doi 10.4436/jass.97004

JASs Invited Reviews
Journal of Anthropological Sciences
Vol. 97 (2019), pp. 91-106

Race in a genome: long read sequencing, ethnicity-specific reference genomes and the shifting horizon of race

Emma Kowal[1] & Bastien Llamas[2]

1) *Alfred Deakin Institute for Citizenship and Globalisation, Deakin University, 221 Burwood Hwy, Burwood Vic 3125, Australia*
e-mail: emma.kowal@deakin.edu.au

2) *Australian Centre for Ancient DNA, School of Biological Science, Environment Institute, The University of Adelaide, Adelaide SA 5005, Australia*
e-mail: bastien.llamas@adelaide.edu.au

Summary - *The sequencing of the human genome at the turn of the 21st century was hailed as revealing the overwhelming genetic similarity of human groups. Scholars of genomics have critiqued the subsequent persistence of race-based genetic science, but were reassured that the wide availability of gene sequencing would end the use of race as a proxy for genetic difference. Once an individual's whole gene sequence could be read, they hoped, their ethnoracial classification would become redundant. At the same time, genome science was recognising that the differences between human genomes went beyond the genome sequence to the structure of the genome itself. 'Structural variation' between genomes, including insertions, deletions, translocations, inversions, and copy number variations, mean that the 'universal' reference genome used for genome sequencing is not so universal. As conventional, 'short-read' sequencing wrongly assumes that all genomes have the same structure, significant genetic variation can be missed. This paper examines the twin phenomena that have been posed as a solution to the biases of short-read sequencing: 'long-read' sequencing and 'ethnicity-specific reference genomes'. Long-read sequencing is a method of generating a genome sequence that can be assembled* de novo *rather than relying on the reference genome. In recent years, a number of countries including China, Korea, and Denmark have used long-read sequencing and* de novo *assembly to develop 'national' reference genomes. Our analysis of one ethnicity-specific reference genome project, the Korean Reference Genome (KOREF), finds that it unduly emphasises the importance of population structural variation, framed in nationalist terms, and discounts the importance of individual structural variation. We argue that the intellectual labour required to make a Korean reference genome a coherent concept works to extend the horizon of race, prolonging the temporality of the 'meantime' in which race remains a seemingly valid concept in genomic science.*

Keywords - *Genomics, Sequencing,* De novo *assembly, Ethnicity, Race.*

Towards ethnicity-specific genomes

In December 2008, the *Korea Times* announced the completion of the "first Korean genome sequence" ('Koreans Complete Human Genome Map', 2008), published the following year in the journal *Genome Research* (Ahn *et al.*, 2009). Dr. Kim Seong-Jin of the Gachon University of Medicine & Science, Incheon, led the study and donated his blood for genome sequencing. Kim said he was inspired to embark on the project after reading one of James Watson's books and was "honored to reveal [his] DNA sequence for the development of medical research". The Times reported that his genome "reveals that genetic variations between humans could be greater than previously thought." Almost half the single nucleotide polymorphisms (single points of variation on the genome, known in the scientific literature by the acronym SNPs) identified through DNA sequencing could not be found in the other three full genomes existing at that time (reported as deriving from American co-discoverer of the DNA structure James Watson, American genomics pioneer Craig Venter, and leading Chinese geneticist Yang Huanming). This unique 'Korean' genetic information revealed in Kim's genome was expected to reap national benefits for personalised medicine and medical research. In place of the reliance on the 'universal' reference genome curated by the Genome Reference Consortium[1], Korean scientists and health consumers would provide their own baseline for genome diversity.

The sequencing milestone marked by the *Korea Times* in 2008 was part of a larger effort that led to the "ethnicity-specific"[2] Korean reference genome (KOREF) published in *Nature* eight years later that was produced using the technologies of long read sequencing and *de novo* assembly ('koreangenome.org'; Seo *et al.*, 2016). The authors argued that KOREF and other "national and ethnic" reference genomes would be "useful in improving the alignment of East-Asian personal genomes", making genome sequencing more relevant to people of East Asian background. As we will see, in the last few years other resource-rich countries have also produced reference genomes specific to their nation or ethnicity.

For those familiar with the long-running debates on the biological basis of ethnic and racial groups, the Korean Reference Genome might induce déjà vu. Wasn't genome sequencing supposed to be the end of race ('June 2000 White House Event', 2000; Angier, 2000)? How is it that "national and ethnic" biological differences have once again become prominent in genomics? This paper will seek to address these questions.

Our argument analyses some social and political implications of the application of current genome sequencing technologies. At this point, we advise readers less familiar with the technical details of genome sequencing to read the accompanying boxed sections on 'short-read' and 'long-read' DNA sequencing, structural variation and *de novo* assembly (see pages 94 and 96). We take this step as we want to communicate both to scientists working in the field and to social scientists who may be less familiar with the technical aspects.

From a technical and scientific point of view, the advent of the KOREF, and other ethnicity-specific reference genomes we mention below, is determined in part by the simultaneous developments of structural variation research and long-read sequencing technologies. However, these factors do not explain the rise of ethnicity-specific reference genomes alone. As we explore in this paper, it would have been entirely possible, and arguably more scientifically accurate, to circumvent population-specific reference genomes and skip straight to individual reference genomes or to methods that facilitate the collation of structural variations from individuals such as 'genome graphs' (Fig. 1A). Genome graphs are an emerging method of representing genomic data, whereby all sequence variations form "bubbles" flanked by stretches of DNA sequences that are conserved (do not vary) among humans. This graphic representation efficiently depicts sequence variation without the need for any single reference sequence. Both the variation between individuals, and within an individual genome (as paternally- and maternally-inherited alleles may differ from each other) are captured in a genome graph. As we will return to in the conclusion, genome graphs offer an alternative to ethnicity-specific reference genomes as a means to reduce the biases inherent to the linear human reference (Garrison *et al.*, 2018). Importantly, genome graphs enable diversity to be represented without automatically reinforcing notions of "national and ethnic" biological differences.

While we can't know for certain why ethnicity-specific reference genomes have so far received far more attention from genome scientists than genome graphs, we suspect that any explanation

[1] The Genome Reference Consortium (GRC) consists of the Wellcome Sanger Institute, The McDonnell Genome Institute at Washington University, The European Bioinformatics Institute, The National Center for Biotechnology Information, and The Zebrafish Model Organism Database.

[2] The terms 'ethnically-specific' and 'ethnically-relevant' have also been used in this emerging area of literature. We consider these other terms to be equivalent to 'ethnicity-specific' although changes in meaning may emerge over time.

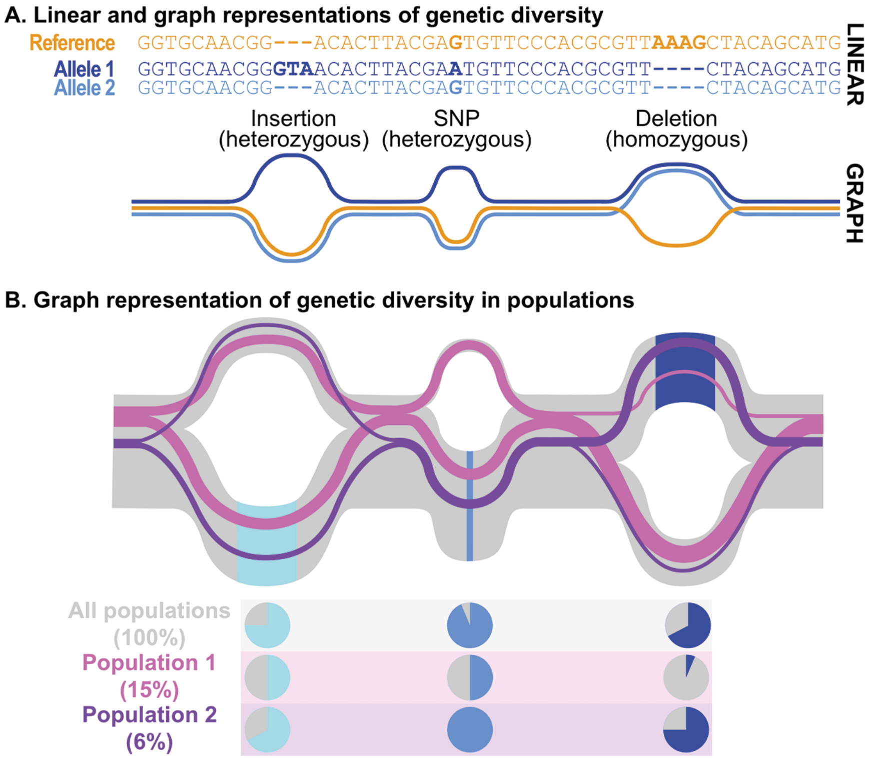
A. Linear and graph representations of genetic diversity
Reference GGTGCAACGG---ACACTTACGAGTGTTCCCACGCGTTAAAGCTACAGCATG
Allele 1 GGTGCAACGGGTAACACTTACGAATGTTCCCACGCGTT----CTACAGCATG
Allele 2 GGTGCAACGG---ACACTTACGAGTGTTCCCACGCGTT----CTACAGCATG
LINEAR
Insertion (heterozygous)
SNP (heterozygous)
Deletion (homozygous)
GRAPH
B. Graph representation of genetic diversity in populations
All populations (100%)
Population 1 (15%)
Population 2 (6%)

Fig. 1- Examples of graph representations of genetic data. A) A single allele is represented in a linear reference sequence (orange) despite potential variation in individual allelic sequences (dark and light blue). This penalises the discovery of non-reference variants. In genome graphs, all variants are represented in "bubbles" flanked by invariant "edges". B) Pink and purple graphs represent individual populations that amount to 15% and 6%, respectively, of the total population (grey graph). The allelic diversity (pie charts) is estimated for each population. The colour version of this figure is available at the JASs website.

needs to recognise the persistent allure of group-based biological differences for various political ends. In the last decade, social science scholars have examined the role of genomics in contests over national and ethnic identity (Sommer, 2010; Egorova, 2010; Nash, 2012). Ruha Benjamin's elaboration of the concept of 'genomic sovereignty' has been particularly influential (Benjamin, 2009; Schwartz-Marin & Mendez, 2012; de Vries & Pepper, 2012). She showed that in addition to the globalising and universalising forces of genetic research (Thacker, 2005), national and ethnic forces were still at play in 21st century genomics. Her focus was on Mexico and India, where national efforts in the early 2000s aimed to protect human genetic diversity—national 'biovalue' (Waldby & Mitchell, 2006)—from global exploitation. The goal for these countries was to build 'a lab of one's own' that maintained control over genetic information for the benefit of the nation. Tupasela has recently extended this analysis to examine populations as 'brands' in a highly competitive research and biotechnology market (Tupasela, 2017).

These interplays are examples of 'co-production', a concept that describes how social and technoscientific phenomena simultaneously shape each other (Jasanoff, 2004). Human biological differences are sites of particularly intense co-production as individual, group, national and global agendas vie with each other to record,

BOX 1: HOW TO SEQUENCE A GENOME

To sequence the human genome means to determine the order of the 4 letters that stand for the genetic information contained within the DNA: the nucleotide bases A (adenine), T (thymine), C (cytosine), and G (guanine). There are 6 billion nucleotide bases in the human genome, arranged along the opposite strands of the DNA double helix in 3 billion A-T and G-C base pairs (bp).The human genome draft sequence was made available to the research community in 2001 (International Human Genome Sequencing Consortium, 2001) after more than a decade of international efforts as part of the Human Genome Project, and for an estimated cost of 3 billion US dollars, roughly $1 per bp. A private venture conducted by Celera Genomics resulted in the simultaneous publication of another working draft of the human genome (Venter *et al.*, 2001), in only 3 years and for a tenth of the cost. For both projects, the methodology consisted in building Bacterial Artificial Chromosomes (BAC), which are large-insert DNA libraries that can be amplified by bacterial cloning, and performing relatively low-throughput sequencing using the Sanger method. Eighteen years later, technological advances have dramatically reduced sequencing time and cost. It is now possible to sequence a complete human genome at a depth of coverage of 30X (each position in the genome is read independently 30 times on average) in less than a week for roughly $1,000US.

The major innovation that made rapid and cheap DNA sequencing a reality is high-throughput short-read sequencing, also dubbed massively-parallel sequencing, next-generation sequencing, or second-generation sequencing, first described in the literature in 2005 (Margulies *et al.*, 2005). Several technologies are commercially available (van Dijk *et al.*, 2014), although the market is arguably dominated by Illumina, and some companies have already closed down. The methodological steps behind short-read sequencing are fairly simple: the DNA is fragmented into short molecules (typically <500 bp) and converted into a sequencing library compatible with the technological platform used. Millions to billions DNA library molecules are then sequenced in parallel to produce short sequencing outputs named ‘reads’. The length of a read is typically less than 500 nucleotides.

In the last few years, the technology of long-read sequencing—also referred to as single-molecule sequencing, or third-generation sequencing—has emerged as the new ‘gold standard’ of genome sequencing. The two technologies currently available for research are commercialised by Oxford Nanopore Technologies (Mikheyev & Tin, 2014) and Pacific Biosciences (Eid *et al.*, 2009). Similar to short-read sequencing, the DNA needs to be prepared into sequencing libraries, except that the fragment size is much larger, from several kilobases—kb—to several tens of kb.

Beyond the size of the DNA fragments that can be sequenced, the fundamental difference between short- and long-read sequencing technologies lies in the techniques to assemble the reads and reconstruct the genome sequence. Short-read datasets necessitate the use of a reference genome as a template (read mapping or reference-based methodology). The millions to billions of short ‘reads’ (fragments of <500bp) are put together—assembled—like a giant jigsaw puzzle. The frame they are assembled on to—the reference genome—may or may not be a perfect fit for all the short reads. Because of structural variation (see Box 2), some of the puzzle pieces will not fit the reference genome template and will be left out of the assembly.

Rather than using a reference genome, long reads are assembled *de novo*, in a reference-free fashion (*de novo* assembly methodology) (Chaisson *et al.*, 2015b). Instead of matching short-read puzzle pieces to a predefined frame (a reference genome)—a process that will leave many puzzle pieces out because they don’t fit the frame—much bigger puzzle pieces (longer reads) are pieced together to make their own frame from scratch (Chaisson *et al.*, 2015a; Rhoads & Au, 2015).

interpret, embrace or denounce signs of the differences between us (Reardon, 2002).

This paper describes the co-production of ethnicity-specific reference genomes through the combination of technoscientific advances and political and cultural agendas that favour the periodic re-articulation of group biological differences. We first briefly outline the recurrent debates within genetics and genomics about group-based biological differences—often understood through the concept of 'race'—since the mid-twentieth century. Ironically, although people on both sides of the debate have argued that the wide availability of genome sequencing would make 'race' obsolete, group-level genomic difference has been re-articulated through the concept of national reference genomes. Predictions of the 'end of race' (June 2000 White House Event, 2000) seem to have settled on a new near-future target: the easy production and availability of individual reference genomes. Through an analysis of published literature and websites related to structural variation, we show how arguments for ethnicity-specific reference genomes tend to emphasise the potential significance of *group* structural variation rather than *individual* structural variation. Our analysis unpacks the ethnicity-specific reference genome as a novel articulation of group-based differences, illustrating how the horizon of race continues to unfold nearly two decades after the sequencing of the human genome.

The end of race?

> "I have deep sympathy for the concern that genetic discoveries could be misused to justify racism. But as a geneticist I also know that it is simply no longer possible to ignore average genetic differences among 'races'." (Reich, 2018)

With these words, leading Harvard geneticist David Reich reignited a fierce and long-running debate about the significance of human genetic differences. His argument was framed as a reaction to what he called the 'orthodoxy' that considers biological differences between racial groups to be irrelevant and sees genetic research into biological differences as a dangerous "slippery slope" leading to pseudoscience, eugenics and the Holocaust. Notwithstanding his own concern about reinforcing racism, the bald facts of biological difference, he argued, could no longer be ignored. Indeed, a century or so of studying human populations has provided ample evidence for patterns of human genetic diversity and some degree of structure of the human gene pool.

In reply, Jonathan Kahn and 66 other natural and social scientists argued against Reich's defence of 'race' (Kahn *et al.*, 2018). Genetic differences exist between different groups of people, they countered, but this does not mean that races are natural biological categories. Genetic differences could be found between any two arbitrary groups of people: the example they use is supporters of different baseball teams. It is when these groups are assigned a biological significance as a race that the problem starts:

> "For centuries, race has been used as potent category to determine how differences between human beings should and should not matter. But science and the categories it constructs do not operate in a political vacuum. Population groupings become meaningful to scientists in large part because of their social and political salience — including, importantly, their power to produce and enforce hierarchies of race, sex, and class." (Kahn *et al.*, 2018)

These scholars see race as a political reality, not a biological one. Of all the possible biological differences that could be discerned between possible human groupings, only very few are considered meaningful. Those that become meaningful do so when they serve a social or political purpose. And although biological population groupings could be and arguably are delineated for positive purposes (e.g. to address inequalities), history has shown that regardless of the intention, the outcome is often detrimental to those biological groupings that are seen as deficient or inferior.

The recent exchange between Reich (Reich, 2018) and Kahn *et al.* (2018) was merely the

BOX 2: Genomic structural variation

A major advantage of long-read sequencing is the ability to accurately describe structural variation. At the time that the sequencing of the human genome was first announced, it would have been reasonable to assume that the most significant kind of genetic variation between humans was SNPs: for example, where one person has an A at a particular point in the genome, another person will have a C. It is estimated that 1 in every 300 bases will be different between two randomly picked human genomes (Feuk *et al.*, 2006a). Therefore, 'SNP chips' (that directly detect which base is present for millions of SNPs) and short-read sequencing seemed to be all that was needed to understand the critical aspects of human genetic variation.

However, the evolution of genomes also involves structural variation, whereby some sequences larger than 1 kb can be deleted (leading to 'missing' sections of the genome), duplicated (repeated sections), inserted (added sections), translocated (sections moved to other locations in the genome*)*, inverted (sections that are backward), and copy number variation—where a particular sequence is repeated in the genome and the number of copies varies between people (Feuk *et al.*, 2006a). In the last decade or so, human genome scientists have increasingly recognised that structural variation undermines the effectiveness of comparing SNPs between people, as two human genomes cannot necessarily be compared base for base.

Short reads will characterise SNPs with high accuracy given the low error rate of short-read sequencing technologies, but structural variants that are larger than the read length will be problematic. For example, a read that falls into a duplication will produce two identical puzzle pieces. Genome assembly using short-read technology and the human reference genome will only map one puzzle piece, as the reference genome has only one space for it (Fig. 2A). The other puzzle piece will be discarded in the assembly process. In this example, structural variation is ignored, even though it may be highly relevant (as explained in this article, many diseases have been associated with structural variation). In order to reliably assemble a genome from sequence data, it is therefore necessary to stitch together long reads *de novo* rather than using the reference genome as a frame (Chaisson *et al.*, 2015a). The increasing recognition of structural variation and parallel rise of long-read sequencing technology has raised questions about the human reference genome sequence, namely: how universal is it? When it is used as the reference sequence for short-read sequencing, how much individual variation is missed?

The human reference genome sequence (Li *et al.*, 2010), is a linear composite sequence assembled from multiple individual genome sequences. It is a key outcome of the Human Genome Project. The reference genome was derived from a number of people in Buffalo, New York state, as the scientist who created the DNA libraries for sequencing was based at the Roswell Park Cancer Institute in Buffalo (Kolata, 2013). The advertisement to recruit donors was published in The Buffalo News on 23 March 1997 and mentioned 20 volunteers (Watson *et al.*, 2017). However, the exact number of anonymous volunteers who provided DNA samples, and ultimately the number of samples that were used in the Human Genome Project, are unknown and not reported in the scientific literature. The only available information is that 5 to 10 samples were collected for each one used in the project (International Human Genome Sequencing Consortium, 2001).

Most of the draft genome sequence (91.6%) was constructed from eight large-insert genome-wide DNA libraries—made from 5 male blood samples, 2 sperm samples, and 1 immortalised cell line (International Human Genome Sequencing Consortium, 2001). One library in particular, RPCI-11 (a.k.a RP11), represented 74.3% of the first draft sequence (International Human Genome Sequencing Consortium, 2001). Evidence shows that RPCI-11 comes from an individual with admixed European and African ancestry (Reich *et al.*, 2009). The same RPCI-11 library represents 70.28% of the latest human genome assembly GRCh38, with another 82 libraries and other sources of sequencing data used for the rest of the assembly (Schneider *et al.*, 2017). Although chromosomes are represented linearly, GRCh38 also provides alternative sequence representations for some highly variable or complex regions of the genome.

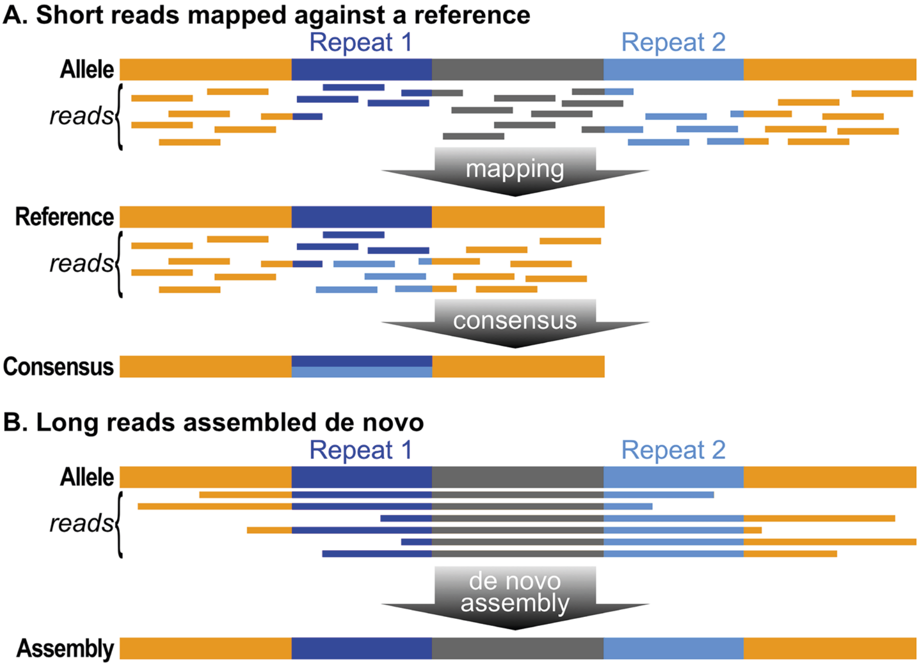

Fig. 2 - Short and long reads alignment across duplicated genomic regions. A) Where an individual has an allele with a duplication (light blue) of a region represented only once in the reference (dark blue), short reads that would normally align to either repeat will align to the sole reference sequence. The resulting consensus sequence is a chimera of the two repeat sequences. B) Long reads can overlap the duplicated region and the true allele is assembled de novo. The colour version of this figure is available at the JASs website.

latest in a long-running debate about the reality or fallacy of race that has proceeded continually with no resolution since the 1930s (Barkan, 1991; M'Charek, 2013). The announcement of the completion of the Human Genome Project in June 2000 is widely cited as a moment of false optimism about the end of race. Bill Clinton, the then President of the United States, famously stated that "one of the great truths to emerge from this triumphant expedition inside the human genome is that in genetic terms, all human beings, regardless of race, are more than 99.9 percent the same." ('June 2000 White House Event', 2000). Fleetingly, it seemed that the debate about race was over, ended by the overwhelming sameness of the species.

In the years since then, the use of race has not abated in popular discourse, and the arguments have continued (Fujimura *et al.*, 2014; Marks, 2017). In the pages of journals, newspapers, bulletin boards and social media, they go back and forth. Those in favour of race often argue it is a straightforward biological concept for breeding populations that progressive scientists deny for political reasons. Those arguing against the use of race commonly say it is an inaccurate and damaging term for the biological phenomenon of gradual population differences that is better understood as clinal genetic variation. Further, the latter argue that the lay, 'common sense' understanding of racial differences is based in sociopolitical forms, not biology, and is a manifestation of how people are unequally treated by others and by institutions.

A common feature of arguments for the utility of race, at least in the last decade, is that it is a temporary 'placeholder', particularly in clinical contexts (Kahn *et al.*, 2018). At such time when it is quick, simple and cheap to determine the relevant 'genotype' (literally, the genetic type) of an individual, then self-identified or externally ascribed race will be obsolete. For example, knowing a patient's particular variants of the *CYP2C9* and *VKORC1* genes could determine their optimal dosage of the blood-thinning drug warfarin without referring to their race.[3] As one commenter on the *American Scientist* website put it, "I agree that a full knowledge of genetics would effectively replace the medical need for race, but, until then, race remains a useful principle of research, diagnosis and treatment" (Dean, 2017). As it is not easy to determine the gene sequence of individuals, the argument goes, their race is a reasonable proxy for their genetic makeup. Race is therefore crucial for stratifying risk and diagnosing diseases that may otherwise be missed without the race prompt.

The use of race as a proxy for genetic differences 'in the meantime' relies primarily on variation data at single points in the genome (SNPs) that are similar *within* populations and different *between* populations. For example, at a single point in the genome, a cytosine (C) base might be found in 70% of 'race A' and only 40% of 'race B'. If this SNP is a marker for a gene variant associated with disease risk, race A is at greater risk than race B for the disease. This does not mean that *no* individuals in race B will get the disease, but that individuals in race A are more likely to get it.

When gene sequencing becomes widely available (either as part of clinical care or as a direct-to-consumer product)—so the 'meantime' argument goes—there will be no need to use the proxy of race. It will naturally be replaced by the far more accurate information contained in the individual genotype. In the last two decades, full genome sequencing has been the 'horizon' of race—the point beyond which race would hypothetically disappear. In the genomic near future, skin colour and appearance will finally become irrelevant when one's genome sequence is easily transported on a USB data storage device or generated at the point of care. This near future seemed to have arrived as advances in next-generation sequencing technology brought down the cost of sequencing a full genome to less than $1,000US, with further falls imminently expected (although data generation and storage still represent serious limitations).

However, as described in Boxes 1 and 2, a parallel stream of research into structural variation (henceforth SV) has undermined the potential of short-read sequencing to transform health and medicine on a mass scale. One implication of research on SV is to increase the estimation of differences between our genomes. The figure of 99.9% identical genome sequence shared by all humans that Clinton cited in 2000 was revised to 99.5% in 2006 (Feuk *et al.*, 2006a,b; Khaja *et al.*, 2006). This means that, on average, two human genomes are five times more different than previously thought, and SVs contributes much more to individual variation than SNPs.[4]

Research into SV is a hot topic with a rapidly expanding literature, but the significance of this 'extra' difference between genomes is not yet clear. SVs are certainly significant in some individuals and are linked to disorders including obesity (Wheeler *et al.*, 2013), diabetes (Cooper *et al.*, 2015), Alzheimer's disease (Corder *et al.*, 1993; Swaminathan *et al.*, 2012), autism (Sebat *et al.*, 2007; Kumar *et al.*, 2008) and schizophrenia (Stefansson *et al.*, 2008; The International Schizophrenia Consortium, 2008). In a 2011 paper, Li *et al.* argue that "structural variations are more specific to individuals than SNPs are. Thus, defining structural variations will be of considerable importance for future analyses of personal genomes,

[3] In practice, even once genetic information is known, many warfarin dosage calculators designed for the United States still retain ethnic and racial categories, see for example: http://www.warfarindosing.org/Source/InitialDose.aspx .

[4] Another emerging field relevant to this discussion, but out of scope for this paper, is epigenetics: biological processes that control the expression and regulation of genes without modification of the DNA sequence. Epigenetic differences are not differences in genomic sequence and it is not yet established whether they are inherited across generations in humans, however, it is quite possible that ethnicity-specific reference epigenomes will emerge in the future.

as structural variations may underlie phenotypic differences between individuals." (Li *et al.*, 2011)

SV also has implications for human groups. Analogous to patterns of variation in genome sequences, there is a great deal of individual variation in SV, but there is also population-level variation. Research into the significance of population-specific SV is in its infancy (Sudmant *et al.*, 2015). However, the lack of extant evidence has not curbed the rise of ethnicity-specific reference genomes.

Ethnicity-specific reference genomes

In the last ten years, long-read sequencing and *de novo* assemblies have been conducted on a small number of individual genomes, revealing substantial structural variation that would have been missed by short-read methods. An early effort to assemble two genomes—one 'Asian' and one 'African'—found ~5Mb of DNA sequences in each of them that is not present in the reference genome (Li *et al.*, 2010). Since then, other studies that have produced 'personal reference genomes' from single individuals have underlined the extent of individual variation in the human genome (Wang *et al.*, 2008; Pendleton *et al.*, 2015).

The advent of ethnicity-specific reference genomes took a little longer. In 2015, Danish researchers created a Danish reference genome from short-read sequencing and reference-based assembly of 10 trios from Copenhagen (Besenbacher *et al.*, 2015). This was the first use of the term 'national pan-genome'. The paper argued that a national pan-genome was justified on clinical and public health grounds:

> "A population-specific inventory of all detectable variation, a 'national pan-genome', has importance for clinical and public health genetics, for example, in facilitating imputation of rare variants in genome-wide association studies and low-pass sequencing studies and in addressing missing heritability due to an incomplete or inadequate human reference genome." (Besenbacher *et al.*, 2015)

This was followed in 2016 by the announcement of a Chinese Reference Genome produced from the first Chinese long-read sequencing/*de novo* assembly from one anonymous donor (Shi *et al.*, 2016). The justification in this paper also emphasised the drawbacks of the reference genome:

> "Previous studies reported that pervasive genetic differences exist across different ethnicity groups, especially on structural variants. For example, through reconstruction of the ancestral human genome, it was reported that megabases of DNA were lost in different human lineages and that large duplications were introgressed from one lineage to another. In addition, genomic elements that are absent from reference genomes may be present in personal genomes. For example, a study estimated that a complete human pan-genome would contain 19–40 Mb of novel sequence not present in the extant reference genome. These novel sequences that are not present in the reference genome may harbour functional genomic elements that are ethnicity-specific, and may affect gene regulations or transcriptional diversity." (Shi *et al.*, 2016)

There is a slippage in both of these quotes between the fact of individual variation and the presumed importance of 'national' or 'ethnic' reference genomes. SV is highly variable between individuals. If novel insertions, deletions, or other SVs that are not part of the human reference genome are found in a sample of a person who identifies with a certain ethnicity, this does not necessarily mean the SV is "ethnicity-specific", as Shi *et al.* imply. For example, half of the so-called 'novel' sequences found in the Chinese 'national' or 'ethnic' study have been recently found in two Swedish genomes, despite their very different ethnicity (Ameur *et al.*, 2018). The literature supporting ethnicity-specific reference genomes recreates 'race' by assuming that variation found in the few individuals they sequence stands in for the structural variation of the nation, ethnicity or racial group.

This slippage between individual and national/ethnic variation is most obvious in papers on Korean national reference genome projects. The key paper on the Korean Reference Genome (KOREF) is entitled "An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes." (Cho *et al.*, 2016) The paper describes the long-read sequencing and *de novo* assembly of the genome of a 'representative' Korean male. As the title makes clear, the argument of that paper parallels older arguments for race as a placeholder until such future time as widespread gene sequencing becomes available and individual genotype can be determined. A 2018 paper by the same group reports on long-read sequencing of 50 "healthy Korean individuals" to create the Korean National Standard Reference Variome (KoVariome). A passage from this paper contains the clearest scientific justification for studying population-specific structural variation:

> "Current efforts to resolve SVs reported several population-scale SVs and CNVs [copy number variations] from whole genome sequencing (WGS) data, and these analyses characterized **population-specific** traits such as amylase gene duplication in high-starch diet populations and revealed associations for **specific diseases** such as hemophilia A, hunter syndrome, autism, schizophrenia, and Crohn's disease with SVs. Nevertheless, SVs identified in healthy individuals also contain a substantial number of **individual- and population-specific** SVs with no disease association. Taken together, these results have demonstrated the importance of constructing **population-specific** SV and CNV profiles for the characterization of disease association and identifying diagnostic markers for precision medicine." (Kim *et al.* 2018, emphases added)

The three sentences of this paragraph illustrate the progression of the argument. The authors start by outlining existing literature linking SV with population-specific traits and individual disorders. The one example provided of population-specific SV traits concerns copy number variation for the amylase gene, the gene that produces a protein for digesting starch (Perry *et al.*, 2007). As this paper is repeatedly cited as evidence for the significance of population-specific SV it is worth examining the study in more detail. Having more copies of the amylase gene may produce more amylase and enhance the digestion of starch. The 2007 study cited by Cho *et al.* compared three populations that traditionally eat a diet high in starch (European-Americans, Japanese and Hadza in Tanzania) with four populations that traditionally eat a diet low in starch (three African groups and the Yakut in north-east Siberia). They found that those in high-starch diet populations had significantly more copies of the amylase gene than low-starch diet populations. They concluded that diet had produced evolutionary pressure that had increased the copy number of the amylase gene in some populations.

Importantly, the 2007 study included populations from a variety of continental ancestries in both the high-starch and low-starch groups. The researchers concluded that copy number variation varies in populations depending on the level of starch in their diet over a long period. This is an environmental difference common to any population that has been exposed to a particular nutritional environment, not an ethnic or national difference *per se*. As such, it is arguably not a strong evidence base from which to argue for the significance of ethnicity-specific reference genomes. Similar reasoning has been used to label sickle cell anaemia as a disease of African Americans, when it is more accurately seen as a disease related to having ancestors who lived in malaria-prone areas, as having one copy of the mutated hemoglobin-Beta gene does not cause sickle cell disease but confers resistance to malaria (Wailoo, 2001; Wailoo & Pemberton, 2006). For both the sickle cell trait and copy number for the amylase gene, an environmental determinant has been mischaracterised as a national, ethnic or racial cause.

The second sentence of the Kim et al quote goes on to argue that there is both individual- and population-specific variation in SV. The third sentence makes the concluding argument that this evidence shows that population-specific reference genomes are important for precision medicine. Crucially for our argument, the narrative transitions from individual-*and* population-specific variation in the first two sentences, to justifying the study of population-specific variation alone in the final sentence.

The earlier paper from this group (Cho *et al.*, 2016) illustrates how a commitment to ethnic-level variation can lead scientists to minimise individual variation. The authors describe the lengthy process of identifying Korean-or Asian-specific SV even though the patterns of variation of SV did not easily map onto ethnic groups. The authors note, for instance, that "YH_2.0 [a Chinese genome] and African genomes shared SVs abundantly, which cannot be explained by our assumption that similar ethnic genomes should have a higher genome structure similarity" (Cho *et al.*, 2016). Rather than abandoning their assumption, the authors use different methods to identify KOREF-specific variation, although they later note that "at the whole-genome variation level, intra-population variation is higher than the inter-population variation in terms of number of variants, supporting the notion that Homo sapiens is one population with no genomically significant subspecies" (Cho *et al.*, 2016). Despite saying this, the KOREF is based on the notion that "[t]he Korean population is regarded as a homogeneous ethnic group in East Asia" that has corresponding meaningful biological differences (Cho *et al.*, 2016).

Social scientists have examined the work that goes into 'genome geographies' that map biological difference onto specific territory and populations (Fujimura & Rajagopalan, 2011; Nash, 2012). The concept of race is a shorthand for this work of tethering space, biology and identity (Gannett, 2014). In the case of Korea, this tethering has been well documented and analysed by sociologist Gi-Wook Shin. Emerging in the late nineteenth century, the ideology of a single Korean race developed as a response to Japanese imperialism. At that time, the legendary founder of the first Korean kingdom, Dangun (Tan'gun) Wanggeom began to be viewed as the common biological ancestor of members of the Korean nation/race, a belief that continues to be dominant today (Shin, 2006; Kyung-Koo, 2007). In this context we can better interpret the words of the chairman of a biotech company cited by Sandra Soo-Jin Lee, the only social scientist thus far to write about Korean genome sequencing. The chairman of the company, that is closely involved with Korean genome projects, considers that the goal of the project is to "decode the biological definition of the word 'Korean'." Lee interprets this as a claim that "in genes resides an elusive essence of Korean identity that is critical to understanding and treating the Korean body." (Lee, 2010)

The representation of Korean national identity as a single biological category is also reflected on the website of the *Genome Asia 100k* project, a genome sequencing project led by commercial company Medgenome (genomeasia100k.com). While all the other 17 Asian countries represented on the 'collaborative' website include an infographic depicting some kind of racial mixture, Korea is described as a country of over 49 million '100%' Korean people, "with about 20,000 Chinese" included in parentheses. Korea is represented as uniquely racially homogenous among Asian nations.

A 2017 version of the Korean Reference Genome website that has since been removed provides perhaps the most detailed description of the biological notion of the Korean race. It begins:

> "Recent research on human diversity using genome information showed that the human race is classified into three super-groups, African, Caucasian and Asian, which is the result of long segregation in the human migration history." (koreagenome.kobic.re.kr)[5]

[5] This website is no longer accessible. The full text of the page we are quoting from is here, accessed 2017: "In earlier archaeology studies, Koreans were known to be the descendants of Altaic or proto-Altaic tribes (Lee *et al.*, 2008; Nelson, 1993). However recent findings based on mtDNA and Y-chromosome showed that current Koreans were originated from both southern and northern parts of East Asia (Jin *et al.,* 2003; Jin *et al.*, 2009; Karafet *et al.*, 2001; Kim *et al.*, 2000). In these DNA based studies, Koreans are known as an admixed population, and the most prevalent Y-chromosome and mtDNA haplogroups were O2b and D4a each (Hammer *et al.*, 2006; Jin *et al.*, 2009). The SRY465 mutation that defines the O2b Y-chromosomal haplogroup (proto-Koreans) is known that it had aroused from an ancestral O2* haplogroup belonging to a man who at least already had belonged to a specific "proto-Tungus-Korean" tribe (or who already had resided within Greater Manchuria) (Hammer *et al.*, 2006; Tymchuk, 2009). After the O2b divergence, another subclade, O2b1, likely had diverged after the

The text goes on to cite mitochondrial DNA (mtDNA) and Y-chromosome studies on Korean populations that found the most common Y-chromosome haplogroup (the 'proto-Korean' haplogroup) to be O2b (Hammer *et al.*, 2006; Jin *et al.*, 2009) and the most common mtDNA haplogroup to be D4 (Jin *et al.*, 2009). These were precisely the Y-chromosome and mtDNA haplogroups found in the single donor of the KOREF genome. The authors of the website conclude: "Thus KOREF can be considered as the direct descendant of proto-Koreans of the Y-chromosome and mtDNA founders."

Hinterberger & Porter (2015) describe how genomic (and viral) sovereignty requires that "variations […] be tethered to specific territories and their corresponding populations and authorities." In this now-discarded website text, the work of tethering Korean genetic variation (mtDNA and Y-chromosomes) to Korean territory and Korean identity is clearly visible. The KOREF donor is represented as a pure Korean, the "direct descendent" of "proto-Korean" ancestors. This tethering is pushed further on the Asian Genome website where the Korean population is depicted as ethnically pure, and Korea a country of ethnic and genetic homogeneity.

Discussion and conclusion

One might argue that the biologized nature of Korean national identity makes it an easy target for critiquing the 'return' of race in population-level genome projects. To be sure, we have not systematically analysed the justifying narratives of other ethnicity-specific reference genomes—including China, Sweden, Denmark, and Vietnam, among others (Besenbacher *et al.*, 2015; Thanh *et al.*, 2015; Shi *et al.*, 2016; Seo *et al.*, 2016; Ameur *et al.*, 2018)—although this would certainly be a worthy study. However, we believe it is highly likely that national and racial biases pervade the narratives and the science of other ethnicity-specific reference genomes in ways determined by the specific histories and politics of those places. The Korean case may be particularly illustrative, but it is surely not unique.

We do not wish to dismiss the potential utility of long-read sequencing and *de novo* assembly of genomes from humans of differing ethnic, cultural or geographic origins. It is also possible that our argument—that rationales for ethnicity-specific reference genome projects unduly emphasise the importance of population structural variation and discount the importance of individual structural variation—will be contradicted by scientific findings to come. Whatever the genomic future holds, scientists and social scientists must remain sensitive to the impact of social factors and interests in the creation and translation of genomic knowledge.

Our goal is not to separate biological 'facts' from social 'fictions', to use anthropologist of science Amade M'Charek's terms (M'Charek, 2013), but to understand how scientific facts about race are co-produced by labour of various kinds and serving various interests. From our perspective, the contradictory arguments in the Kim *et al.* paper for and against the significance of population-specific SV are interesting because they reflect the intellectual labour required to make a Korean reference genome, or any ethnicity-specific reference genome, a coherent concept. This is the labour of extending the horizon of race, prolonging the temporality of the 'meantime' in which race is still needed.

proto-Koreans formation about 1,640 ~ 7,960 years ago. KOREF's Y-chromosome had two proto-Korean markers (SRY465 and IMS-JST022454) that define the O2b haplogroup. Markers of the O3 haplogroup prevalent in China and the 47z mutation of the O2b1 prevalent in Japan were not detected. Annotated fifteen maternal mutations are known as mtDNA subhaplogroup D4a markers. The haplogroup, D4, was reported as the most prevalent haplogroup in Korea (Lee *et al.*, 2006; Umetsu *et al.*, 2005). Thus KOREF can be considered as the direct descendant of proto-Koreans of the Y-chromosome and mtDNA founders. At the same time, the autosomal linage drawn with NJ method indicated that the Korean donor can be regarded as the one who has the most common genetic traits within Koreans because he was clustered with other Korean samples between Japanese in Tokyo (JPT) and Han Chinese in Beijing (CHB) (Fig. 1). We think that our sequencing project will contribute to the overall human genetics including the understanding of human diversity especially in northern Asia."

Alongside concerns about how race persists, there is also the question of why. Following science and technology studies scholar Susan Leigh Star, it is important to ask *Cui bono*—who benefits—from particular representations of scientific facts? (Star, 1995) In the case we have described in this paper, the Korean state is one clear beneficiary of ethnicity-specific reference genomes. Through the Korean reference genome, science reinforces and strengthens Korean national identity. Other clear beneficiaries are the commercial sector offering tools for long-read sequencing. Pacific Biosciences, in particular, currently dominate this industry niche and would view ethnicity-specific reference genomes as a business opportunity. Accordingly, a 2017 article by Pacific Biosciences Senior Director for Human Biomedical Applications Luke Hickey argues that "[t]he idea of 'ethnic reference genome' appears to be more commonly discussed in the commercial sector" (Hickey, 2017).

Minority groups have long suffered negative consequences of race, a fact that has motivated the generations of scientists and social scientists who have critiqued the concept. Ethnicity-specific reference genomes illustrate how the 'durability' of race allows it to jump platforms and take on different forms rather than disappearing (Pollock, 2012). In years to come, we should expect epigenomic, proteomic and transcriptomic versions of race to emerge, and we should continue to ask difficult questions about the justification of these concepts and who benefits from them.

As we have described in this paper, 21st century genomic science found that genome variation was more extensive and more significant than many believed at the time the human genome was first sequenced. Nations like Korea have made use of this to promote a brand of national genomics that reinforces the idea of biological race and extends its horizon. Gene sequencing was once held up as the horizon of race, the point at while race would cease to exist. With ethnicity-specific reference genomes, the horizon of race has shifted to a future time when personalised reference genomes become standard practice.

It was not necessary or inevitable that the horizon of race be extended, maintaining a place for race. Things could be done otherwise. In concluding this paper by advocating for a better way to account for population-level differences, we are inspired by M'Charek's challenge "to denaturalize [race] without dematerializing it, and to simultaneously attend to materiality without fixing race" (M'Charek, 2013). We propose that genome graphs are one way to account for individual and population diversity without dematerializing or fixing race (Church *et al.*, 2015).

Ethnicity-specific reference genomes are always already racialized as they begin with the assumption that specific human populations are the most important and meaningful level at which to map variation. By contrast, genome graphs have the potential to avoid racialization by pooling individual diversity into a single human pan-genome. In advocating for genome graphs as an alternative tool for human genome research, we may be accused of simply creating another false horizon beyond which race will cease to exist. We do not wish to underestimate the limitations and dangers of this alternative method of representing variation.

First, the utility of a human genome graph, like any other form of recording and analysing variation, relies on the data it is based on. A genome graph that draws on data from a single population, for example, shares all the limitations of an ethnicity-specific reference genome. Second, even if a genome graph draws on a wide variety of human data and presents these as human variation (and not 'national' or 'ethnic' differences), there is always the potential for graphs to become racialized when they are used to measure differences between populations. We see this in Figure 1(B) where a genome graph is used to measure differences between 'Population 1' and 'Population 2'. When 'Population 1' and 'Population 2' are national or ethnic groups, a potential tool for representing pan-human variation reverts to a tool for race.

We hope that this paper prompts discussion about ethnicity-specific reference genomes and genome graphs in the scientific and social science communities. Our ultimate goal, however, is not to definitively solve the problem of race through genome graphs, or any other emerging tool, but to illustrate why and how we might seek alternative approaches to mapping human difference. Ethnicity-specific reference genomes reflect 'science-as-usual', a way of doing science that is likely to reproduce the power relations and inequalities embedded in society (Harding, 1986). If we approach human genome variation responsibly with the political and scientific history of race in mind, we will—hopefully—do things differently, producing more equitable and less harmful futures.

Acknowledgements

The title of this article is inspired by Kahn (2012). Emma Kowal's research is supported by an Australian Research Council Future Fellowship (FT160100093). Bastien Llamas's research is also supported by an Australian Research Council Future Fellowship (FT170100448) .

References

Ahn S.-M., Kim T.-H., Lee S. *et al.* 2009. The first Korean genome sequence and analysis: Full genome sequencing for a socio-ethnic group. *Genome Res.,* 19: 1622–1629.

Ameur A., Che H., Martin M. *et al.* 2018. *De Novo* Assembly of Two Swedish Genomes Reveals Missing Segments from the Human GRCh38 Reference and Improves Variant Calling of Population-Scale Sequencing Data. *Genes,* 9: 486.

Angier N. 2000. Do Races Differ? Not Really, Genes Show. *N. Y. Times*. August 22, 2000.

Barkan E. 1991. *The Retreat of Scientific Racism: Changing Concepts of Race in Britain and the United States between the World Wars*. Cambridge University Press.

Benjamin R. 2009. A Lab of Their Own: Genomic sovereignty as postcolonial science policy. *Policy Soc.,* 28: 341–355.

Besenbacher S., Liu S., Izarzugaza J.M.G. *et al.* 2015. Novel variation and *de novo* mutation rates in population-wide *de novo* assembled Danish trios. *Nat. Commun.,* 6: 5969.

Chaisson M.J.P., Huddleston J., Dennis M.Y. *et al.* 2015a. Resolving the complexity of the human genome using single-molecule sequencing. *Nature,* 517: 608–611.

Chaisson M.J.P., Wilson R.K. & Eichler E.E. 2015b. Genetic variation and the *de novo* assembly of human genomes. *Nat. Rev. Genet.,* 16: 627–640.

Cho Y.S., Kim H., Kim H.-M. *et al.* 2016. An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat. Commun.,* 7: 13637.

Church D., Schneider V., Steinberg K. *et al.* 2015. Extending reference assembly models. *Genome Biol.,* 16: 13.

Cooper N.J., Shtir C.J., Smyth D.J. *et al.* 2015. Detection and correction of artefacts in estimation of rare copy number variants and analysis of rare deletions in type 1 diabetes. *Hum. Mol. Genet.,* 24: 1774–1790.

Corder E.H., Saunders A.M., Strittmatter W.J. *et al.* 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science,* 261: 921–923.

Dean A. 2017. Is Race Real? (Comments). *Am. Sci. www.americanscientist.org/article/is-race-real*

van Dijk E.L., Auger H., Jaszczyszyn Y. *et al.* 2014. Ten years of next-generation sequencing technology. *Trends Genet.,* 30: 418–426.

Egorova Y. 2010. Castes of genes? Representing human genetic diversity in India. *Genomics Soc. Policy,* 6: 32-49.

Eid J., Fehr A., Gray J. *et al.* 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science,* 323: 133–138.

Feuk L., Carson A.R. & Scherer S.W. 2006a. Structural variation in the human genome. *Nat. Rev. Genet.,* 7: 85–97.

Feuk L., Marshall C.R., Wintle R.F. *et al.* 2006b. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.,* 15: R57–R66.

Fujimura J.H., Bolnick D.A., Rajagopalan R. *et al.* 2014. Clines Without Classes: How to Make Sense of Human Variation. *Sociol. Theory,* 32: 208–227.

Fujimura J.H. & Rajagopalan R. 2011. Different differences: The use of 'genetic ancestry' versus race in biomedical human genetic research. *Soc. Stud. Sci.,* 41: 5–30.

Gannett L. 2014. Biogeographical ancestry and race. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.*, 47 Part A: 17184.

Garrison E., Sirén J., Novak A.M., Hickey G., Eizenga J.M., Dawson E.T., Jones W., Garg S., Markello C., Lin M.F., Paten B., Durbin R. 2018 Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotech*., 36: 875–879.

Hammer M.F., Karafet T.M., Park H. *et al.* 2006. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J. Hum. Genet.,* 51: 47–58.

Harding S. 1986. *The Science Question in Feminism*. Cornell University Press, Ithaca, NY, USA.

Hickey L. 2017. Hunting Structural Variants: Population by Population. *Front Line Genomics Mag.,* 15: 43-45.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409: 860–921.

Jasanoff S. 2004. *The Co-Production of Science and the Social Order*. Routledge, New York, NY.

Jin H.-J., Tyler-Smith C. & Kim W. 2009. The Peopling of Korea Revealed by Analyses of Mitochondrial DNA and Y-Chromosomal Markers. *PLoS One,* 4: e4210.

June 2000 White House Event 2000. *Natl. Hum. Genome Res. Inst. NHGRI*.

Kahn J. 2012. *Race in a Bottle: The Story of BiDil and Racialized Medicine in a Post-Genomic Age*. Columbia University Press, New York.

Kahn J., Nelson A., Graves Jr. J.L. *et al.* 2018. How Not To Talk About Race And Genetics. *BuzzFeed*. March 30th 2018. www.buzzfeednews.com/article/bfopinion/race-genetics-david-reich

Khaja R., Zhang J., MacDonald J.R. *et al.* 2006. Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.,* 38: 1413–1418.

Kolata G. 2013. The Human Genome Project, Then and Now. *N. Y. Times*. April 15th.

Koreans Complete Human Genome Map, 2008. *koreatimes*. 12th April 2008. www.koreatimes.co.kr/www/news/tech/2008/12/133_35578.html

Kumar R.A., KaraMohamed S., Sudi J. *et al.* 2008. Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.,* 17: 628–638.

Kyung-Koo H. 2007. The Archaeology of the Ethnically Homogeneous Nation-State and Multiculturalism in Korea. *Korea J.,* 47: 8–31.

Lee S.S.-J. 2010. The Asian Genome: Racing in an Age of Pharmacogenomics. In *Sleeboom-Faulkner* M. (ed): Frameworks of Choice: Predictive & Genetic Testing in Asia, pp. 211-222. Amsterdam University Press.

Li R., Li Y., Zheng H. *et al.* 2010. Building the sequence map of the human pan-genome. *Nat. Biotechnol.,* 28: 57–63.

Li Y., Zheng H., Luo R. *et al.* 2011. Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome *de novo* assembly. *Nat. Biotechnol.,* 29: 723–730.

Margulies M., Egholm M., Altman W.E. *et al.* 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437: 376–80.

Marks J. 2017. *Is Science Racist?* Polity Press, Cambridge, UK; Malden, MA, USA.

M'Charek A. 2013. Beyond Fact or Fiction: On the Materiality of Race in Practice. *Cult. Anthropol.,* 28: 420–442.

Mikheyev A.S. & Tin M.M.Y. 2014. A first look at the Oxford Nanopore MinION sequencer. *Mol. Ecol. Resour.,* 14: 1097–1102.

Nash C. 2012. Genome geographies: mapping national ancestry and diversity in human population genetics. *Trans. Inst. Br. Geogr.,* 38: 193–206.

Pendleton M., Sebra R., Pang A.W.C. *et al.* 2015. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods,* 12: 780–786.

Perry G.H., Dominy N.J., Claw K.G. *et al.* 2007. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.,* 39: 1256–1260.

Pollock A. 2012. *Medicating Race: Heart Disease and Durable Preoccupations with Difference*. Duke University Press, Durham and London.

Reardon J.E. 2002. *Race to the Finish: Identity and Governance in an Age of Genetics*. Cornell University.

Reich D. 2018. How Genetics Is Changing Our Understanding of 'Race'. *N. Y. Times*.DAY???

Reich D., Nalls M.A., Kao W.H.L. *et al.* 2009. Reduced Neutrophil Count in People of African Descent Is Due To a Regulatory Variant in the

Duffy Antigen Receptor for Chemokines Gene. *PLoS Genet.,* 5: e1000360.

Rhoads A. & Au K.F. 2015. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics,* 13: 278–289.

Schneider V.A., Graves-Lindsay T., Howe K. *et al.* 2017. Evaluation of GRCh38 and *de novo* haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.,* 27: 849–864.

Schwartz-Marin E. & Mendez A.A. 2012. The Law of Genomic Sovereignty and the Protection of Mexican Genetic Patrimony. *Med. Law,* 31: 283–294.

Sebat J., Lakshmi B., Malhotra D. *et al.* 2007. Strong Association of *De Novo* Copy Number Mutations with Autism. *Science,* 316: 445–449.

Seo J.-S., Rhie A., Kim J. *et al.* 2016. *De novo* assembly and phasing of a Korean human genome. *Nature,* 538: 243–247.

Shi L., Guo Y., Dong C. *et al.* 2016. Long-read sequencing and *de novo* assembly of a Chinese genome. *Nat. Commun.,* 7: 12065.

Shin G.-W. 2006. *Ethnic Nationalism in Korea: Genealogy, Politics, and Legacy*. Stanford University Press, Stanford, CA.

Sommer M. 2010. DNA and cultures of remembrance: Anthropological genetics, biohistories and biosocialities. *BioSocieties,* 5: 366–390.

Star S.L. 1995. *Ecologies of knowledge : work and politics in science and technology*. State University of New York Press, Albany.

Stefansson H., Rujescu D., Cichon S. *et al.* 2008. Large recurrent microdeletions associated with schizophrenia. *Nature,* 455: 232–236.

Sudmant P.H., Mallick S., Nelson B.J. *et al.* 2015. Global diversity, population stratification, and selection of human copy-number variation. *Science,* 349: aab3761.

Swaminathan S., Shen L., Kim S. *et al.* 2012. Analysis of Copy Number Variation in Alzheimer's Disease: the NIA-LOAD/NCRAD Family Study. *Curr. Alzheimer Res.,* 9: 801–814.

Thacker E. 2005. *The Global Genome: Biotechnology, Politics, and Culture*. MIT Press, Cambridge, MA.

Thanh N.D., Trang P.T.M., Hai D.T. *et al.* 2015. Building Population-Specific Reference Genomes: A Case Study of Vietnamese Reference Genome. *Seventh International Conference on Knowledge and Systems Engineering (KSE)*, pp. 97–102.

The International Schizophrenia Consortium. 2008. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature,* 455: 237–241.

Tupasela A. 2017. Populations as brands in medical research: placing genes on the global genetic atlas. *BioSocieties,* 12: 47–65.

Venter J.C., Adams M.D., Myers E.W. *et al.* 2001. The Sequence of the Human Genome. *Science,* 291: 1304–1351.

de Vries J. & Pepper M. 2012. Genomic sovereignty and the African promise: mining the African genome for the benefit of Africa. *J. Med. Ethics,* 38: 474–478.

Wailoo K. 2001. *Dying in the City of the Blues: Sickle Cell Anemia and the Politics of Race and Health*. The University of North Carolina Press, Chapel Hill and London.

Wailoo K. & Pemberton S. 2006. *The Troubled Dream of Genetic Medicine: Ethnicity and Innovation in Tay-Sachs, Cystic Fibrosis, and Sickle Cell Disease*. John Hopkins University Press, Baltimore, Maryland, USA.

Waldby C. & Mitchell R. 2006. *Tissue Economies: Blood, Organs, and Cell Lines in Late Capitalism*. Duke University Press, Durham, NC.

Wang J., Wang W., Li R. *et al.* 2008. The diploid genome sequence of an Asian individual. *Nature,* 456: 60–65.

Watson J.D., Berry A. & Davies K. 2017. *DNA: The Story of the Genetic Revolution*. Knopf, New York.

Wheeler E., Huang N., Bochukova E. *et al.* 2013. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat. Genet.,* 45: 513–517.

Associate Editor, Alan Goodman