Variations on a Chip:
Technologies of Difference in Human Genetics Research

Ramya M. Rajagopalan[1]
Joan H. Fujimura[2]

Abstract

In this article we examine the history of the production of microarray technologies and their role in constructing and operationalizing views of human genetic difference in contemporary genomics. Rather than the "turn to difference" emerging as a post-Human Genome Project (HGP) phenomenon, interest in individual and group differences was a central, motivating concept in human genetics throughout the 20th century. This interest was entwined with efforts to develop polymorphic "genetic markers" for studying human traits and diseases. We trace the technological, methodological and conceptual strategies in the late twentieth century that established single nucleotide polymorphisms (SNPs) as key focal points for locating difference in the genome. By embedding SNPs in microarrays, researchers created a technology that they used to catalog and assess human genetic variation. In the process of making genetic markers and array-based technologies to track variation, scientists also made commitments to ways of describing, cataloging and "knowing" human genetic differences that refracted difference through a continental geographic lens. We show how difference came to matter in both senses of the term: difference was made salient to, and inscribed on, genetic matter(s), as a result of the decisions, assessments and choices of collaborative and hybrid research collectives in medical genomics research.

[1] Institute for Practical Ethics, University of California, San Diego, 9500 Gilman Drive, San Diego, CA, 92093 rmrajagopalan@ucsd.edu
[2] Department of Sociology and Holtz Center for Science and Technology Studies, University of Wisconsin-Madison, 1180 Observatory Drive, Madison, WI, 53706

Keywords: sociotechnical practice, infrastructure, biomedical genomics, human genetic variation, single nucleotide polymorphisms (SNPs), population, DNA microarray, SNP chip, haplotype map, public-private research partnerships, genomics consortia

Introduction

In the early 2000s, the first large-scale data-generating genomics initiatives driven by the Human Genome Project (HGP) were drawing to a close. The data that the HGP and related projects released into the public domain, and the technologies whose development they had fostered through the 1990s, initiated significant organizational, scientific, and technical transformations in genetic studies of disease, shifting the scale of analysis from gene to genome. Under the banner of the HGP, researchers had envisioned, anticipated, and actively planned for a future in which they could search genome-wide for genetic factors involved in the chronic diseases that they had labeled the most significant public health threats. Among the new tools enabling genome-wide studies in the new millennium was the microarray (abbreviated "array"), a miniaturized silicon wafer-shaped device about the size of a thumbnail. Because microarrays drew design principles from and resembled integrated circuits or microchips used in the semiconductor industry, they were colloquially referred to as "chips" (Guo et al., 1994).

Microarrays significantly altered the landscape of medical genetics research in the mid to late 2000s, buttressing high-throughput efforts to identify and characterize genomic differences known as single nucleotide polymorphisms (SNPs). Researchers designated SNPs as those points in the genome where humans differed in their DNA sequence by a single nucleotide. For example, where some individuals might have an "A," others might have a "G." Microarrays designed to assay these SNP differences in a highly parallel fashion (nicknamed "SNP chips." About the size of a penny, they initially contained thousands (and later, millions) of "spots" arrayed in an ordered fashion. Each spot contained thousands of copies of a unique DNA sequence 20 to 25 nucleotides long,

known as a "probe." When DNA collected from a human donor was applied to the chip, each probe would selectively bind only those bits of the donor DNA that were perfectly complementary in sequence. This binding reaction would generate a positive optical signal at certain wavelengths of light, which could be measured and read by a detection instrument. A single base-pair difference between the individual's DNA and a given probe would prevent binding, resulting in a negative optical readout from that probe. Each SNP was assayed by a set of partially overlapping probes that could distinguish single nucleotide differences. By collating information gleaned through the combination of positive and negative readouts across the probe set, researchers could determine the donor DNA's nucleotide identity at a given SNP locus. SNP microarrays generated data like this across thousands of SNP loci simultaneously. When analyzed on a computer, a single microarray experiment yielded a profile of a donor's DNA, detailing the nucleotide identity (or "variant") present at each of the many hundreds or thousands of SNP loci assayed by the chip.[3]

Researchers in public, academic and corporate labs worked to improve chip synthesis, miniaturization, and manufacturing chemistries over time. They revised the statistical methods that guided their choices of which SNP loci in the genome to examine, developing approaches for accommodating increasing numbers of SNP assays on a single chip. These choices mattered because, according to researchers, the precise representation of SNPs on any given chip affected its utility for the study of diseases across different human "populations" (Pe'er et al. 2006). By the middle of the 2000s, some researchers had concluded that the first wave of generic SNP chip designs were inadequate for querying genetic variation across all human genomes and that new chip designs needed to represent more and more human genetic diversity to facilitate genomic studies of disease. By the end of the 2000s, these articulated needs had spurred a new wave of chips designed specifically for different human groups, which

[3] In laboratory vernacular, this process of genotyping SNP markers in the genome was known as "querying" or "(base)calling the SNP."

became commonplace across the human genomics research landscape.

This article examines the theoretical frames, scientific work practices, institutional logics, and financial incentives that fashioned SNP differences and embedded them in microarray technologies that both produced and analyzed genetic difference across human groups. A focus on SNPs and group-level differences have become defining features of contemporary medical genomics in the United States. We first trace the technological, methodological, and conceptual developments in the late twentieth century that established SNPs as key focal points for tracking and locating human difference in the genome. Though genomics researchers involved in the Human Genome Project settled on SNPs in the 1990s as their genetic markers of choice for studies of the human genome, and later designated chips as the preferred tool with which to measure them, these outcomes were not inevitable. They were the product of a series of research priorities and choices, in addition to theoretical and empirical considerations, that made certain kinds of genetic markers more valuable or informative to researchers than others. Just as important were the contingent decisions, assessments, and forecasts about genomics research needs made in the 1990s within a set of sociotechnical conditions shaped partially by the available technologies, partially by desires for and speculations about future technologies, and partially by changing institutional and work structures that researchers and organizations initiated within the field of human genomics.

SNP chips also drew in particular ways on social assumptions about human difference built into theoretical frames used in their design. We examine how these articulations of human difference became embedded in the SNP chip. By analyzing how SNP chips became instruments for measuring genetic difference, we show how their design, production, and use shaped (and were shaped by) ideas about the boundaries and limits of population groupings, ideas that often rendered distinctions among human groups along continental lines.

Our analysis relies on both ethnographic and historical approaches. We conducted fieldwork from 2007 to 2013 in five leading research institutes and genomics laboratories in the United States.[4] We observed genome scientists at work in laboratories and at meetings and conferences, conducted interviews and oral histories with researchers, and analyzed their published and unpublished scientific papers. We also draw on materials within the newly established NHGRI Genome History Archive, including consortia project draft and official reports, published literature, administrative documents, and meeting minutes, as well as emails, memos, handwritten notes, and personal correspondence shared among lead researchers, NHGRI administrators, consortia project coordinators and funders, and industry partners. Some of the laboratories and institutes where we conducted fieldwork were involved in the big genomics consortia projects initiated or supported by the National Human Genome Research Institute (NHGRI) and its predecessor, the National Center for Human Genome Research, including the Human Genome Project, The SNP Consortium, the International Haplotype Mapping Project, and the 1000 Genomes Project.

Writing histories of the contemporary period can offer unique opportunities for interrogating archival and ethnographic data through each other, and so more richly capture the complexity of institutional decisions and scientific choices. But it also has the potential to amplify contradictions, presenting unique challenges for verifying and authenticating conflicting source material. Writing an ethnographic history of the contemporary period means that many of the research subjects ("respondents") are still alive, still working professionally and building careers. Many of the governmental and academic institutions, private companies, and projects are also still "alive." While this offers opportunities to query respondents at different stages of their work, and so get multiple perspectives on a given event or project, it also magnifies the possibility of

[4] Due to human subjects confidentiality agreements with our respondents, in some instances our use of ethnographic material has necessitated the omission of respondent names.

contradictions or inconsistencies. In addition, since genomics has matured from a nascent field largely in tandem with the rise of the digital era, archival documents pertaining to more recent history of the field include non-traditional materials, such as email exchanges, web pages, and PowerPoint slides. Though these have analogs among more traditional historical materials, such as handwritten or typed correspondence and meeting notes, they do demand novel approaches to thinking about what constitutes historical data, how to analyze it, and the stability and verifiability of different sources and different accounts of the same events.

The challenge for the researcher is to develop techniques to resolve any contradictions. Triangulating across data sources, or seeking interviews with actors themselves to add to the historical record, can help corroborate previous statements and resolve ambiguities. But it can also introduce problems of what sociologists call "self-report," where respondents may seek to change the historical record. This does not assume ill will; for example, people often remember the past differently in the present moment for various reasons. Sociologists have methods for dealing with this. Nevertheless, post-hoc interviews can usefully reveal actors' stated motivations and intentions before and even after events in question, which cannot be directly interrogated through historical records alone. Post-hoc interviews combined with ethnographic fieldwork data can also inform the analysis of historical materials; they can, for example, help us discern the intended versus actual trajectories of technologies researchers were developing or using, and also help identify key decision points in that process that might not be obvious from the historical record alone. We specifically use the term ethnographic data here to emphasize that as sociologists we combine fieldwork observations and interviews, which gives more multifaceted perspectives on technical work practices, the chronology of events in scientific research, and the continuities and discontinuities between what scientists say and do than interviews on their own. However, care must be taken not to telescope presentist frames of reference to the past. Many of these challenges extend across historical studies of the recent sciences (Doel and Söderqvist

2006).

Genetic Markers and Genetic Maps

Early 20th century genetics focused on the question of how variations in phenotypic traits could be inherited and transmitted across generations. As a way to answer this question, generating methods for mapping the physical basis of inherited characteristics was from the outset a defining preoccupation of modern genetics, characterized by practices of interpretation, conceptual framing, inscription, and representation.[5] Genetic markers were fashioned as core elements of mapping efforts, beginning with the work of Thomas Hunt Morgan, Alfred Sturtevant, and colleagues at Columbia University, who developed trait markers as a means to construct genetic maps of chromosomes linking phenotypic difference to genotypic difference in fruit flies. The first major class of markers applied to human genome mapping, known as "restriction fragment length polymorphisms" (RFLPs), were developed in the 1970s with the advent of recombinant DNA technologies. Restriction enzymes generated unique patterns of DNA fragments by cutting DNA in sequence-specific locations across the genome. Researchers used these cuts and fragment patterns to "mark" chromosomal locations where individuals differed in genome sequence. Groups of markers were assigned locations relative to each other along the chromosomes, and in this way a picture of human sequence differences along chromosomes was produced. In the process, sequence differences were transmuted into genetic markers, each with its own genomic address.

Geneticists David Botstein, Raymond White, Mark Skolnick, and Ronald Davis (1980) proposed the first strategy for generating a map of the human genome, using the relative chromosomal locations of RFLP differences. They argued that a genome map could help researchers identify genomic loci involved in inherited diseases. If

[5] For detailed historical analyses of 20th century genetic mapping efforts across species, see the edited volumes of Gaudillière and Rheinberger (2004).

researchers could show that a particular marker and a particular disease phenotype were co-inherited by studying an affected family, the genomic location of the marker could act as a lamppost illuminating nearby gene(s) involved in the disease. Botstein and colleagues commented that, to date, "No method of systematically mapping human genes has been devised, largely because of the paucity of highly polymorphic marker loci." "Polymorphic" referred to the multiple different forms a marker could take (also known as variants, or alleles). An individual genome, with two copies of each chromosome, could exhibit at most two of these many possible forms. The variable character of RFLPs rendered them highly valuable to geneticists, because it offered a means to access and assay genetic variation. Thus, from the earliest human genome maps, researchers actively sought marker loci that exhibited extensive variability across individuals. Further, they framed the search for markers of human genetic variation as a necessary pre-requisite for gene finding through family linkage studies in the genome era, a problem akin to locating a "needle in a haystack" (Collins, 1991).

By mapping RFLPs in families affected by highly hereditary diseases in the late 1980s and early 1990s, geneticists zoomed in on the protein-coding genes responsible for diseases whose causes could be traced to sequence mutations in a single gene, including Huntington's disease, cystic fibrosis, and Duchenne muscular dystrophy. Nevertheless, geneticists soon found many gaps in the RFLP maps, because RFLPs were sparse in some regions of the genome. Vast tracts of the human genome remained inaccessible to genetic scanning, signalling an early disciplinary anxiety about the number of markers that genetic maps needed to contain in order to be "comprehensive," that is, to represent enough of the genome sequence to help researchers identify disease-associated loci with statistical confidence. In discussions in the literature throughout the 1980s and 1990s, estimates of the number of polymorphic markers required for disease gene finding continued to increase.[6]

[6] See US Office of Technology Assessment (1988) and Kruglyak (2008)

These discussions expanded under the aegis of the Human Genome Project (HGP), officially commenced in 1990 under the leadership of the NIH's National Center for Human Genome Research (which became the National Human Genome Research Institute, or NHGRI, in 1997). The HGP aimed to generate the definitive maps for understanding genetic diseases. One of the HGP's early ambitions was to pinpoint new classes of polymorphic genetic markers that were more numerous across the genome and more easily assayed than RFLPs, to produce genetic maps that covered more parts of the genome (US OTA 1988). The HGP's genetic maps were initially based on sequence tagged sites (STSs), unique sequences with known locations in the genome between 200 to 500 nucleotides in length that were not always polymorphic but were easily assayed by the new polymerase chain reaction (PCR) technique (Collins and Galas 1993). These included a class of markers known as short tandem repeat polymorphisms (STRs, or microsatellites), first identified in 1989 (Smeet et al., 1989) and characterized by James Weber's medical genetics lab at the Marshfield Clinic in Wisconsin (Weber and May 1989, Weber 1990). However, like RFLPs, microsatellites did not occur frequently enough along the chromosomes to enable high resolution mapping in the vicinity of individual genes (Weissenbach 1993, Weiss 1998).

It had been known since the 1960s that single base substitutions in protein-coding genes was sufficient to trigger disease phenotypes associated with some of the known hereditary disorders (McCusick 1992). These were sites in the genome where the nucleotide sequence (A, T, G, or C) varied across individuals. In the expanding vernacular of genetic mapping and genetic markers, single base substitutions became more commonly known as "single nucleotide polymorphisms" (SNPs). SNPs were considered less informative than RFLPs and microsatellites because they exhibited less variability. Whereas a RFLP or microsatellite could vary widely in length and sequence, anywhere between 50 to 100 concatenated occurrences of a repeating sequence motif, most known SNPs appeared to be bi-allelic (occurring only in two forms, such as A and G). Despite exhibiting lower sequence diversity across individuals, researchers increasingly

began to favor SNPs in genetic studies of disease, partly because they appeared to be far more abundant in the genome. The most widely-accepted estimate pegged them as occurring every 1000 nucleotides or so (Kruglyak 1997). Perhaps more importantly than their estimated density in the genome, several labs began to demonstrate that SNP analysis could be automated much more efficiently than RFLP or STR analysis, by adapting existing commercial technologies like PCR and high-throughput DNA sequencing machines (Nickerson et al. 1990; Nikiforov et al. 1994; Kwok et al. 1994; Delahunty et al. 1996). SNPs required less time-intensive manual labor to analyze than RFLPs and microsatellites and were more easily aligned with emerging technologies of automation. Coupled with their preponderance in the genome, this rendered SNPs increasingly popular objects of study in human genetic variation research. Below, we show how SNPs became conventional tools for marking genetic difference. The kinds of differences researchers claimed they represented and revealed emerged through the technologies in which they embedded SNPs, like SNP arrays.

DNA Microarrays and Catalogs of Human Genetic Difference

As genetic markers were being refined into research tools in the late 1980s and early 1990s, biotech and academic labs began searching for ways to speed up and automate DNA analysis. Traditional approaches, like the Southern blotting technique used to analyze RFLPs, were slow, requiring the immobilization of genomic DNA on a membrane and then adding a radioactively labeled probe of a known sequence to detect any complementary sequences in the DNA of the individual whose genome was being studied. This could take days or weeks. Instead, an alternative approach sought to invert this process, immobilizing many probes on a small silicon wafer (called a microarray), testing for their presence simultaneously in an individual's genome to see which of the many possible probes would bind to complementary DNA, then stripping that individual's genomic DNA away and repeating with another individual's DNA. Many researchers helped develop these ideas and techniques through the early 1990s, but

their incorporation into a portable, scalable technology has been credited to two teams. Stephen Fodor and his colleagues, working at the Affymax Research Institute in Palo Alto in the early 1990s, developed a way to use light-directed, spatially-addressed chemical synthesis to "print" peptides in an ordered fashion on glass slides, which they called "arrays" (Fodor et al. 1991). Shortly after, Mark Schena, Pat Brown, and colleagues at Stanford University developed arrays to look at comparative gene expression, with new techniques to accelerate the printing process using robotics and pre-synthesized DNA. They visualized gene expression differences by using differentially labeled fluorescent dyes (Schena et al. 1995). In 1993, Fodor co-founded a company, Affymetrix, to commercialize arrays for biological research.

The concept of patterning and printing amino acid and oligonucleotide sequences onto microarrays built on technical advances in the growing microelectronics and semi-conductor industry centered in Silicon Valley, including the use of photolithography for printing digital circuits. These inspired the possibility of "high-throughput" DNA analysis, whereby hundreds or even thousands of DNA sequences could be assayed in parallel. This earned them the nickname "lab on a chip." Although DNA microarrays saw their first applications in comparative studies of gene expression (Shostak 2005; Rogers and Cambrosio 2007), Brown's original intention and inspiration for working with microarrays grew out of a desire to better understand sequence variation and complex traits. When his team originally began developing the microarray technology, "It was to enable a new method for relating sequence differences in genes to complex traits in people… heritable differences in sequences".[7] This question of sequence differences, and the potential for microarray technology to address it, spurred the research aspirations of human genetic variation researchers in the late 1990s.

[7] Brown, P. 2003. "Why we developed the microarray." DNA Interactive video interview 15036, DNA Learning Center at Cold Spring Harbor Laboratory. https://www.dnalc.org/view/15036-Why-we-developed-the-microarray-Patrick-Brown.html.

Against the backdrop of the HGP, by the mid-1990s some prominent geneticists began targeting SNPs as key to characterizing patterns of genetic difference among individuals. Calls for ever-more comprehensive marker maps of the human genome were spurred by shifts in the diseases of interest to geneticists, which began to include complex conditions more widespread in the general population (Lander and Botstein 1986). These were chronic age-related diseases like heart disease, cancers, and neurological conditions, quickly becoming the major health concerns in wealthy industrialized nations. Although geneticists acknowledged that environmental factors might play a much larger role than genetics in these diseases, they nevertheless assumed that DNA sequence differences between affected and unaffected individuals would prove medically relevant.

In 1996, statistical geneticist and a leader within the HGP Eric Lander (1996) and statistical geneticist Neil Risch and epidemiologist Kathleen Merikangas (1996), formalized a genetic strategy for the study of complex diseases. They argued that the future of genetics lay in "genetic association studies." These would make use of large sets of polymorphic markers and large numbers of affected but unrelated individuals. The aim was to pinpoint small but possibly additive genetic contributions to complex disease from many loci across the genome (as opposed to Mendelian studies of disease, which isolated single genes responsible for highly hereditary conditions). Their proposal became the basis for a new class of genetic studies at the turn of the millennium, known as genome-wide association studies (GWAS). The primary obstacles, they argued, were the lack of a sufficiently large number of polymorphic genetic markers with which to conduct such studies and the high throughput technologies to assay them.

The concept of GWAS gave traction to "the idea of being able to spread markers right across the genome, and the power of doing so at an ever-higher density."[8] To that end,

[8] NHGRI's Oral History Collection: Interview with David Bentley, 2017, available at https://www.genome.gov/27552689/all-videos/

even as the HGP continued to focus on developing a single reference genome sequence, NIH's National Center for Human Genome Research began to consider more seriously the scientific resources that would need to be in place to enable GWAS. They established funding programs intended to accelerate the HGP by supporting innovations in large-scale DNA sequencing and analysis. From 1995 to 1997, technology development at Affymetrix benefitted from these programs. Among many other DNA and RNA analysis applications, Affymetrix was pursuing SNP microarray technology development in anticipation of GWAS. Many labs were on the hunt for SNPs, and at the time, the rate at which researchers were identifying SNPs far outpaced the ability of any technology to genotype[9] them for disease studies (Kruglyak and Nickerson 2001). Through the late 1990s Affymetrix experimented with methods for chip construction and genotyping and assay protocols, as well as building the computational software for managing and analyzing the sequence data, with a view to eventually commercializing a chip technology that could analyze ever-increasing numbers of SNPs in parallel.

Just as important, like some but not all genomics companies in the 1990s, Affymetrix saw value in public-private research collaborations to drive product development. There was ample HGP-related federal funding to support such collaborations, which were also strategic because the academic research community and federal research labs comprised the initial target market of end-users for Affymetrix products. In 1997, with funds from a P01 Genome Science and Technology Centers (GESTEC) program grant, Affymetrix established a user center at their company campus for academic and industrial scientists to experiment with their technology prior to its commercialization.[10] Affymetrix researchers collaborated with NHGRI director Francis Collins and colleagues at NIH to demonstrate the utility of microarrays to screen for single nucleotide mutations in the *BRCA1* gene implicated in breast cancer (Hacia et al. 1996).

[9] To "genotype" a SNP meant to assess which nucleotides or variants were present at that SNP locus, at both chromosomal copies in an individual's genome. As SNPs and methods for assessing them became more widespread, this verbing of the term "genotype" became common usage among genomics researchers.

[10] NIH RePORTER, Grant # 5P01HG001323-03.

Lander's lab also collaborated with Affymetrix. With funding from NHGRI, they began to search for and catalog SNPs across the genome, which they saw as representing a new pool of polymorphic markers for genetic mapping and disease studies.[11] In addition to being the first high-throughput study to identify a large set of SNP loci in the genome, the resulting publication also built on earlier efforts initiated at Affymetrix to adapt their chips to SNP genotyping (Wang et al. 1998). Affymetrix researchers had demonstrated that researchers could use arrays with sufficient redundancy of probes[12] to identify medically-relevant SNPs in human genomes (Cronin et al. 1996). Importantly, this was the first demonstration that microarray technology might play a key role in genetic variation studies of disease. The subsequent experiments with Lander's lab illustrated how chips could be used both to identify and catalog novel SNP loci in the genome ("SNP discovery") and to genotype multiple SNP loci simultaneously across many human genomes ("SNP characterization"). Soon other groups were using chips to genotype SNPs in candidate genes for various diseases.[13]

NHGRI had begun investing millions on extramural grants for SNP discovery, awarded through multi-institutional Requests for Applications (RFAs).[14] Concurrently, NHGRI director Francis Collins and other leading geneticists were calling for large-scale efforts to develop a high-resolution, freely available SNP map of the genome, hoping for private sector participation (Collins et al. 1997; Weiss 1998). Many academic researchers feared SNPs would be patented by multinational biotech and pharma companies, already racing to develop proprietary SNP maps to guide disease research and the development of new diagnostics and therapeutics (Marshall 1997).

---

[11] Talk presented at the American Society of Human Genetics annual meeting (see D. Wang *et al*. (1996) "Toward a third generation genetic map of the human genome based on biallelic polymorphisms," *American Journal of Human Genetics* 59: A3).

[12] Redundancy was encoded into the chips to improve the reliability of the genotyping data, allowing for multiple measurements for each SNP, as discussed below.

[13] See, for example, Halushka et al. 1999.

[14] Email from Francis Collins to NIH Director Harold Varmus, "The SNP Consortium," March 20, 1999, #1702, p131, NHGRI History of Genomics Program Archival Resource.

It was clear that such an effort would require significant international coordination and funding to achieve. In 1999, ten pharmaceutical companies in the United States and Europe along with the Wellcome Trust, a privately endowed biomedical research charity based in the UK and a leading supporter of the HGP's sequencing efforts through the Sanger Center, decided to jointly finance The SNP Consortium (TSC). Their aim was to generate an "industrial standard" open access SNP map for genomics researchers.[15] The TSC benefitted from NHGRI's growing managerial expertise with large genomics consortia projects. The experimental benchwork of the $53 million TSC was carried out by NHGRI-supported HGP academic sequencing centers (including the Whitehead Institute for Biomedical Research/MIT Center for Genome Research, Washington University Genome Center, the Sanger Center, Stanford University, and Cold Spring Harbor Laboratory), collectively known as the International SNP Map Working Group (Holden 2002).

The TSC began generating pilot data in early January 1999 and officially launched in April 1999. To discover SNPs, researchers targeted their searches to specific regions of the genome and identified polymorphisms by re-sequencing and comparing DNA sequence from many of the individuals whose DNA was being used for the HGP (Altshuler et al. 2000). [16] These methods could help identify previously unknown SNPs in human genomes but could not map their genomic locations. There was no high-throughput method for doing this until the HGP's draft human genome sequence was released in 2000. TSC-identified SNP sequences were then computationally aligned *in silico* to sequences in the HGP's draft genome, producing a SNP map.

By the conclusion of their work, the TSC estimated that the number of SNPs that might

[15]"NIH Membership Discussions: The SNP Consortium," July 8 1999, #1704, p28, NHGRI History of Genomics Program Archival Resource.
[16] See also NHGRI's Oral History Collection: Interview with David Bentley, 2017, available at https://www.youtube.com/watch?v=lk2q4KdlfQQ.

underlie common diseases numbered between 3-4 million (Holden 2002). But at the outset of the TSC, as with the RFLP maps, researchers debated and disagreed on the number of SNPs needed for a map comprehensive enough to power disease studies. Some estimated as few as 30,000 might be needed (A. Collins et al. 1999), while others proposed numbers closer to 500,000 (Kruglyak 1999). Both the TSC's CEO Arthur Holden and NHGRI director Francis Collins pegged the number at 500,000. The TSC initially aimed to identify 300,000 SNPs, but the availability of the HGP data facilitated surpassing these goals. TSC's scientific work neared completion in 2001, with a publicly available SNP map describing over 1.7 million human SNPs (International SNP Map Working Group 2001; Holden 2002), about 11-12% of estimated SNPs in the genome (Kruglyak and Nickerson 2001; Brooks 2003). All SNP information was deposited in the public database hosted at NIH, dbSNP.

As the TSC was gearing up, the NHGRI sponsored a special supplement of the journal *Nature Genetics*, with reviews considering the potential impacts and challenges facing the use of microarray technology within the wide range of genomics applications for which it was being developed (A. Collins et al. 1999). At the time, SNP arrays were about 1.28cm square and could assay a few thousand SNPs in parallel. Before the TSC formally concluded, Affymetrix began marketing the first commercial SNP chips for life sciences research (Lipshutz et al. 1999). Though their entry-level arrays could genotype 2000 SNPs simultaneously in an individual's genome, this SNP count was still considered too sparse to fully support GWAS.

The early developments at Affymetrix did not, however, assure a monopoly in the commercial microarray market. A rival biotech, Illumina, had formed in 1998 around the "bead array" technology created by chemist David Walt and colleagues at Tufts University (Michael et al. 1998). Illumina researchers, some of whom were former Affymetrix scientists, began adapting bead arrays for a range of commercial-scale DNA and RNA analytic applications, including SNP genotyping. Instead of printing and

immobilizing short DNA probes on glass slides, the Illumina microarrays consisted of silica beads, each covered with hundreds of thousands of copies of a DNA probe sequence that could hybridize and bind to complementary sequences in an individual's genomic DNA. Beads with the same probe were assembled in a microscopic pit called a microwell, and several thousand microwells (each with probes of unique sequence) were etched on a substrate (initially fiber optic bundles and later glass slides). The Illumina technology was even more miniaturized than that of Affymetrix, capable of assaying several times more SNPs in parallel on the same size chip and touting an even higher "information density" (Oliphant et al. 2002). In July 2001, Illumina launched a SNP genotyping service for academic and industry researchers to make use of its pilot arrays.[17] Illumina had, like Affymetrix, also received federal funds under NHGRI's "SNP RFA" grants program, intended to support SNP discovery and the development of technology to reduce genotyping cost and increase throughput (the number of SNP loci that could be assayed simultaneously). Other commercial grantees supported by these funding streams included Orchid Biosciences and Genometrix, both developing high-throughput array technologies for SNP genotyping[18] and both potential competitors to Affymetrix and Illumina.

SNP Chips: Tools for Producing and Interpreting Human Genetic Variation

Genetic markers have been applied to the purpose of marking difference between individuals, families, and eventually groups of individuals, playing a central role in experimental genetics since its beginnings. As TSC researchers expanded the SNP map, they drew on human population genetics to establish theoretical frameworks for rendering SNP markers into tools for human genetic variation studies. Their thinking was influenced by the "out of Africa" theory in population genetics, developed in the 1980s. The theory suggests that anatomically modern humans originated in Africa, that

[17] Correspondence from Illumina co-founder John Stuelpnagel to Francis Collins, February 26 2002, #2295, p98, NHGRI History of Genomics Program Archival Resource.
[18] "SNP RFA Awards," (1999), #1704, p9, NHGRI History of Genomics Program Archival Resource.

between 50,000-200,000 years ago a subset of humans migrated out of Africa, and that this subset was further subdivided into groups that settled on each of the other continents on which humans live today. The theory lent itself to the view that continental barriers provided a major axis for human differentiation. It played a central part in the TSC's efforts to set the epistemological boundaries for how human genetic diversity ought to be studied and how it ought to be conceptually ordered and structured.

Drawing on this theory, TSC researchers advanced the view that SNPs could represent patterns of human genetic variation if SNP variants adjacent to each other on a chromosome were grouped together into "blocks" of DNA. They asserted that these blocks differed across continental lines, and that knowledge of these blocks along the chromosome, known as "haplotypes" in the parlance of population genetics, could accelerate GWAS and other studies of disease. Concomitant with the TSC, from 1998 to -2000, several groups had studied haplotypes in a handful of human genes (Reich et al. 2001; Daly et al. 2001; Patil et al. 2001; International SNP Map Working Group 2001; Gabriel et al. 2002), generating enthusiasm among human genome scientists for further work in the area (Brooks 2003). The idea that some SNP variants might reflect continent-specific differences was also promoted by Mark Shriver and colleagues, who in 1997 had begun to propose a handful of SNP loci that they argued could distinguish continental ancestries.[19]

A central tenet of this work was the assertion by TSC researchers that haplotype blocks would vary in content and length when examined in individuals from "different human groups," which they delineated in terms of major geographic regions (such as continents and sub-continents) (Kruglyak 1999; Reich et al. 2001; Gabriel 2002). This approach to ordering human variation reflected the way TSC DNA donors had been described. The

[19] See Fullwiley (2008) and Rajagopalan and Fujimura (2012) for more thorough treatment of the politics and limitations of these ideas.

TSC SNP map was developed using the DNA Polymorphism Discovery Resource (DPDR), which consisted of DNA collected by NIH researchers from 24 unrelated donors in the United States. They described the DPDR as a "representative" set of DNAs from donors of Caucasian, African, Asian, or "other" ancestry "reflecting the ethnic diversity of humankind" (Holden 2002). Although the DNA in DPDR was not labeled with racial or ethnic identifiers, the TSC posited human genetic variation as connected to continental ancestry, which in turn became incorporated into genomic databases and tools. Indeed, the TSC was drawing on a longer history of studies within population genetics, framed around the distribution of global human variation among continental race groups.[20]

Fueled by scientists' growing interest in the global extent and distribution of human genetic variation, [21] NHGRI decided, in the latter half of 2001, to continue the TSC's efforts through what became known as the International Haplotype Mapping Project (HapMap). Advocated by the International Human Genome Sequencing Consortium (2001), the "haplotype map of the human genome" would catalog haplotype patterns in DNA from donors located around the world, described as the "next key step of the HGP" for medical genetics and genome-wide association studies of disease.[22]

The TSC had already laid the groundwork for such an effort, as had Affymetrix. In

[20] For example, see the histories of the Human Genome Diversity Project, which generated controversy because of its race-based population genetics approach (Gannett 2001; M'Charek 2005, Reardon 2005). For historical analyses of postwar studies of genetic variation organized around blood and chromosomal polymorphisms in racialized groups, including prisoners, patients, survivors of radiation fallout from atomic bombs, ethnic groups, and inhabitants of remote islands, see Gannett and Griesemer (2004), Marks (2012), Mukharji (2014), de Chadarevian (2014), Bangham (2014), Widmer (2014), Suárez-Diáz (2014), and Lipphardt (2014).

[21] For example, in May 2001 Jim Weber wrote to NHGRI director Francis Collins urging the Institute to fund the establishment of a common resource of DNA and cell lines to further studies of linkage in the human genome, outlining a detailed plan for how and from which people to collect DNA. Weber's plan advocated for studying families from five "populations" (European, Asian, sub-Saharan, and two reproductively isolated populations) (#888, NHGRI History of Genomics Program Archival Resource). The NHGRI's HapMap ultimately pursued a different study design, analyzing some groups as trios of two parents and a child, and others as unrelated individuals, and deciding not to study any group deemed "isolated."

[22] "A Haplotype Map of the Human Genome: Project Summary," July 23 2001, #1696, p20, NHGRI History of Genomics Program Archival Resource.

October 2000, Affymetrix spun out a genomics subsidiary called Perlegen Sciences, Inc., which completed a large-scale scan of the 24 human genomes in the DPDR, to uncover broad patterns of genetic variation that might be related to health and disease (Patil et al. 2001). Perlegen built a proprietary database of SNP variant frequencies of about 1.6 million SNPs. Their goal was to partner with pharmaceutical companies to develop new drugs around patient subpopulations whose genomes carried these SNP variants (Peacock and Whiteley 2005). More immediately, the NHGRI's HapMap project bought and used their SNP data and services, as described below. Fodor, as chairman of both Affymetrix and Perlegen, anticipated a large role for both companies in future GWAS, stating in 2000 that no other entity had "adequate technology to look at whole genome patterns found across many individuals."[23]

Once again, debates ensued over the number of SNP markers that needed to be characterized for a haplotype map of the human genome. As the International SNP Map Working Group (2001) had noted, "The required density of markers will depend on the complexity of the local haplotype structure, and the distance over which these haplotypes extend." Initially, NHGRI anticipated having to genotype at least two million SNPs, including those identified by the TSC and those available in the public catalog dbSNP. But in pilot studies, researchers at HapMap genotyping centers in the USA (many of which had participated in both the HGP and TSC) concluded that these would prove insufficient to properly describe the haplotype structure of human variation around the world. Thus, despite the efforts of the TSC to make sufficient numbers of SNPs public, a significant fraction of the HapMap budget was set aside for purchasing proprietary SNPs (which were later deposited into the public domain as part of HapMap). Proprietary SNPs came from two companies: during HapMap Phase I from Applied Biosystems which was working on its own haplotype map for GWAS studies (De

[23] "Affymetrix Announces Formation of New Genomics Company: Perlegen Sciences," Affymetrix press release, Oct. 3 2000, https://ir.thermofisher.com/investors/news-and-events/news-releases/news-release-details/2000/Affymetrix-Announces-Formation-of-New-Genomics-Company-Perlegen-Sciences/default.aspx

La Vega et al. 2002), and for HapMap Phase II from Perlegen Biosciences, which had developed industrial-scale methods for SNP discovery. This brought the total number of characterized HapMap SNPs to almost 8 million by 2004.[24]

HapMap researchers decided to organize their catalog of variation by geographic source of donors' DNA. After extensive discussion and consideration of the ethical, legal, and social implications of naming and specifying human groups for such research, HapMap organizers decided that the first phase of the project would examine DNA donated by 45 individuals in Tokyo, Japan (designated as JPT), 45 Han Chinese individuals in Beijing, China (designated as CHB), 90 individuals among the Yoruba in Ibadan, Nigeria (designated as YRI), as well as DNA that had already been collected in the 1980s from 90 Utah residents with ancestry from northern and western Europe (designated as CEU). HapMap's official guidelines for use emphasized that researchers should refer to the DNA sets using the entire geographic location where they had been collected, specifying why neither the DNA nor the data should be interpreted as representative of larger populations, either national, ethnic, or racial. But both producers and users of the HapMap data began to operationalize the DNA sets as if they were representative of continental genotypes. Researchers came to view the SNP patterns and haplotypes present in the YRI set as representative of the kind and extent of genetic diversity they might discern across Africa. Similarly, the CEU DNA was seen as indicative of patterns of human genetic variation in people living in Europe, and the CHB and JPT DNA were seen as representative of variation among people in East Asia. These ways of framing difference by continental ancestry extended the efforts of the TSC and the DPDR. Furthermore, they presaged the conceptual frameworks and logics of population difference that researchers would operationalize in the development of SNP chips for GWAS, despite concerns and efforts to "avoid race."[25] This would have important

[24] NHGRI's Oral History Collection: Interview with Jim Mullikin, 2017, available at https://www.genome.gov/27552689/all-videos/

[25] Researchers' understanding of ancestry differs from their understanding of race and ethnicity. But the line has blurred, particularly when the labels used for different DNA collections, operationalized as

consequences for how researchers would design and use SNP chips for disease studies, as we describe next.

The NHGRI established sub-contracts to enlist private cooperation in generating HapMap data, such as with Perlegen. These funding streams in turn also rendered it a key financial supporter for the development of high-throughput commercial technologies for genotyping, including microarrays, which the NHGRI and many prominent geneticists felt was desperately needed if researchers were to effectively make use of SNP variation for disease studies (Hacia and Collins 1999). The grants programs thus facilitated symbiotic relationships between federal and private research entities during the HapMap project. Chips were by no means the inevitable choice for genotyping technology, but a constellation of factors, including their utility to researchers constructing the HapMap and planning for GWAS, contributed to their eventual dominance.

The first phase of HapMap, which launched in October 2002, was as much an exploratory period for assessing different genotyping technologies that were under private development as it was the inaugural data-generating period for the fledgling project. The NHGRI issued an RFA in 2002 to allow "ample opportunity for competition amongst genotyping platforms, as [the NHGRI Council] had some concerns about prematurely crowning any particular technology as the dominant one at this point."[26] Several companies won awards through the RFA, including Affymetrix (which was developing its "120K chip") and Illumina (developing its "40K BeadArray"), the first two prototypes of SNP chips, as well as non-chip technologies undergoing development at Sequenom, Perkin-Elmer/Applied Biosystems, Orchid Biosciences, and Third Wave Technologies. In pilot studies, as researchers characterized DNA from HapMap donors

---

populations, have traveled into broader social and political domains where they are interpreted in sometimes unintended ways (see Fujimura and Rajagopalan 2011).

[26] Correspondence from Francis Collins to Arthur Holden, Feb. 12, 2002, #2295, p90, NHGRI History of Genomics Program Archival Resource.

using these high-throughput genotyping technologies, they also assessed the quality of the data they were obtaining and the efficiency of the organizational and work processes they had to innovate to use each technological platform. Across platforms, they compared genotyping accuracy and efficiency, cost per SNP, and the proprietary analytic software packages for data collation and interpretation. Companies were competing for potentially lucrative downstream NIH contracts, as HapMap was expected to expand its efforts in later phases.

In side-by-side comparisons, HapMap researchers distinctly favored the Affymetrix and Illumina SNP chips for speed, reliability, ease of use, accuracy, and reproducibility of genotyping. These two companies went on to provide many of the key technological foundations and resources for the public efforts to investigate human genetic variation. Because researchers and laboratories involved in generating data for the HapMap were also among those best positioned (in terms of funding and resources), and indeed among the first, to launch high-throughput disease association studies, the preference for Affymetrix and Illumina solidified the competitive success of both platforms in what became the burgeoning field of GWAS.[27] The financial and institutional support from NHGRI and the imprimatur of NHGRI's HapMap project proved a significant validation of both companies' technologies, which by 2004 emerged as leading commercial suppliers (and competitors) for the growing research-based market for SNP genotyping. Affymetrix and Illumina designed and mass-produced fixed content chips (to assay pre-selected SNPs), but also provided services to help researchers custom-design their own chips to assay SNPs of their choosing for GWAS on particular diseases. Both companies continued to develop business models in which they collaborated closely with researchers, aligning new versions of the technology to users' needs. As we discuss below, academic researchers were just as involved in driving the diversification of the technology as corporate researchers and innovators.

[27] GWAS remains one of the most common study designs in human genomics research, and virtually all of the hundreds of GWAS published each year employ SNP chips.

Importantly, the epistemic traffic was two-way; the commercial SNP chips were not simply tools for future studies like GWAS, based on HapMap data, but rather generative of HapMap itself. The HapMap project was as much a user as an enabler of microarray technology. As we discuss below, building the successive versions of SNP chips and the HapMap was an iterative, self-reinforcing process, where data and outputs from HapMap informed new chip designs, and where new chip designs were mobilized to generate additional genotypes for successive phases of HapMap.

Composing Groups and Group Differences Chip by Chip

HapMap data featured prominently in microarray design considerations as companies built the first SNP chips for genome-wide association studies of disease. The Affymetrix GeneChip Human Mapping 100K[28] array was released in 2004; the Affymetrix 500K Array. released in September 2005, could assay a fixed set of 500,000 SNP loci chosen "quasi-randomly" on the basis of technical performance from among the SNPs identified in HapMap phase I. In contrast, Illumina's first chips for GWAS, the Human Hap300 BeadChips, contained about 300,000 SNPs selected based on haplotypes that HapMap researchers had described in the CEU DNA (Pe'er et al. 2006). Researchers preparing to conduct GWAS viewed the Affymetrix 500K array as the first SNP chip to have sufficient genome coverage to permit high-powered association studies of disease,[29] and the earliest GWAS used the Affymetrix 100K (Klein et al. 2005) and 500K arrays (Saxena et al. 2007; WTCCC 2007) to generate their data. Affymetrix chips were much less expensive and ready for market earlier than Illumina chips, but the latter became increasingly popular through their use in GWAS conducted by deCODE Genomics using

[28] SNP chip naming conventions specified how many SNPs a chip could assay, as this continued to be seen as directly indicative of their utility for disease studies; for example, "100K" meant the chip assayed 100,000 SNPs in parallel.

[29] Interview with a genotyping researcher who assisted with chip re-design, month and day? 2007.

DNA in its Icelandic biobank beginning in 2007 (Gudmundsson et al. 2007).

It was (and remains) not technically feasible to genotype all known human SNPs in an individual's DNA. Although efforts to enhance miniaturization had improved the chips' capacity, they continued to have a finite amount of space, accommodating only a small subset of known SNPs, what one researcher called "real estate."[30] Several constraints shaped researchers' determinations of which and how many SNPs to include in any given chip design. For example, some SNPs "behaved poorly" during array-based genotyping, repeatedly generating inconclusive results, so researchers eliminated these SNPs from consideration during quality control checks. In addition, SNPs were not considered equally informative; some were seen to give more information than others, as discussed below. With the goal of "maximizing coverage" across the genome, SNP choices were made according to researchers' computational estimates of the statistical "power" they could extract from these choices in disease association studies. This was done by pitting the possible SNP sets against each other in side-by-side comparisons and estimating the likelihood of finding a disease-associated SNP in one set versus another.

Researchers also made these decisions by appealing to the HapMap data, guided by population genetic theories about historical continental migrations, pegging technical constraints to the frameworks they had devised for making sense of haplotypes and the information they could glean from the correlations between adjacent SNPs in haplotypes. These frameworks underwrote a shared understanding of the patterns and extent of human genetic variation, in which SNP variant frequencies and haplotypes were seen to differ across continental human groupings, influencing chip design.

With the first few GWAS underway, researchers prominently involved in the HapMap began to re-assess the SNPs that were represented on the early Affymetrix 500K and Illumina Hap300 Beadchip arrays. They concluded that both chips vastly

[30] Interview with a lead scientist on the Affymetrix chip re-design project, month and day, 2007.

underrepresented the much more diverse SNP variation of people whose ancestry was ascribed to Africa or Asia. They argued that the 500K chip performed better in the HapMap YRI DNA than Illumina chips did, but both worked best in CEU DNA (Pe'er et al. 2006). Although Affymetrix had not chosen the 500K SNPs to represent any particular DNA, in practice their 500K chip was seen as more useful for genotyping individuals with "European" ancestry and less useful for individuals with "African" ancestry. "Usefulness" was assessed both quantitatively and qualitatively; the 500K chip did not have enough markers on it, or the right kind, to capture the haplotypes that were seen as more relevant to individuals of "African" ancestry.

This notion that one chip could not fit all genomes drew on an idea popularized by HapMap, that genomes from individuals of European ancestry looked less diverse in their SNP variants, while genomes from individuals whose ancestry was ascribed to Asia or Africa exhibited greater sequence variation.[31] "African genomes" were said to display features of an "older" population, since peoples on other continents were thought to have originated from small founder populations that split from a single parental population in Africa. That original population, they posited, had experienced a longer evolutionary history during which DNA recombination could scramble and shrink haplotypes, more so than in groups that had experienced population bottlenecks while migrating to the other continents. The shorter haplotypes of genomes with African origin, researchers claimed, had to be compensated for on the chips by genotyping many more SNP loci in individuals of African ancestry than would be needed in genomes of people with ancestries from Asia or Europe.

This view of human genetic variation, constructed through the lens of SNPs and chips, gave space for some perspectives on population and difference to flourish and constrained others. In particular, what was at stake in these decisions was the

[31] These views drew on earlier propositions in human population genetics; see Halushka et al. (1999), Cavalli-Sforza et al. (1994), Cargill et al. (1999), and Zietkiewicz (1997).

definition and significance of human genetic differences, the conceptual frameworks humans should use to consider, organize, work with, and ultimately act on genetic differences. These choices had significant implications and consequences for our understandings of genetic difference and disease. They produced continental-level variation as a significant factor for disease studies, and the result was that subsequent designs of SNP chips became explicitly associated with different continental-level ancestries.

Since Illumina chips were more expensive, researchers reasoned that they could leverage the information in HapMap haplotypes to more "intelligently" choose which SNPs to genotype on the chip. This could dramatically increase the "information content" of Affy chips, bringing them more in line with Illumina's pricier alternatives. Motivated by a desire to address the limitations of the Affymetrix chips, a laboratory prominently involved in HapMap initiated a multi-year collaboration with Affymetrix to design the next generation of SNP chips together. They argued that continued use of the existing Affymetrix and Illumina chips to study genomes of non-European ancestry would hamper researchers' ability to identify key SNPs associated with disease. In their first attempt to redesign the Affymetrix 500K chip, they addressed its built-in redundancies. The 500K chip included multiple probes for assessing each SNP, with some probes measuring one strand of the DNA and the rest measuring the other strand. Computational analysis, averaged across the redundant measurements from all of the probes, yielded a final assessment (or "call") of the variant at any given SNP. The chip redesigners refined the computational algorithms that specified how to call a variant, and they empirically assessed the minimal configuration of probes needed to yield the most reproducible results for each SNP. This allowed elimination of redundant probes, freeing up space on the chip to include assays for additional SNPs. The redesigned microarray was called the Affymetrix 5.0 chip, commercially released to the research community in February 2007.

To redesign the Affymetrix 5.0 chip, the collaborators exploited the information content of what they called "tag SNPs." Each haplotype block in HapMap was believed to have at least one proxy (or "tag") SNP, which, if identified and then assayed, could provide information on chromosomally adjacent SNP variants in the haplotype "for free." Researchers argued that by genotyping a single tag SNP in an individual's DNA, they could infer (or in the scientific vernacular, "impute") the adjacent co-inherited variants that person carried by referring back to HapMap data. Thus, instead of including on the chip all the SNPs comprising a haplotype, they only needed to include DNA probes for genotyping tag SNPs. Tag SNPs were framed as an economical resource. They increased efficiency while retaining statistical power (genotyping as few SNPs as possible to get maximal information in a particular region of the genome). Tag SNPs allowed researchers to justify genotyping fewer SNPs in any given region of the genome. This reduced what they called the "genotyping burden" for that region and made more of the chip's limited "real estate" available to interrogate other regions of the genome.

The collaborators experimentally screened and evaluated millions of additional SNPs that were polymorphic in the HapMap groups, identifying and selecting tag SNPs for inclusion on the chip. They then assessed the new chip's genotyping abilities, using as a benchmark the HapMap YRI group, deemed to represent the most genetically diverse (and most analytically complex) reference DNAs available to them. Commercially released in spring 2007, the Affymetrix Genome-Wide Human SNP Array 6.0 represented nearly 1 million SNPs, twice as many as its predecessor. With the extra "real estate" opened up during development of the 5.0 chip, the 6.0 chip also included probes to detect almost a million copy number sequence variations, another type of genetic variation that chip designers argued would make the chip even more useful in studies of disease among people with ancestry from Asia or Africa. As one chip designer noted, the Affymetrix 6.0 chip captured about "70%" of the YRI haplotypes in the HapMap, compared to "40%" captured by the 500K chip.[32] Others added that the

[32] Interview with a genotyping researcher who assisted with chip re-design, month and day, 2007

Affymetrix 6.0 chip was "more inclusive" but still limited in its ability to capture the full extent of genetic variation in peoples other than those few who had donated DNA to the HapMap, particularly in areas of the world where genetic variation is thought to be much more extensive, such as among peoples in Africa. For some projects, the Affymetrix 500K chips were seen as sufficient and cost-effective, and versions remain in use today. But because the new 6.0 chips could assess more of the genome than the older chips, the data they generated was seen as more comprehensive and up-to-date, and it became an industry standard. Many labs conducting GWAS who could afford the redesigned (and more expensive) 6.0 microarrays adopted them for their studies regardless of how they viewed the possible ancestry of the DNA they were studying. Illumina had also developed a new version of the BeadChip released in 2006, called the "Sentrix HumanHap 650Y BeadChip," which they marketed as extending researchers' ability to assess genetic variation and conduct more robust GWAS in all individuals but especially those with ancestry from Africa.[33]

The HapMap was a crucial resource in the process of redesigning the chips, both as a dataset from which to select SNPs for inclusion on newer chip designs and as a benchmark against which to gauge the chips' comprehensiveness. But again, these new chips were also *generative* for HapMap; while researchers used the HapMap Phase II data to guide the design of the Affymetrix 6.0 chip, the Affymetrix 6.0 chip was used, along with Illumina Infinium Human1M-single BeadChips, to generate the HapMap Phase III data (International HapMap 3 Consortium 2010).

SNP chips represented a technology that, through the multiplication of its own variations or versions, aimed to iteratively capture and assess human genetic difference, but also to configure it in specific ways. The research priorities and agendas of leading

[33] "Illumina Introduces Sentrix(R) HumanHap650Y Genotyping BeadChip; Product Sets New Standard for SNP Density and Genomic Coverage on a Single Array," Illumina press release, June 29 2006, https://www.businesswire.com/news/home/20060629005187/en/Illumina-Introduces-Sentrix-HumanHap650Y-Genotyping-BeadChip-Product.

genomics consortia after the HapMap, including the NHGRI-led 1000 Genomes project, continued to drive SNP chip specialization for studies of researcher-delineated population groups in different parts of the world, for which they deemed existing chips statistically under-powered. For example, in 2013 the NHGRI-initiated "Population Architecture using Genomics and Epidemiology" (PAGE) consortium undertook a collaboration with Illumina to design a new array "to empower GWAS in diverse ancestry populations." Funded by the NIH, the re-designed array built on Illumina's Infinium HumanCore BeadChip, which had been described and marketed as most suited to DNA from individuals of European ancestry. Known as the "Multi-Ethnic Genotyping Array" (MEGA), the redesigned chip was created primarily to study DNA from individuals of African and/or Hispanic/Latino ancestry (Bien et al. 2016). The Kaiser Permanente Research Program on Genes, Environment and Health (RPGEH)[34] used Affymetrix's "Axiom" custom array design service to create what they called population-specific arrays, including a "European-optimized SNP array" (Hoffmann et al. 2011a) and arrays "optimized for individuals of East Asian, African America and Latino race/ethnicity" (Hoffmann et al. 2011b). Other research groups have used Affymetrix's custom design service to create arrays like the "Japonica" array designed for GWAS in people of Japanese descent (Kawai et al. 2015).

The collaborative design of microarrays for genetic variation studies demonstrates the significance of hybrid collectives of many types of research organizations working across public and private domains for the development of new genomics technologies. As successive iterations of chip designs have expanded in the numbers of SNPs they can assay, they have also increasingly reflected both prevailing disciplinary thinking about genetic variation and the historical traces of chipmakers' choices and decisions under particular constraints. Importantly, the scope and distribution of human genetic

[34] The RPGEH received a substantial amount of their funding from the NIH's 2009 Grand Opportunity, which supported large capital investments in biomedical research infrastructure. RPGEH established a biobank of DNA collected from over 100,000 volunteers in the Kaiser Permanente managed healthcare group and designed their ethnic-specific arrays in order to genotype these specimens (Hoffmann 2011b).

variation from the vantage of RFLPs looked different compared to genetic variation viewed from the vantage of SNPs and SNP chips. The highly polymorphic RFLPs were used to follow genetic difference and disease in families, while the less polymorphic SNPs became part of a view of genetic variation and disease that was built around "populations," loosely defined but often along continental lines. Alongside this shift, researchers reframed the kinds of diseases of interest to genomics, to those with complex etiologies and ambiguous genetic involvement. These were mutually reinforcing shifts. As any choice of methodological approach does, the chip constrained researchers' views of that which they studied -- genetic variation and disease; it restricted operational views of SNP variation to a million or so SNPs out of the estimated 80-100 million in the genome. Thus, it helped foster a permissive environment for viewing a small subset of all SNP loci (those that the chip assayed for) as a stand-in for the universe of human variation, reinforcing the idea that SNP chips could capture genetic variation in a continent-specific way.

Conclusion

This article has examined the particular visions of human groupings that motivated and guided how genome scientists in the United States in the 1990s and early 2000s thought of and produced human genetic variation, and how epistemic considerations regarding its apportionment among groups influenced the design of their studies of disease. In the process of making genetic markers and array-based technologies to track variation for disease studies, scientists also made commitments to particular ways of describing, cataloging, and "knowing" human genetic variation (ways that align with data from related fields about the geographic, temporal, and archaeological moorings of human groups across time and space). SNPs and SNP chips were mobilized along a trajectory in genomics research that exerted a kind of path dependence, locking in particular views of human genetic variation. By examining how SNP chips were used to operationalize

population genetic theories about DNA variation and ancestry, we illuminate some of the historical roots of and routes through which different contemporary human groupings have come to be bounded and understood in terms of sequence differences at the level of DNA.

We have traced how population genetics theories assumptions about the age and movements of populations during human history, and their relative extent of genetic variation, was given materiality and meaning through the production of genetic markers, genetic maps, and haplotypes. The design and uses of SNPs and SNP chips together cemented an understanding of genetic variation in the early 2000s that relied heavily on continent-based ideas about the organization of human differences. The differences of interest that took shape in the context of twenty-first century medical genomics had as much to do with differences in frequency or rate of occurrence described along continental lines as they did in absolute nucleotide differences.

Importantly, SNPs followed a particular historical trajectory, gaining prominence through large international consortia projects oriented around the genome in the 1990s, within a particular web of institutional priorities, disciplinary conventions, bureaucratic choices, and articulated scientific needs. Sequence differences were framed as a necessary corollary for genetic mapping and the study of diseases, and mapping difference later became an end in itself. Geneticists' professed needs for maps of variation articulated a vision of human genetic diversity that, coupled with technological shifts and research preferences, enabled SNPs and SNP chips to flourish and become the de facto standard for measuring human genetic difference in disease studies. None of these developments were inevitable. Rather, we have shown how these were the product of specific decisions and choices through decades of research, and how genetic tools for assessing difference were fashioned within particular sociotechnical contexts and under certain conditions.

By examining the kinds of considerations, decisions, and choices that went into the design and construction of microarray technologies for GWAS, we highlight how, as disease genetic studies moved away from family linkage studies to population-wide studies of unrelated individuals, an emphasis on "population" became dominant. Instead of individuals, (population) groups, themselves heterogeneous and impossible to circumscribe in any precise way, became circumscribed and tamed for the laboratory, standardized as objects of analysis through the tools of population-specific chips. These same conceptual framings of population prompted researchers to ascribe certain limits to their tools, such that which SNPs were chosen for inclusion on the chips came to have fundamental consequences for disease research, prompting iterative SNP chip designs for human groups specified through the very work for which the SNPs and the chips were designed. Thus, the story is one of both standardization and differentiation; technologies became simultaneously standardized and differentiated.

Tracking the development of the SNP chip reveals how continent-based notions of human difference and genetic diversity have become encoded and embedded within the new technologies of genomics. As discussed, epistemic commitments (to differences, coded in particular ways) become embedded in techniques and objects, which are then mobilized to do work that reinforces those commitments, such as GWAS. Thus, techniques and objects are not only material but also epistemic, partly constituted of the physicality of materials and biology and partly driven by (and productive of) the epistemic commitments of their designers. Their particular uses, boundaries, and specific formulations are also shaped by scientists' imaginations, measurements, negotiations, and collaborative work.

Acknowledgements:

We gratefully thank respondents at our field sites who generously shared their time and allowed us to observe their work. We also thank Christopher Donohue and the staff at the NHGRI History of Genomics Program for facilitating access to the NHGRI Archival

Resource, and for organizing the workshop “Capturing the History of Genomics,” where an early version of this article was presented. The article also benefited from questions and comments at presentations in the U.S, Wales, and Norway. We thank the two anonymous reviewers who offered valuable feedback to strengthen the manuscript. RMR acknowledges research support from the Life Sciences Foundation (now the Science History Institute). JHF acknowledges research support from the Russell Sage Foundation and the Center for Advanced Study in the Behavioral Sciences. This work was supported by NSF grant 0621022, NIH/NHGRI grant R03HG005030, NIH/NHGRI grant R03HG006571, a University of Wisconsin-Madison Vilas Life Cycle Professorship Award, and grants from the University of Wisconsin Institute for Race and Ethnicity and the University of Wisconsin Graduate School.

References

Altshuler, D., V. J. Pollara, C. R. Cowles, W. J. Van Etten, J. Baldwin, L. Linton, and E. S. Lander. 2000. “A SNP map of the human genome generated by reduced representation shotgun sequencing.” *Nature* 407 (6803): 513-516.

Bangham, J. 2014. “Blood groups and human groups: Collecting and calibrating geneticdata after World War Two,” *Studies in the History and Philosophy of Biological and Biomedical Sciences, Part A* 47: 74-86.

Bien S. A., G. L. Wojcik, N. Zubair, C. R. Gignoux, A. R. Martin, J. M. Kocarnik, L. W. Martin, *et al*. 2016 “Strategies for enriching variant coverage in candidate disease loci on a multiethnic genotyping array.” *PLoS ONE* 11(12): e0167758.

Botstein D., R. L. White, M. Skolnick, and R. W. Davis. 1980. “Construction of a genetic linkage map in man using restriction fragment length polymorphisms.” *American Journal of Human Genetics*. 32(3): 314-331.

Brooks, L. D. 2003. “SNPs: Why do we care?” In *Single Nucleotide Polymorphisms: Methods and Protocols*, edited by P-Y. Kwok, 1-14. Totowa: Humana Press.

Cargill M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, *et al*. 1999. “Characterization of single-nucleotide polymorphisms in coding regions of human genes.” *Nature Genetics* 22: 231-238.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press.

Collins, A., C. Lonjou, and N. E. Morton. 1999. “Genetic epidemiology of single-nucleotide polymorphisms.” *Proceedings of the National Academy of Sciences USA.* 96(26): 15173–15177.

Collins F. C. 1991. “Of needles and haystacks: Finding human disease genes by positional cloning,” *Clinical Research* 39(4): 615-623.

Collins F. C. and D. Galas. 1993. “A new five-year plan for the U.S. Human Genome Project.” *Science* 262(5130): 43-46.

Collins, F. C., M. S. Guyer, and A. Chakravarti. 1997. “Variations on a theme: cataloging human DNA sequence variation.” *Science* 278: 1580–1581.

Cronin, M. T., R. V. Fucini, S. M. Kim, R. S. Masino, R. M. Wespi, and C. G. Miyada. 1996. “Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Human Mutation* 7(3): 244-255.

Daly M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. 2001. “High-resolution haplotype structure in the human genome.” *Nature Genetics* 29(2): 229-232.

de Chadarevian, S. 2014. “Chromosome surveys of human populations: Between epidemiology and anthropology.” *Studies in the History and Philosophy of Biological and Biomedical Sciences Part A* 47: 87-96.

Delahunty C., W. Ankener, Q. Deng, J. Eng, and D. A. Nickerson. 1996. “Testing the feasibility of DNA typing for human identification by PCR and an oligonucleotide ligation assay.” *American Journal of Human Genetics.* 58(6): 1239-1246.

De La Vega F. M., D. Dailey, J. Ziegle, J. Williams, D. Madden, and D. A. Gilbert. 2002. “New generation pharmacogenomic tools: a SNP linkage disequilibrium map, validated SNP assay resource, and high-throughput instrumentation system for large-scale genetic studies.” *Biotechniques* 32: S48-S54.

Doel, R. E. and T. Söderqvist. 2006. *The Historiography of Contemporary Science, Technology, and Medicine: Writing Recent Science*. New York: Routledge.

Fodor S. P., J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. 1991. “Light-directed, spatially addressable parallel chemical synthesis.” *Science* 251(4995): 767-773.

Fujimura, J.H. and R. Rajagopalan. 2011. “Different differences: The use of ancestry versus race in biomedical human genetic research.” *Social Studies of Science*, 41(1): 5-30.

Fullwiley D. 2008. “The biologistical construction of race: 'Admixture' technology and

the new genetic medicine." *Social Studies of Science,* 38(5): 695-735.

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, *et al*. 2002. "The structure of haplotype blocks in the human genome." *Science* 296(5576): 2225–2229.

Gannett, L. 2001. "Racism and human genome diversity research: The ethical limits of 'population thinking.'" *Philosophy of Science* 68(3): S479-S492.

Gannett L. and J. Griesemer. 2004. "The ABO blood groups: mapping the history and geography of genes in *Homo sapiens*." In *Classical Genetic Research and its Legacy: The mapping cultures of twentieth-century genetics*, edited by J.-P. Gaudillière and H.-J. Rheinberger. New York: Routledge.

Gaudillière J.-P. and H.-J. Rheinberger. (eds.). 2004. *From Molecular Genetics to Genomics: The Mapping Cultures of Twentieth-Century Genetics*. New York: Routledge.

Gudmundsson J., P. Sulem, A. Manolescu, L. T. Amundadottir, D. Gudbjartsson, A. Helgason, T. Rafnar, *et al*. 2007. "Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24." *Nature Genetics* 39(5): 631-637.

Guo Z, R. A. Guilfoyle, A. J. Thiel, R. Wang, and L. M. Smith. 1994. "Direct fluorescence analysis of genetic polymorphisms by hybridization with oligonucleotide arrays on glass supports." *Nucleic Acids Research* 22(24): 5456-5465.

Hacia J. G., L. C. Brody, M. S. Chee, S. P. Fodor, and F. S. Collins. 1996. "Detection of heterozygous mutations in *BRCA1* using high density oligonucleotide arrays and two-colour fluorescence analysis." *Nature Genetics* 14(4): 441-447.

Hacia J. G. and F. C. Collins. 1999. "Mutational analysis using oligonucleotide microarrays," *Journal of Medical Genetics* 36: 730-736.

Halushka M. K., J. B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, *et al.* 1999. "Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis." *Nature Genetics* 22(3): 239-247.

Hoffmann T. J., M. N. Kvale, S. E. Hesselson, Y. Zhan, C. Aquino, Y. Cao, S. Cawley, *et al*. 2011a. "Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array." *Genomics* 98(2): 79-89.

Hoffmann T. J., Y. Zhan, M. N. Kvale, S. E. Hesselson, J. Gollub, C. Iribarren, Y. Lu, *et al*. 2011b. "Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm." *Genomics* 98(6): 422-30.

Holden, A. L. 2002. “The SNP Consortium: Summary of a private consortium effort to develop an applied map of the human genome.” *BioTechniques* 32: S22-26.

International HapMap 3 Consortium. 2010. “Integrating common and rare genetic variation in diverse human populations,” *Nature* 467(7311): 52-58.

International SNP Map Working Group. 2001. “A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.” *Nature* 409(6822): 928-933.

Kawai Y., T. Mimori, K. Kojima, N. Nariai, I. Danjoh, R. Saito, J. Yusada, *et al*. 2015. “Japonica array: improved genotype imputation by designing a population-specific SNP array with 1070 Japanese individuals.” *Journal of Human Genetics.* 60(10): 581–587.

Klein R. J., C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, *et al*. 2005. “Complement factor H polymorphism in age-related macular degeneration.” *Science* 308 (5720): 385–389.

Kruglyak L. 1999. “Prospects for whole-genome linkage disequilibrium mapping of common disease genes.” *Nature Genetics* 22: 139-144.

Kruglyak L. and D. Nickerson. 2001. “Variation is the spice of life.” *Nature Genetics* 27, 234-236.

Kruglyak L. 2008. “The road to genome wide association studies.” *Nature Reviews Genetics* 9: 314-318.

Lander E. S. 1996. “The new genomics: global views of biology.” *Science*. 274(5287): 536-539.

Lander E. S. and D. Botstein D. 1986. “Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms.” *Proceedings of the National Academy of Sciences USA* 83: 7353-7357.

Lipphardt V. 2014. “Geographical distribution patterns of various genes: Genetic studies of human variation after 1945.” *Studies in History and Philosophy of Biological and Biomedical Sciences Part A* 47: 50-61.

Lipshutz, R. J., S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart. 1999. “High density synthetic oligonucleotide arrays.” *Nature Genetics* 21: 20-24.

Marks J. 2012. “The origins of anthropological genetics.” *Current Anthropology* 53(S5): S161-S172.

Marshall E. 1997. "'Playing Chicken' over Genetic Markers." *Science* 278(5346): 2046-2048.

Matson R., J. Rampal, S. L. Pentoney, P. D. Anderson, and P. Coassin. 1995. "Biopolymer synthesis on polypropylene supports: Oligonucleotide arrays." *Analytical Biochemistry* 224(1): 110-116.

McCusick, V. 1992. *Mendelian inheritance in man: catalogs of autosomal dominant, autosomal recessive and X-linked phenotypes* (10th ed). Baltimore: Johns Hopkins University Press.

M'charek, A. 2005. *The Human Genome Diversity Project: An ethnography of scientific practice*. Cambridge: Cambridge University Press.

Michael, K. L., L. C. Taylor, S. L. Schultz, and D. R. Walt. 1998. "Randomly ordered addressable high-density optical sensor arrays." *Analytical Chemistry* 70(7): 1242-1248.

Mukharji, P. B. 2014. "From serosocial to sanguinary identities: caste, transnational race science and the shifting metonymies of blood group B, India c. 1918-60", *Indian Economic and Social History Review*, 51(2): 143-176.

Nikiforov T. T., R. B. Rendle, P. Goelet, Y. H. Rogers, M. L. Kotewicz, S. Anderson, G. L. Trainor *et al*. 1994. "Genetic Bit Analysis: A solid phase method for typing single nucleotide polymorphisms." *Nucleic Acids Research* 22(20): 4167-4175.

Oliphant, A., D. L. Barker, J. R. Stuelpnagel, and M. S. Chee. 2002. "BeadArray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping." *Biotechniques* 32: S56-61.

Patil N., A. J. Berno, D. A. Hinds, W. A. Barrett, J. M. Doshi, C. R. Hacker, C. R. Kautzer, *et al*. 2001. "Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21." *Science* 294(5547): 1719-1723.

Peacock E. and P. Whiteley. 2005. "Perlegen Sciences, Inc." *Pharmacogenomics* 6(4): 439-442.

Pe'er I., P. I. de Bakker, J. Maller, R. Yelensky, D. Altshuler, and M. J. Daly. 2006. "Evaluating and improving power in whole-genome association studies using fixed marker sets." *Nature Genetics* 38(6): 663-667.

Rajagopalan, R. and J.H. Fujimura. 2012. "Making history via DNA, making DNA from history: Deconstructing the race-disease connection in admixture mapping." In *Genetics*

and the Unsettled Past: The Collision between DNA, Race and History, edited by K. Wailoo, C. Lee, and A. Nelson. New Brunswick: Rutgers University Press.

Reardon, J. 2005. *Race to the Finish: Identity and Governance in an Age of Genomics.* Princeton: Princeton University Press.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. C. Richter, T. Lavery, *et al*. 2001. "Linkage disequilibrium in the human genome." *Nature* 411(6834): 199-204.

Risch N., and K. Merikangas. 1996. "The future of genetic studies of complex human diseases." *Science* 273(5281): 1516-1517.

Rogers S. and A. Cambrosio A. 2007. "Making a new technology work: The standardization and regulation of microarrays," *Yale Journal of Biology and Medicine* 80: 165-78.

Saxena R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, J. J. Roix, *et al*. 2007. "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." *Science* 316(5829): 1331-1336.

Schena M., D. Shalon, R. W. Davis, and P. O. Brown. 1995. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270(5235): 467-470.

Shostak S. 2005. "The emergence of toxicogenomics: A case study of molecularization." *Social Studies of Science* 35(3): 367-403.

Suárez-Diáz, E. 2014. "Indigenous populations in Mexico: Medical anthropology in the work of Ruben Lisker in the 1960s." *Studies in History and Philosophy of Biological and Biomedical Sciences* 47: 108-117.

The 1000 Genomes Project Consortium. 2015. "A global reference for human genetic variation." *Nature* 526(7571): 68-74.

U.S. Congress Office of Technology Assessment. 1988. "Mapping our genes. The genome projects: how big, how fast?" Washington: U.S. Government Printing Office.

Wang D. G., J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, *et al*. 1998. "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome." *Science* 280(5366): 1077-1082.

Weber J. L. and P. E. May. 1989. "Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction." *American Journal of Human Genetics.* 44(3): 388-396.

Weber J. L. 1990. “Human DNA polymorphisms and methods of analysis.” *Current Opinion in Biotechnology* 1(2): 166-171.

Weissenbach J. 1993. “Microsatellite polymorphisms and the genetic linkage map of the human genome.” *Current Opinion in Genetics and Development* 3(3): 414-417.

Weiss K. M. 1998. “In search of human variation.” *Genome Research* 8(7): 691-697.

Wellcome Trust Case Control Consortium. 2007. “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.” *Nature* 447(7145): 661-678.

Widmer, A. 2014. “Making blood ‘Melanesian’: Fieldwork and isolating techniques in genetic epidemiology (1963-1976),” *Studies in the History and Philosophy of Biological and Biomedical Sciences, Part A* 47: 118-129.

Zietkiewicz E., V. Yotova, M. Jarnik, M. Korab-Laskowska, K. K. Kidd, D. Modiano, R. Scozzari, *et al*. 1997. “Nuclear DNA diversity in worldwide distributed human populations.” *Gene* 205(1-2): 161−171.