News & views

Human genome

Bridging the gaps

Karen H. Miga

Since the human genome was published in 2001, many of the gaps in the original sequence have been filled in, offering a more detailed understanding of genome regulation, structure and function.

The release of drafts of the human genome in 2001 was a landmark achievement[1,2]. Scientists could, for the first time, study long stretches of each human chromosome, base by base. As such, researchers could begin to understand how individual genes were ordered, and how the surrounding non-protein-coding DNA was structured and organized. Despite this amazing progress, the draft genomes were still incomplete, with more than 150 million bases missing[3]. Technological advances in the intervening years have allowed researchers to add to the draft, with the complete sequencing of a chromosome finally being achieved[4,5] in 2020. As a result, new and uncharacterized parts of the human genome are beginning to surface, ushering in another exciting period of biological discovery.

What exactly was included in the draft genomes? The original draft contained many previously unexplored intergenic regions. It also encompassed the vast majority of genes. The International Human Genome Sequencing Consortium[1] initially estimated that the genome contained 30,000–40,000 protein-coding genes, although the publication of an updated genome[6] in 2004, along with improved gene-prediction approaches[7], led the figure to be revised to about 20,000. The 2004 genome gave a high-resolution map of 2.85 billion nucleotides from euchromatin — the more loosely packaged regions of DNA, which are enriched in genes and make up roughly 92% of the human genome.

The reference genome launched the scientific community into an era of genome exploration, shifting the focus from single genes to more-complete, genome-wide studies. However, gaps remained on each of the 23 pairs of human chromosomes, estimated to contain more than 150 megabases of unknown sequence[3] (Fig. 1). The largest gaps were at locations enriched with highly repetitive DNA or sequences for which there are many near-identical copies. These sections were originally difficult to clone, sequence and correctly assemble. As a result, the human genome project intentionally under-represented these repetitive sequences. Although researchers had a very basic idea of the nature of sequences in these regions, the regions' high-resolution genomic organization remained elusive.

Early attempts to close the gaps used long sequence reads to span the repetitive sequences — but such reads were initially highly error-prone. In the 2010s, new opportunities arose, thanks to advances in the ability to read longer stretches of sequence (outlined in refs. 8 and 9, for instance), along with the development of scalable bioinformatic tools. Sequence reads of tens to hundreds of kilobases allowed the study of the genomic organization of many moderately sized gaps. This provided insights into some subtelomeric regions[9] — repeat-rich DNA adjacent to the telomere structures that cap the ends of chromosomes. It also enabled the study of the first centromeric satellite array[10], in which short sequences are repeated in tandem for about 300 kilobases. A subset of segmental duplications (stretches of sequence that share 90–100% of their bases and are found in multiple locations) was also resolved, many containing genes previously missing from the reference genome[9,11]. However, many of the largest, multi-megabase-sized repeat-rich regions remained intractable.

Over the past few years, the combination of both ultra-long reads[9] and highly accurate long-read data[12] has proved a game-changer for resolving these regions[13,14], revealing, for the first time, extremely long tracts of tandem repeats and regions enriched in segmental duplications. By breaking down these technological barriers, scientists are now discovering extensive repeat-rich regions that can span millions of bases, and make up the entire short arms of chromosomes.

Researchers do not yet fully understand why parts of the human genome are organized in this way. But gaining such an understanding will undoubtedly be valuable, because these repeat-rich sequences are often positioned at sites that are crucial for life. For example, long tracts of ribosomal DNA (rDNA) repeats encode RNA components of the cell's protein-synthesizing machinery and have an important role in nuclear organization[15]. And the repetitive DNA of structures called centromeres is essential for proper chromosome segregation during cell division[16].

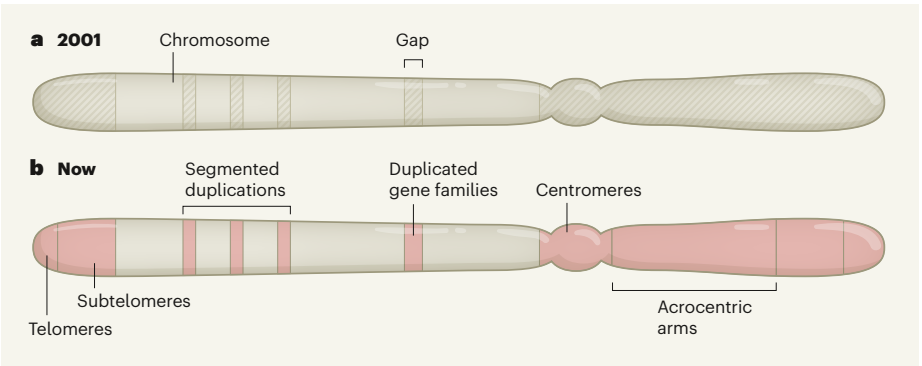

Figure 1 | Filling in the missing sequence in the human genome. **a**, The 2001 draft human genome[1,2] covered most of the gene-rich DNA, which is loosely packaged in the nucleus. But many gaps remained in tightly packaged regions rich in repetitive DNA sequences, which are often untranscribed (the overall extent of the gaps is exaggerated here, for ease of interpretation). **b**, Thanks to advances in sequencing and bioinformatics, researchers can now study all of these missing sequences. These include the telomere and subtelomere regions that cap chromosomes; centromere structures that are essential for cell division; and particularly short and highly repetitive chromosome arms known as acrocentric arms. Regions in which DNA is duplicated, either in one location or in a segmented way, can also now be analysed.

These large swathes of repetitive DNA come with different sets of rules, in terms of their genomic organization and evolution. They are also subject to different epigenetic regulation (molecular modifications to DNA and associated proteins that do not alter the underlying DNA sequence), which leads repetitive DNA to differ from euchromatin in terms of its organization, replication timing and transcriptional activity[17–19]. Many genome-wide tools and data sets cannot yet fully capture all this information from extremely repetitive DNA regions, and so scientists do not yet have a complete picture of what transcription factors bind to them, how these regions are spatially organized in the nucleus, or how regulation of these parts of our genome changes during development and in disease states. Now, much like the initial release of the genome decades ago, researchers are faced with a new, unexplored functional landscape in the human genome. Access to this information will drive technology and innovation to be inclusive of these repeat regions, once again broadening our understanding of genome biology.

In the past year, scientists have used extremely long and highly accurate sequence reads to reconstruct entire human chromosomes from telomere to telomere[4,5]. Last year also saw the release of a near-complete human reference genome from an effectively 'haploid' human cell line, with only five remaining gaps that mark the sites of rDNA arrays (go.nature.com/3rgz93y). In this line, cells have two identical pairs of chromosomes, simplifying the challenge of repeat assembly compared with typical human cells (which are diploid, with different chromosomes inherited from the mother and father). These maps together offer the first high-resolution glimpse of centromeric regions, segmental duplications, subtelomeric repeats and each of the five acrocentric chromosomes, which have very short arms made up almost entirely of highly repetitive DNA at one end.

It is tempting to think scientists are finally approaching the finish line. However, a single genome assembly, even if complete with near-perfect sequence accuracy, is an insufficient reference from which to study the sequence variation that exists across the human population. Existing maps that chart the diversity across the euchromatic parts of the genome must be extended to fully capture repetitive regions, where copy number and repeat organization vary between individuals. Doing so will require the development of strategies for routine production and analysis of complete human diploid genomes. The aspirational goal of reaching a more-complete and comprehensive reference of humanity will undoubtedly improve our understanding of genome structure and its role in human disease, and align with the promise and legacy of the Human Genome Project.

Karen H. Miga is at the UC Santa Cruz Genomics Institute, University of California, Santa Cruz, California 95064, USA.
e-mail: khmiga@ucsc.edu

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al. Science* **291**, 1304–1351 (2001).
3. Schneider, V. A. *et al. Genome Res.* **27**, 849–864 (2017).
4. Miga, K. H. *et al. Nature* **585**, 79–84 (2020).
5. Logsdon, G. A. *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.09.08.285395 (2020).
6. International Human Genome Sequencing Consortium. *Nature* **431**, 931–945 (2004).
7. Frankish, A. *et al. Nucleic Acids Res.* **47**, D766–D773 (2019).
8. Koren, S. *et al. Nature Biotechnol.* **30**, 693–700 (2012).
9. Jain, M. *et al. Nature Biotechnol.* **36**, 338–345 (2018).
10. Jain, M. *et al. Nature Biotechnol.* **36**, 321–323 (2018).
11. Chaisson, M. J. P. *et al. Nature* **517**, 608–611 (2015).
12. Wenger, A. M. *et al. Nature Biotechnol.* **37**, 1155–1162 (2019).
13. Nurk, S. *et al. Genome Res.* **30**, 1291–1305 (2020).
14. Bzikadze, A. V. & Pevzner, P. A. *Nature Biotechnol.* **38**, 1309–1316 (2020).
15. Zentner, G. E., Saiakhova, A., Manaenkov, P., Adams, M. D. & Scacheri, P. C. *Nucleic Acids Res.* **39**, 4949–4960 (2011).
16. Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K. & Willard, H. F. *Science* **294**, 109–115 (2001).
17. Sullivan, B. A. & Karpen, G. H. *Nature Struct. Mol. Biol.* **11**, 1076–1083 (2004).
18. Gilbert, D. M. *Curr. Opin. Cell Biol.* **14**, 377–383 (2002).
19. Prior, C. P., Cantor, C. R., Johnson, E. M., Littau, V. C. & Allfrey, V. G. *Cell* **34**, 1033–1042 (1983).

Human genome

A genetic revolution in rare-disease medicine

Fowzan S. Alkuraya

Mendelian diseases are caused by mutations in a single gene. The first draft of the human genome, published in 2001, had broad implications for how these diseases are diagnosed, managed and prevented.

When the first draft of the human genome was published[1,2], it was expected to have a transformative impact on medicine. Bold predictions were made about a paradigm shift in which medicine became personalized, predictive and preventive[3]. To many, no such transformation materialized, probably because of a focus on common diseases such as diabetes and coronary artery disease. But the predictions were right on target for Mendelian diseases — those caused by mutations in single genes — such as hereditary cancers and many forms of developmental delay in children.

"The true game-changer came when the draft genome was used in combination with 'next-generation' sequencing technologies."

Before the draft genome, basic information about the sequence and genomic location of a mutated gene had to be worked out through a process called cloning, in which short chromosomal segments were cut from human DNA using enzymes, and replicated in bacteria to produce sufficient quantities for analysis. Cloning was a stupendously laborious exercise that often took years and could be performed by only a few laboratories. The genetic underpinnings of most Mendelian diseases were therefore unknown, making diagnosis extremely difficult. Even for the few that did have a known underlying genetic basis (such as fragile X syndrome), a specialist was still likely to fail to make a diagnosis, because of the remarkable variability of the diseases' clinical presentation and their rarity[4].

In the 1990s, the development of 'positional mapping' methods made it easier to identify genes associated with Mendelian diseases. Early positional-mapping efforts involved comparing the DNA of several people who had the same disease, using a primitive genome map containing a few known sequences that vary between individuals; these acted as location markers to help researchers zero in on a candidate disease-causing region[5]. The primitive map, which dates back to 1987, was essential to early gene-discovery efforts. Nonetheless, its low resolution was a major obstacle to gene-discovery efforts.

It is hard, then, to overstate just how influential the human genome draft was for people with Mendelian diseases and their families. The draft did not directly link individual genes to diseases, but it did provide the necessary elements for a revolution in diagnosis. Initially, it provided a rich map of markers that permitted a much higher resolution in positional mapping. However, the true game-changer