exons. This provided a compendium of hundreds of candidate RNA structural families and nearly a million conserved transcription factor-binding sites, which intersected conserved candidate promoter, enhancer and insulator regions. The catalogue also revealed positive selection at the level of several amino acid residues, large numbers of novel mammalian lineage-specific non-coding elements and hundreds of accelerated regions in primates and humans.

An expanded 29 mammals team is currently collaborating with a large set of experts to collect, sequence and assemble an additional ~150 Eutherian species, which will complement the ~50 mammalian genomes available to create a set of 200 mammals. This enlarged data set includes more primates, rodents, lagomorphs and bats, and provides the increased sensitivity required for identifying and refining regions of genome constraint, as well as detecting accelerated regions on any of the lineages that may be of interest. Although the key focus is comparative analysis across the mammalian genome and among all the species, these novel genomes will be released for public interrogation during 2017. Similar collaborative efforts between Genome 10K investigators are under way for birds and fish.

The study of genetic diversity and conservation of endangered species. The International Union for Conservation of Nature (IUCN; accessed 2016) lists 205 critically endangered mammalian species, 30 of which are tagged as 'possibly extinct'. In addition, 783 mammalian species (14% of those evaluated) are listed as 'data deficient', meaning that there is insufficient information for a full assessment of conservation status. Similar numbers apply to other vertebrate groups. Much of the work involved in saving endangered species is related to protection of habitat and wild populations. However, the field of genomics is well placed to offer insight into genetic diversity and movement of different populations. These processes include isolation, in which populations are reproductively separated, be it by physical barriers or other mechanisms, drift, in which

Accelerated regions
Regions of the genome that are typically conserved across species, but where novel changes have happened in one or more related species. This suggests that the region is under positive selection for the novel variant (or variants).

population allele frequencies alter by random chance, and admixture, in which previously isolated populations have interbred.

Whole-genome sequencing can quickly survey the full genomic diversity of remaining individuals of a species or subspecies. For example, the global population of Northern white rhinoceros (NWRs; *Ceratotherium simum cottoni*) comprises only three individuals. The one male and two females probably lack the fitness to reproduce on their own. Propagating the species becomes complicated, involving artificial methods for generating a viable embryo and sourcing a suitable surrogate. It has been suggested that the sister species, the Southern white rhinoceros, may aid in the early stages of NWR recovery[80]. This recovery process involves understanding the level of diversity that exists in frozen tissues, cell lines and spermatozoa, so that the most diverse individuals are used for reproduction, combined with novel stem cell and assisted reproductive technologies[81].

The California condor is on the brink of extinction, with an embryonic lethal chondrodystrophy (a form of dwarfism affecting the skeleton) ravaging the population. Although genotyping a large pedigree at 17 microsatellite loci has provided a better assessment of the current population's genetic variation[82], developing genomic tools and resources (including a genetic map) is a prerequisite for the identification of candidate loci for this deadly disease. The understanding of this trait could lead to more comprehensive conservation plans and also provide clues for understanding the mechanisms that affect genetic variation, adaptation and evolution.

Conclusion and outlook

With 315 sequenced vertebrate genomes currently available in NCBI (National Center for Biotechnology Information) and less than 80 reference-quality vertebrate genomes available for query (for example, University of California, Santa Cruz (UCSC) Genome Browser), we have just scratched the surface of vertebrate genome biology (FIG. 1). As the lag between genome generation and genome publication closes, it will become more feasible to use the sequencing efforts of others to inform or augment the new sequencing efforts required to answer novel biological questions. In the future, this could mean hybrid assemblies of SRS and LRS, designed to be affordable but to balance the current need for information with long-term use. Consortia, such as 200 mammals or 1000 bulls, can help the community to drive research that answers questions on diverse species and targeted populations. The generation of high-quality genomes, or at least representative genomes, for many species will propel the understanding of genome content, enabling genotype versus phenotype correlations in conserved, divergent or accelerated regions. To map traits and to catalogue variation, different kinds of population and/or functional genomics data are essential. The year 2017 should see the number of vertebrate genomes that are available to the community increase to more than 600, emphasizing the need to work collaboratively and comparatively, sharing not only data but also tools and technology across species and projects.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Istrail, S. *et al.* Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc. Natl Acad. Sci. USA* **101**, 1916–1921 (2004).
3. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
4. Wade, C. M. *et al.* The mosaic structure of variation in the laboratory mouse genome. *Nature* **420**, 574–578 (2002).
5. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
This study describes the canine genome project, which addressed both comparative genome analysis and trait mapping in dogs.
6. Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–867 (2009).
7. Bovine Genome Sequencing and Analysis Consortium *et al.* The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–528 (2009).
8. Rhesus Macaque Genome Sequencing and Analysis Consortium *et al.* Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**, 222–234 (2007).
9. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).
10. Melé, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* **27**, 27–37 (2017).
11. Valouev, A. *et al.* A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* **18**, 1051–1063 (2008).
12. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
13. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol.* **12**, R112 (2011).
14. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
15. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
16. Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M. & Jaffe, D. B. Direct determination of diploid genome sequences. *Genome Res.* **27**, 757–767 (2017).
17. Lu, H., Giordano, F. & Ning, Z. Oxford nanopore minION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14**, 265–279 (2016).
18. Cao, H. *et al.* Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology. *Gigascience* **3**, 34 (2014).
19. Howe, K. & Wood, J. M. D. Using optical mapping data for the improvement of vertebrate genome assemblies. *Gigascience* **4**, 10 (2015).
20. Ganapathy, G. *et al.* High-coverage sequencing and annotated assemblies of the budgerigar genome. *Gigascience* **3**, 11 (2014).
21. Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable *de novo* reference assembly of the domestic goat genome. *Nat. Genet.* **49**, 643–650 (2017).
This paper presents an example of a hybrid reference genome, with particular attention paid to gains of continuity through a combination of sequencing methods.
22. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
23. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
This is an in-depth review of different sequencing technologies and their pros and cons.
24. Koepfli, K.-P., Paten, B. & O'Brien, S. J. The Genome 10K Project: a way forward. *Annu. Rev. Anim. Biosci.* **3**, 57–111 (2014).
25. Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature* **518**, 371–375 (2015).
26. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
This study presents the comparative analysis of 29 mammals to annotate the human genome.
27. Hu, Y. *et al.* Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl Acad. Sci. USA* **114**, 1081–1086 (2017).
This is an elegant paper that describes convergent evolution in two distantly related pandas.
28. Reichwald, K. *et al.* High tandem repeat content in the genome of the short-lived annual fish *Nothobranchius furzeri*: a new vertebrate model for aging research. *Genome Biol.* **10**, R16 (2009).
29. Carneiro, M. *et al.* Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. *Science* **345**, 1074–1079 (2014).
30. Rubin, C.-J. *et al.* Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587–591 (2010).
31. Alföldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011).
32. Newman, C. E., Gregory, T. R. & Austin, C. C. The dynamic evolutionary history of genome size in North American woodland salamanders. *Genome* **60**, 285–292 (2017).
33. Huang, H. W., NISC Comparative Sequencing Program, Mullikin, J. C. & Hansen, N. F. Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinformatics* **16**, 235 (2015).
This article presents an overview of variant detection methods that are used for sweep analysis.

34. Oleksyk, T. K., Smith, M. W. & O'Brien, S. J. Genome-wide scans for footprints of natural selection. *Phil. Trans. R. Soc. B Biol. Sci.* **365**, 185–205 (2010).
35. Weir, B. S. & Cockerham, C. C. Estimating F-statistics for the analysis of population structure. *Evolution* **38**, 1358 (1984).
36. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
37. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
38. Boyko, A. R. *et al.* Complex population structure in African village dogs and its implications for inferring dog domestication history. *Proc. Natl Acad. Sci. USA* **106**, 13903–13908 (2009).
39. Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).
40. Friedenberg, S. G. & Meurs, K. M. Genotype imputation in the domestic dog. *Mamm. Genome* **27**, 485–494 (2016).
41. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
42. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
43. Sargolzaei, M., Chesnais, J. P. & Schenkel, F. S. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478 (2014).
44. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nat. Genet.* **48**, 84–88 (2016).
45. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).
This study describes the single-molecule sequencing of a human genome, which enables the deciphering of both haplotypes and complex genomic regions.
46. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
47. Hoeppner, M. P. *et al.* An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS ONE* **9**, e91172 (2013).
48. Ramsköld, D., Kavak, E. & Sandberg, R. How to analyze gene expression using RNA-sequencing data. *Methods Mol. Biol.* **802**, 259–274 (2012).
49. Sandberg, R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods* **11**, 22–24 (2014).
50. Ricaño-Ponce, I. & Wijmenga, C. Mapping of immune-mediated disease genes. *Annu. Rev. Genom. Hum. Genet.* **14**, 325–353 (2013).
51. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
52. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297–1309 (2015).
53. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
This article describes the ENCODE project, in which functional elements are assigned to the human genome.
54. Andersson, L. *et al.* Coordinated international action to accelerate genome-to-phenome with FAANG, the Functional Annotation of Animal Genomes project. *Genome Biol.* **16**, 57 (2015).
55. Tuggle, C. K. *et al.* GO-FAANG meeting: a Gathering On Functional Annotation of Animal Genomes. *Anim. Genet.* **47**, 528–533 (2016).
56. Lonsdorf, E. V. *et al.* Socioecological correlates of clinical signs in two communities of wild chimpanzees (*Pan troglodytes*) at Gombe National Park, Tanzania. *Am. J. Primatol.* http://dx.doi.org/10.1002/ajp.22562 (2016).
57. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
58. Cockett, N. E. *et al.* Polar overdominance at the ovine *callipyge* locus. *Science* **273**, 236–238 (1996).
59. Hutchings, M. R., Knowler, K. J., McAnulty, R. & McEwan, J. C. Genetically resistant sheep avoid parasites to a greater extent than do susceptible sheep. *Proc. Biol. Sci.* **274**, 1839–1844 (2007).
60. Davis, B. W. & Ostrander, E. A. Domestic dogs and cancer research: a breed-based genomics approach. *ILAR J.* **55**, 59–68 (2014).
61. Karlsson, E. K. & Lindblad-Toh, K. Leader of the pack: gene mapping in dogs and other model organisms. *Nat. Rev. Genet.* **9**, 713–725 (2008).
62. Munson, L. & Moresco, A. Comparative pathology of mammary gland cancers in domestic and wild animals. *Breast Dis.* **28**, 7–21 (2007).
63. Menotti-Raymond, M. & O'Brien, S. J. in *Sourcebook of Models for Biomedical Research* (ed. Conn, P. M.) 221–232 (Humana Press, 2008).
64. Soares, M. *et al.* Molecular based subtyping of feline mammary carcinomas and clinicopathological characterization. *Breast* **27**, 44–51 (2016).
65. O'Neill, D. G. *et al.* Epidemiology of diabetes mellitus among 193,435 cats attending primary-care veterinary practices in England. *J. Vet. Intern. Med.* **30**, 964–972 (2016).
66. Lyons, L. A. *et al.* Whole genome sequencing in cats, identifies new models for blindness in AIPL1 and somite segmentation in HES7. *BMC Genomics* **17**, 265 (2016).
67. Yamamoto, J. K., Sanou, M. P., Abbott, J. R. & Coleman, J. K. Feline immunodeficiency virus model for designing HIV/AIDS vaccines. *Curr. HIV Res.* **8**, 14–25 (2009).
68. Vail, D. M. & MacEwen, E. G. Spontaneously occurring tumors of companion animals as models for human cancer. *Cancer Invest.* **18**, 781–792 (1999).
69. Axelsson, E. *et al.* The genomic signature of dog domestication reveals adaptation to a starch-rich diet. *Nature* **495**, 360–364 (2013).
70. Andersson, L. S. *et al.* Mutations in *DMRT3* affect locomotion in horses and spinal circuit function in mice. *Nature* **488**, 642–646 (2012).
71. Petersen, J. L. *et al.* Genome-wide analysis reveals selection for important traits in domestic horse breeds. *PLoS Genet.* **9**, e1003211 (2013).
72. Promerová, M. *et al.* Worldwide frequency distribution of the 'gait keeper' mutation in the *DMRT3* gene. *Anim. Genet.* **45**, 274–282 (2014).
73. Lamichhaney, S. *et al.* Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. *Proc. Natl Acad. Sci. USA* **109**, 19345–19350 (2012).
74. Wei, C. *et al.* Genome-wide analysis reveals adaptation to high altitudes in Tibetan sheep. *Sci. Rep.* **6**, 26770 (2016).
75. Wang, G.-D. *et al.* Genetic convergence in the adaptation of dogs and humans to the high-altitude environment of the Tibetan plateau. *Genome Biol. Evol.* **6**, 2122–2128 (2014).
76. Zhang, G. *et al.* Comparative genomic data of the Avian Phylogenomics Project. *Gigascience* **3**, 26 (2014).
77. Jarvis, E. D. *et al.* Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).
78. Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).
79. Montague, M. J. *et al.* Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc. Natl Acad. Sci. USA* **111**, 17230–17235 (2014).
80. Saragusty, J. *et al.* Rewinding the process of mammalian extinction. *Zoo Biol.* **35**, 280–292 (2016).
81. Ben-Nun, I. F. *et al.* Induced pluripotent stem cells from highly endangered species. *Nat. Methods* **8**, 829–831 (2011).
82. Romanov, M. N. *et al.* The value of avian genomics to the conservation of wildlife. **10** (Suppl. 2), S10 (2009).
83. Andrén, T. *et al.* in *The Baltic Sea Basin* (eds Harff, J., Björck, S. & Hoth, P.) 75–97 (Springer Berlin Heidelberg, 2011).
84. Martinez-Barrio, A. *et al.* The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. *eLife* **5**, e12081 (2016).
85. Cui, Y., Sheng, Y. & Zhang, X. Genetic susceptibility to SLE: recent progress from GWAS. *J. Autoimmun.* **41**, 25–33 (2013).
86. Wilbe, M. *et al.* Genome-wide association mapping identifies multiple loci for a canine SLE-related disease complex. *Nat. Genet.* **42**, 250–254 (2010).
87. Strang, A. 1. & Macmillan, G. *The Nova Scotia Duck Tolling Retriever* (Loveland, 1996).
88. Kozyrev, S. V. *et al.* Functional variants in the B cell gene *BANK1* are associated with systemic lupus erythematosus. *Nat. Genet.* **40**, 211–216 (2008).
89. Wilbe, M. *et al.* Multiple changes of gene expression and function reveal genomic and phenotypic complexity in SLE-like disease. *PLoS Genet.* **11**, e1005248 (2015).
90. Eriksson, D. *et al.* Extended exome sequencing identifies *BACH2* as a novel major risk locus for Addison's disease. *J. Intern. Med.* **280**, 595–608 (2016).
91. Denas, O. *et al.* Genome-wide comparative analysis reveals human-mouse regulatory landscape and evolution. *BMC Genomics* **16**, 87 (2015).

Acknowledgements
J.R.S.M. was supported by the Swedish Research Council, FORMAS (221-2012-1531). K.L.-T. was supported by the Swedish Research Council, European Research Council (ERC) Starting Grant and Knut och Alice Wallenberg Foundation.

Competing interests statement
The authors declare no competing interests.

DATABASES
1000 Bulls Genome: www.1000bullgenomes.com
Dog 10K genome project: dog10kgenomes.org
Encyclopedia of DNA Elements (ENCODE): www.encodeproject.org/
ExAC consortia database: exac.broadinstitute.org
Genome 10K consortium: genome10k.soe.ucsc.edu
RefSeq: www.ncbi.nlm.nih.gov/refseq/
Nature Encode Explorer: www.nature.com/encode/
NCBI: www.ncbi.nlm.nih.gov/
The Functional Annotation of Animal Genomes consortium (FAANG): www.faang.org/
The International Union for Conservation of Nature: www.iucnredlist.org/about/summary-statistics
UCSC Genome Browser: genome.ucsc.edu/
UK10K database: www.uk10k.org/

FURTHER INFORMATION
CRISP: github.com/vibansal/crisp/
LoFreq: csb5.github.io/lofreq/
National Human Genome Research Institute 1991 financial year: www.genome.gov/11006943/
National Human Genome Research Institute sequencing costs: www.genome.gov/sequencingcostsdata/
Online Mendelian Inheritance in Animals (OMIA): omia.angis.org.au
Online Mendelian Inheritance in Man (OMIM): www.omim.org
VarScan: varscan.sourceforge.net/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF