Review

CellPress

It's more than stamp collecting: how genome sequencing can unify biological research

Stephen Richards

Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

The availability of reference genome sequences, especially the human reference, has revolutionized the study of biology. However, while the genomes of some species have been fully sequenced, a wide range of biological problems still cannot be effectively studied for lack of genome sequence information. Here, I identify neglected areas of biology and describe how both targeted species sequencing and more broad taxonomic surveys of the tree of life can address important biological questions. I enumerate the significant benefits that would accrue from sequencing a broader range of taxa, as well as discuss the technical advances in sequencing and assembly methods that would allow for wide-ranging application of whole-genome analysis. Finally, I suggest that in addition to 'big science' survey initiatives to sequence the tree of life, a modified infrastructure-funding paradigm would better support reference genome sequence generation for research communities most in need.

Biology fundamentals from the genome reference

Freely available whole-genome reference sequences – the genome sequences in the public domain (Table 1) with annotated gene models and viewable in browsers – have been so immensely successful, valuable, and accessible that they are now taken for granted in many research communities. Despite what is clearly a paradigm shift, the number of available sequences is actually low, and access to well-annotated genomes is limited. For example, some relatively common model organisms have only incomplete or poorly annotated genomes, such as maize, and others have no publicly available genome, including *Xenopus laevis*; the sequence of which is still awaiting publication. Here, I propose that additional references surveying the tree of life are a necessary foundation for the study of biology in the 21st century and will enable biology to transcend its observational roots and become more of an engineering discipline. I begin by illustrating the extent of the transformation genome references enable in biology by noting the successes and techniques brought about by the sequencing of the human genome. I then discuss how reference genome sequences could bring about a similar revolution for the remainder of the tree of life.

In assessing the impact of the human reference sequence, it is instructive to remember a time when the number of human protein coding genes was thought to be as high as 120 000 (although sensible approaches placed the number lower [1]). A GeneSweep pool [2] was held at the 2000 Cold Spring Harbor Laboratory Biology of Genomes meeting, and all estimates of human gene number – by the world's assembled genomics experts – were significantly higher than the actual number revealed in 2003, which has since been refined down further [3]. The genome sequencing revolution is still in its infancy; however, we must acknowledge it as the major driver of biology since the start of the 21st century. Much of the credit for these successes is due to the US National Human Genome Research Institute (NHGRI) and its surrounding community, whose leadership has driven sequencing technology, investigation of genome biology, and general human and model organism biology for the past two decades.

Reference genomes also enable analysis of RNA sequencing (RNAseq) data. In the human genome, we now contrast protein coding sequence comprising ~1% of the genome with extensive transcription of large amounts of the genome and the assignment of activity to as much as possibly 80% of the genome [4]. Combining RNAseq and a reference to align those data enabled the discovery of new classes of noncoding RNA such as ~8000 human long noncoding RNAs (lncRNAs) [5]. The genome sequence is also the structural framework for the transcriptional machinery and the source of information to be transcribed. The ENCODE project extended our functional understanding of the human reference genome by annotating transcription-factor binding sites, enhancers, chromatin accessibility and modification patterns, and the identification of expression quantitative trait loci (eQTLs). These have facilitated deeper understanding of epigenetic regulation of RNA processing, noncoding RNA, and regulatory networks, and sparked the growing appreciation for the importance of the 3D structure of the functioning cellular genome [4]. Overall, observational descriptions of the human genome have resolved previous misunderstandings (such as gene number) and unknowns (such as transcriptional capability), but most importantly they provide the

Corresponding author: Richards, S. (stephenr@bcm.edu).

Keywords: genome reference sequences; tree of life; taxanomic genome surveys; genome infrastructure funding; DNA sequencing; Long read sequencing technologies; genome assembly; genome analysis; eukaryotic genomics; eukaryotes.

0168-9525/

 http://dx.doi.org/10.1016/j.tig.2015.04.007

Table 1. Where are the reference genomes?

Focus	Database	URL	Notes
All Sequences	NCBI Genbank	www.ncbi.nlm.nih.gov/genbank/	The International Nucleotide Sequence Database Collection (INSDC) collects all sequences
	EMBL-ENA	www.ebi.ac.uk/ena	
	DNA Databank of Japan	www.ddbj.nig.ac.jp	
Genome Annotation Portals	Ensemble Genomes	http://ensemblgenomes.org	
	NCBI-Refseq/entrez	www.ncbi.nlm.nih.gov/refseq/	
Example Large Community based Databases	UCSC Genome Browsers	http://genome.ucsc.edu	Focused on mammals
	Mouse Genome Informatics	www.informatics.jax.org	These model organism based databases link genome and gene sequences to other reagents and mutant lines, publications and, for *E. coli*, systems biology
	Flybase	http://flybase.org	
	Wormbase	www.wormbase.org	
	Saccharomyces Genome db	www.yeastgenome.org	
	EcoCyc *Escherichia coli* database	http://ecocyc.org	
	Plant genome database	www.plantgdb.org	
Ortholog databases	OrthoDB	http://orthodb.org	Rapid lookup of orthologous genes across many species
	PhylomeDB	http://phylomedb.org	

necessary foundation for current and future progress in fundamental biology and clinical medicine.

Technology designed around the human reference genome leads the way

Humans, like much of the tree of life, do not share the traits of classic genetic models such as *Drosophila*, mice, and yeast, which have short life spans and whose gene expression can be experimentally controlled. Thus, human genetic analyses based on short-read alignment to reference genomes are directly applicable to the majority of species. For example, resequencing a single patient can identify natural Mendelian causative alleles or *de novo* mutations. Sequencing 2000 exomes from patients referred to a medical genetics clinic led to a diagnosis for 25% of patients [6]. Genome sequencing of individuals is routine in model organisms [7–9], but has also been used for other species, such as dogs [10], where it was used to identify mutations underlying the neurodegenerative disorder neuronal ceroid lipofuscinosis, and shed light on the same disease in humans. Genome-wide association studies (GWASs) based on single nucleotide polymorphism, exome, and genome sequencing of cohorts have contributed to our understanding of complex disease genetics identifying >15 000 regions associated with the majority of common human diseases [11]. GWASs are also applicable to quantitative traits in non-model species including crops [12] and farm animals for agricultural traits such as fertility and milk production [13,14]. Single-cell sequencing and alignment of resultant short reads to the human reference has primarily been used to understand how mutation variation and mutant cell lineage within human tumors affects cancer treatment [15]. The same technique also enables molecular study of individual microbial species that cannot be grown outside of microbial communities [16]. Population sequencing can identify (and sometimes date) recent selection on genomes such as altitude adaptation (for review, see [17]) and convergent adaptation of human lactase persistence in both Africa and Europe ~7000 years ago [18]. In birds, population sequencing associated selection of the ALX1 crainofacial transcription factor to beak shape, clarifying species delineations in Galapagos island finches [19]. Sequencing domesticated dog populations identified selection on nervous system development genes for behavior and genes enabling adaptation to a starch-rich diet; both crucial for domestication [20]. Genome sequencing of ancient Neanderthal DNA [21] identified remnants of historical gene flow from Neanderthal, Denisovan populations, and possibly *Homo erectus*, into *Homo sapiens*. Similarly, investigation of small genomic regions containing yellow skin chicken domestication genes in DNA from 280 BC dated fixation of domestication alleles to the last 500 years [22]. Sequence from a 600 000-year-old horse bone preserved in permafrost [23] changed divergence time estimates for the horse lineage, and identified putative domestication loci.

A small sampling of life

A measure of the incredible success of genome references is that for many researchers their availability is taken for granted: it is assumed that the sequence of gene X, its paralogs, alternative splice forms, and its chromosomal location are all known. It is important to remember, however, that the vast majority of species cannot be studied effectively due to lack of a genome reference. The extent of reference sequence coverage of the eukaryotes is shown in Figure 1. Within the relatively well-studied vertebrates, fifty percent of primate families have a reference, comprehensive sampling of bird species has recently started [24], and the mammals are well covered, but reptiles and amphibians have few genome references. Outside of the vertebrates, there is a dearth of genomes throughout the tree of life. Approximately half of the insect orders have no representative genome. The water flea *Daphnia* [25] has the only high-quality crustacean genome available. The myrapods are represented by a single centipede genome [26], Chicilierates (spiders, mites and ticks) are currently represented by only three published genomes: the agricultural pest spider mite [27], a social spider, and tarantula [28]. Outside of the arthropods, invertebrate genome representation drops again. While there is at least one or two of each invertebrate phylum, it is the equivalent of having a chicken and a fish sequence as the closest representative to the human sequence. For example, the mollusks, among the most diverse animal phyla, are currently represented by a limpet, a polychaete, and a leech [29]. While this is a start (and an excellent scientific paper), it is not useful for those studying cephlapods such as octopi and cuttlefish for their alien intelligence, liquid crystal display skins,