Human genome

Expanding diversity in genomics

Charles N. Rotimi & Adebowale A. Adeyemo

In the 20 years since the first drafts of the human genome were made public, an explosion in genome sequencing has revealed how our evolutionary history and health can be understood by analysing the diversity in our genomes.

The successful sequencing of the human genome[1,2] in 2001 is considered by many to be one of the greatest achievements in biology. The published sequences were generated from the DNA of a few anonymous volunteers of differing ethnic backgrounds. However, a single genome (even one generated from many individuals) can provide only so much information. It was immediately clear that we would need to generate and compare more sequences from different people, if we were to harness information coded in genomes to better understand our health and heritage. So far, we have genomes for hundreds of thousands of individuals — more than was imaginable 20 years ago. Even so, we are just beginning to sequence diverse populations in the numbers needed to realize the promise of genomics.

Although human genomes are 99.9% similar, they also contain millions of single nucleotide polymorphisms (SNPs) — single bases where there is genetic variation between individuals. A map of about 1.42 million SNPs was published alongside the draft genome[3], generated in part from differences found between the individuals who contributed their DNA for the draft. Thus, the Human Genome Project provided a framework for larger-scale projects to analyse human variation.

In 2003, a consortium of researchers set out to generate a genetic map of SNPs from diverse individuals — an endeavour known as the International HapMap Project[4]. The first iteration of the map, published in 2007, was a major milestone that documented more than 3 million SNPs discovered in 270 individuals from Japan, China, the United States and Nigeria[5]. The work shed light on how the genome is organized, revealing how segments of our DNA are inherited together as blocks, and highlighting how these blocks vary within and between populations. The HapMap was eventually expanded to include 11 population groups[6], emphasizing differences in the way in which common human genetic variants (HGV) are distributed worldwide.

The HapMap project also aided the development of biotechnological and computational approaches such as genome-wide association studies (GWAS), which allow scientists to search thousands of individual genomes to discover genetic variants that are linked to specific traits. GWAS have successfully identified genomic regions that increase the risks of common conditions such as diabetes, coronary artery disease and Crohn's disease[7]. But GWAS have been performed mainly in people of European ancestry[7], and as of December 2020, 78% of individuals in all GWAS were of such ancestry (go.nature.com/3ocyhql). Several factors account for this bias, including a reliance on existing cohorts, preference for homogeneous population groups, limited funding for enrolling under-represented groups and early perceptions that findings from Europeans should be generalizable to other groups. The lingering lack of diversity in GWAS has been highlighted as one of the main roadblocks to the scientific and equitable realization of the promise of genomics[8,9].

The 1000 Genomes Project was created in 2008 to generate a more comprehensive catalogue of HGV by systematically sequencing the genomes of thousands of individuals from diverse geographical locations, to identify both common and rare genetic variants[10]. Because of the ever-diminishing cost of sequencing, by its completion, the project had amassed 2,504 individuals from 26 population groups on 5 continents (including several groups with mixed ancestries), providing a detailed catalogue of genetic variants on a scale previously unimaginable.

The data generated led to an unprecedented range of discoveries about the global distribution of HGV. For instance, it emerged that most common variants are shared globally, but rarer variants are shared by closely related populations, with 86% of rare variants restricted to a single continental group. The project also confirmed that there is greater genetic diversity in African populations than in other groups.

The small group of humans that left Africa about 100,000 years ago to populate the rest of the world carried only a subset of the variations that existed at the time; this means that the subset of HGV left behind can be studied only in Africans[11,12]. Africa is historically under-represented in genomic studies owing to inadequate funding and investment by African governments. Until a few years ago, there was a limited number of African scientists with genomics expertise, and inadequate biomedical research and computational infrastructure. Last year, the Human Heredity and Health in Africa (H3Africa) consortium, of which we are both members, reported whole-genome sequences

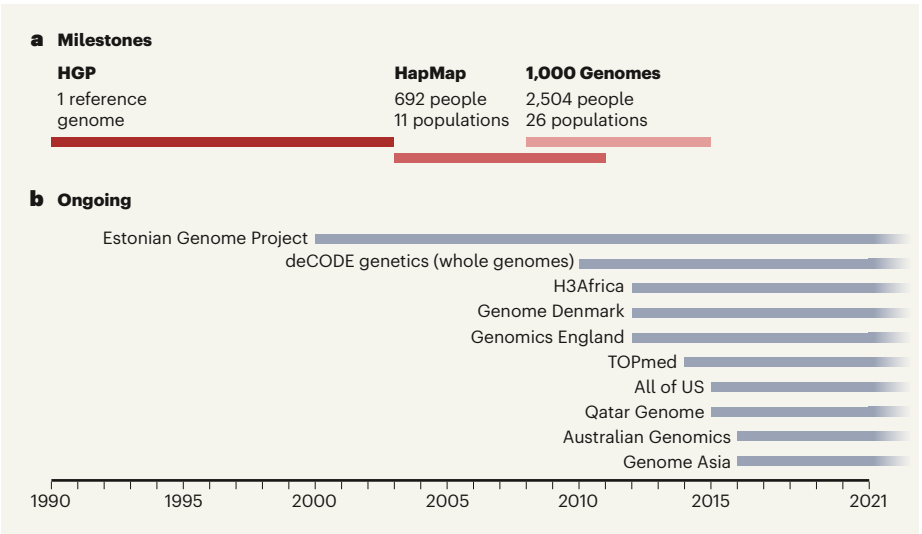

Figure 1 | Increasing diversity in genomics. a, The Human Genome Project (HGP) was established in 1990 and completed in 2003, with the first draft of the human genome[1,2] published in 2001. Since then, collaborative efforts have resulted in the analysis of large numbers of genomes from increasingly diverse populations. Milestones of note include the International HapMap Project[4] and the 1000 Genomes Project[8]. **b**, Today, there are many ongoing projects to sequence populations around the world.

of 426 individuals from 50 ethnolinguistic groups in Africa. H3Africa discovered more than three million variants[13] — mainly in previously unrepresented ethnolinguistic groups. It also observed complex patterns in mixing of ancestries and identified 62 regions of the genome that have been evolutionarily maintained at high frequency, perhaps because of protective roles in viral immunity, DNA repair and metabolism.

These findings highlight — as we and others[8] have argued for years — the need to increase diversity in genome science (Fig. 1). Clearly, Eurocentric studies will not be broadly applicable to all populations. Some disease risk variants are specific to certain populations, and polygenic risk scores (which quantify the risk that an individual will develop a given trait or disease, based on the aggregate or sum of variants they carry) might not generalize well across multiple populations[9,14–16]. Type 2 diabetes is a common disease that demonstrates this observation. Despite a set of well-known risk variants that are shared across populations, seemingly population-specific variants have been identified in East Asian, Mexican and African groups[15,16].

Understanding how differences between our genomes cluster according to the ancestral backgrounds of individuals and groups is undoubtedly valuable. However, inferred clusters might not overlap with social descriptors such as 'Black', 'Latino', 'Asian' and 'European' — an assumption that has been used by some to justify racial categorization[17]. The best evidence so far suggests that social categories and genetic clusters are inconsistent[17,18]. Indeed, one study[19] that identified 21 global ancestries reported that its 6,000 individuals had, on average, DNA from 4 ancestries. This indicates the need for caution when using labels such as African/Black, Hispanic/Latino, Asian or European/white in genome science. Indeed, the use of these terms in genome science should be discouraged except as self-reported descriptors or to provide socio-demographic context. Using these terms risks distorting our understanding of the distribution of HGV in history and health.

In future, we will increasingly use genomics to understand our evolutionary history, predict individual disease risks, develop therapeutics such as vaccines, and cure diseases including sickle-cell anaemia using DNA-editing technologies. To fully realize these expectations, we must address several continuing challenges, including, but not limited to, three factors related to diversity. The first is increasing the participation of individuals from diverse ancestral backgrounds in genome research. This challenge is being met through support for genomics research and capacity building from both large consortia (such as the H3Africa consortium) and national genome projects in under-studied populations. The second is developing global collaborations to establish crucial inter-country biomedical infrastructure, ethical frameworks and equitable data sharing — common barriers to international collaboration. The third is equitable deployment of genomic advances to avoid exacerbating health disparities, especially in resource-challenged settings across the world.

Achieving these goals will greatly improve our knowledge of human genetic diversity, aid disease-gene discovery efforts and facilitate our understanding of human biology. The road from one genome reference to hundreds of thousands of genomes has provided unprecedented insight into human genetic variation and the complex tapestry of our ancestry, leading to many practical scientific and medical benefits. Making these benefits available to all humanity is the next frontier.

Charles N. Rotimi and **Adebowale A. Adeyemo** are at the Center for Research on Genomics and Global Health, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.
e-mails: rotimic@nih.gov; adeyemoa@nih.gov

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al. Science* **291**, 1304–1351 (2001).
3. The International SNP Map Working Group. *Nature* **409**, 928–933 (2001).
4. The International HapMap Consortium. *Nature* **426**, 789–796 (2003).
5. The International HapMap Consortium. *Nature* **449**, 851–861 (2007).
6. The International HapMap 3 Consortium. *Nature* **467**, 52–58 (2010).
7. Popejoy, A. B. & Fullerton, S. M. *Nature* **538**, 161–164 (2016).
8. Bustamante, C. D., De la Vega, F. M. & Burchard, E. G. *Nature* **475**, 163–165 (2011).
9. Martin, A. R. *et al. Nature Genet.* **51**, 584–591 (2019).
10. The 1000 Genomes Project Consortium. *Nature* **526**, 68–74 (2015).
11. McClellan, J. M., Lehner, T. & King, M.-C. *Cell* **171**, 261–264 (2017).
12. Rotimi, C. N. *et al. Hum. Mol. Genet.* **26**, R225–R236 (2017).
13. Choudhury, A. *et al. Nature* **586**, 741–748 (2020).
14. Genovese, G. *et al. Science* **329**, 841–845 (2010).
15. Adeyemo, A. A. *et al. Nature Commun.* **10**, 3195 (2019).
16. The SIGMA Type 2 Diabetes Consortium. *Nature* **506**, 97–101 (2014).
17. Keita, S. O. Y. *et al. Nature Genet.* **36**, S17–S20 (2004).
18. *Nature Biotechnol.* **20**, 637 (2002).
19. Baker, J. L., Rotimi, C. N. & Shriner, D. *Sci. Rep.* **7**, 1572 (2017).

Metabolism

New-found brake calibrates insulin action in β-cells

Rohit N. Kulkarni

Insulin is produced by pancreatic β-cells. The identification of a regulator of insulin signalling in these cells cements the long-standing idea that this pathway has a key role in β-cell biology. **See p.326**

It is almost a century since insulin was first used to treat diabetes[1]. Since then, a great deal has been learnt about the complex metabolic pathways that are regulated by insulin and the related molecule insulin-like growth factor 1 (IGF1), acting through receptor proteins (reviewed in ref. 2). But it is less clear how the activity of these receptors is regulated in the cells that actually produce insulin, the pancreatic β-cells. Such knowledge is urgently needed because reduced β-cell function is a key contributor to diabetes. Deciphering the molecular pathways that regulate β-cells might therefore help to better manage, or even prevent, this disease. On page 326, Ansarullah *et al.*[3] identify a previously unknown regulator of β-cells, and outline the mechanism by which this protein can 'tailor' expression of the insulin receptor.

First, the authors analysed levels of messenger RNA in mouse cells, to identify genes that were highly expressed specifically in the embryonic pancreas. This revealed an abundantly expressed mRNA encoded by a gene on chromosome 3. The corresponding human gene is named oestrogen-induced gene (*EIG121*), and the mouse and human proteins are highly evolutionarily conserved.

Ansarullah *et al.* renamed the protein insulin inhibitory receptor (inceptor), because of its similarities to the insulin and IGF1 receptors. All three receptors span the cell membrane and have similar extracellular domains. But, unlike the insulin and IGF1 receptors, the short cytoplasmic tail of inceptor carries an amino acid sequence known to bind to the assembly polypeptide 2 (AP2) protein complex. AP2 is involved in a process called clathrin-mediated endocytosis, through which molecules and receptors at the cell surface are transported into the cell.

Then, the authors examined the function of inceptor by generating mice completely lacking the inceptor gene, and mice in which the gene could be deleted specifically in β-cells. The two models show generally similar traits.