The International HapMap Project (2002–2016)

Matthew Tontonoz
2025-08-13

Launched in 2002, the International HapMap Project was a collaborative effort among scientists from around the world to create a map of common patterns of genetic variation in the human genome. HapMap stands for haplotype map. A haplotype is a stretch of DNA nucleotides, or letters, that individuals inherit as a block because they lie relatively close together along a chromosome. For any particular region of a chromosome, there may be multiple different haplotypes present among humans, each characterized by a slightly different DNA sequence. By collecting and sequencing the DNA of initially 270 individuals from several different geographic regions, HapMap scientists were able to identify common haplotypes that exist among those individuals, as well as reliable markers to distinguish them. That collection of haplotypes and identifying markers—the HapMap—provided a shortcut for researchers who wanted to identify associations between those inherited DNA variants and particular human traits, especially diseases.

1. SNPs, Haplotypes, and Tag SNPs
2. The HapMap: Background and Context
3. Development of the HapMap
4. Ethical Considerations of HapMap
5. Use and Impact of the HapMap

SNPs, Haplotypes, and Tag SNPs

The HapMap is a tool that shows common patterns of genetic variation, in the form of haplotypes, located throughout the three billion base pairs of DNA in the human genome. While the human genome sequence is 99.9 percent identical from one person to the next, 0.1 percent of the DNA sequence varies from person to person. The most common type of DNA variation is a difference in one DNA letter, called a single nucleotide polymorphism, or a SNP. For example, where one person's DNA reads AAATCCG, another's might read AACTCCG. SNPs occur approximately once every 1,000 base-pairs when any two people are compared.

Because of how chromosomes pass from parents to children, SNPs that are relatively close to each other along a chromosome tend to be inherited together in a block—the haplotype. Because SNPs in the same haplotype are linked, scientists can infer the presence of other SNPs when they know the identity of a subset of SNPs in the haplotype. Researchers refer to that subset of SNPs that distinguish one haplotype from another as tag SNPs. Scientists estimate that the total number of common SNPs within the human genome is on the order of ten million. The number of tag SNPs needed to identify all common haplotypes is around 500,000.

By identifying many of the haplotypes that exist in humans as well as the tag SNPs needed to distinguish them, the HapMap makes it much easier for scientists to look for associations between genetic variations and particular traits, such as having a particular disease. Rather than needing to sequence the entire genome of many people to identify all of their SNPs, scientists can instead rely on sequencing a small number of tag SNPs to identify haplotypes and look for associations between those haplotypes and specific traits.

The HapMap: Background and Context

The International HapMap Project belongs to the tradition of genomic research that began with the Human Genome Project, or HGP, which ran from 1990 to 2003. The HGP's initial aim was to sequence the entire three billion base-pairs of DNA in one human genome, as well the genomes of several other species. By 1998, the HGP's goals had expanded to include a search for DNA sequence variation, with an initial focus on SNPs. The goal of identifying SNPs was to enable researchers to find associations between particular SNPs and human traits, particularly diseases. From 1998 to 2001, scientists looked for SNPs within the DNA from 450 anonymous individuals whose cells they made into a panel of cell lines called the DNA Polymorphism Discovery Resource. When scientists published the first draft of the human genome sequence in 2001, they also published along with it the identity and location of 1.4 million SNPs spread across the human genome. By October 2002, when the HapMap launched, there were 2.8 million known SNPs out of an estimated total of at least 10 million SNPs.

The HapMap emerged out of a recognition among genetic researchers that there were far too many SNPs to study individually in their search for ones associated with disease. They needed alternative approaches. The idea to focus on haplotypes came out of empirical studies showing that the average haplotype size in the human genome is 11,000 to 22,000 base-pairs long, with three to five possible haplotypes per region. Those findings suggested that researchers could identify DNA sequences associated with diseases by focusing initially on associations between haplotypes and diseases.

For example, if a researcher wanted to identify genetic variants that increase a person's risk of diabetes, they could compare the frequency of haplotypes between a group of individuals with diabetes and a group without diabetes. If any haplotypes appear to be more common in the group with diabetes, then the researcher has reason to believe that those haplotypes contain genetic variation that may influence a person's risk of developing diabetes. They can then focus their efforts on DNA variations located in that particular haplotype. The HapMap researchers themselves reported no plans to perform those sorts of comparisons. Rather, their aim was to permit others to use the HapMap in that way.

The particular diseases that HapMap scientists hoped their map would be useful for understanding were common, complex diseases such as heart disease, cancer, and diabetes. Those diseases differ from simple, Mendelian disorders in several important ways. For one, whereas Mendelian disorders such as cystic fibrosis and sickle-cell disease are determined by mutations in single genes that have particularly pronounced effects, common, complex diseases are influenced by many genes that individually have small effects and that interact with environment factors. As a result, the inheritance patterns for complex diseases are more complex and harder to predict than Mendelian diseases.

A further assumption undergirding the HapMap Project was the common disease–common variant hypothesis, which states that the genetic contributors to common diseases like heart disease, cancer, and diabetes are likely to be common in the human population.

Development of the HapMap

The United States National Institutes of Health, or NIH, held an initial planning meeting to discuss the feasibility and ethics of creating a haplotype map on 18 and 19 July 2001 in Washington, DC. The meeting included about 165 individuals from various backgrounds, including researchers who studied human genetics, population genetics, and anthropology, pharmaceutical and biotechnology representatives, social scientists, ethicists, representatives from various communities and disease groups, administrators from NIH institutes and international funding agencies, and journalists. Participants met in groups to consider ethical issues, discuss scientific feasibility, and decide which human populations to include in the project. Following a period of review, the organizers formally announced the creation of the International HapMap Consortium to pursue the HapMap Project at a second meeting in Washington, DC, held 27 and 29 October 2002. Organizers estimated the jointly run public-private project would cost 120 million dollars and take three years, until 2005, to complete. They planned to make the results quickly and freely available on the internet.

To create the haplotype map, HapMap scientists proposed collecting blood samples from 270 individuals from several populations living around the globe, with the goal of turning those blood samples into stable cell lines for later storage in a biobank. The Project defined a

population as a group of people with a shared ancestry and therefore a shared history and pattern of geographical migration. From those stable cell lines, HapMap researchers planned to identify all SNPs in the genome of each of the 270 individuals, and then identify haplotypes. Because humans are so genetically similar, HapMap scientists stated that they likely did not need to sample more than a single population to identify common haplotypes across the entire human population. However, they stated they decided to include several populations to ensure that the HapMap would include most of the common genetic variation as well as some of the less common, regionally specific, variation present in humans.

In addition to those scientific reasons, HapMap scientists also stated that including individuals from multiple populations was a matter of justice. They explained that if their Project did lead to some medical benefits, including people from a single area might lead to real or perceived benefits accruing to that group alone. Therefore, it was of ethical importance to include more than one population in the study.

The populations the HapMap researchers initially decided to study were ninety individuals living in Ibadan, Nigeria, who belonged to the Yoruba ethnic group, ninety individuals living in Utah, US, forty-five individuals living in Tokyo, Japan, and forty-five Han Chinese individuals living in Beijing, China, for a total of 270 individuals and blood samples. The sampled individuals from Nigeria and the US were related trios consisting of two parents and an adult child, while the individuals from other populations were unrelated. The choice of those particular populations was partly a practical matter. Researchers at Howard University in Washington, DC, and the University of Ibadan in Ibadan, Nigeria, already had an established research collaboration and had built a relationship of trust with the Yoruba people in Ibadan. Chinese investigators chose to collect samples at the Beijing Normal University residential community in Beijing, China, because it represented a diverse yet socially cohesive population of mostly ethnic Han individuals from all over China. Japanese researchers collected samples from five different communities in Tokyo within which participants previously participated in biomedical studies. Each research group was responsible for collecting the blood samples from the communities with whom they already had a relationship.

Once the researchers had collected the samples, a process that took place between 2002 and 2004, they sent them to the Coriell Institute for Medical Research in Camden, New Jersey, a non-profit biomedical research center that specializes in storing living cells and making them available to scientists for further study. The samples have population and sex identifiers listed but no information that could link them to individual donors. As the goal of the HapMap project was solely to identify patterns of genetic variation, the researchers did not include any medical or other phenotypic information with the samples. Scientists also collected more samples from members of each population than they ultimately placed in the biobank to help preserve anonymity. Therefore, sample donors could not know for sure that

their samples were included in the study. HapMap scientists also specified that scientists cannot commercialize the samples themselves, though future researchers may patent specific discoveries based on HapMap data.

For sequencing of SNPs, the HapMap organizers assigned different chromosomes or chromosome regions to different sequencing centers located around the world. For example, researchers at the University of Tokyo in Japan sequenced SNPs on chromosomes 5, 11, 14, 15, 16, 17, and 19, while researchers at the US-based biotech company Illumina in San Diego, California, sequenced chromosomes 9, 22, and X and regions 8q and 18q. Each group had access to all the cell lines they needed to complete their analyses. Researchers from all sequencing sites deposited their SNP and haplotype data into a public database.

The HapMap project went through three phases. Phase one ran from 2002 through 2005 and assembled the first haplotype map. Although the researchers collected 270 samples for phase one, they excluded one sample due to technical problems for a final total of 269 samples. Phase two, which ran from 2005 to 2007, added over 2.1 million SNPs to the original map in the same individuals as phase one of the project. In 2010, researchers released phase three, which provides SNP and haplotype information for nearly one thousand additional individuals from seven additional populations. Those populations are Maasai individuals in Kinyawa, Kenya, Luhya individuals in Webuye, Kenya, Chinese individuals in metropolitan Denver, Colorado, Gujarati Indian individuals in Houston, Texas, Tuscani individuals in Italy, individuals of African ancestry in the US Southwest, and individuals of Mexican ancestry in Los Angeles, California. In addition, the researchers obtained blood samples from several hundred additional individuals from the original four populations from phases one and two of the HapMap.

Ethical Considerations of HapMap

Ethical considerations were part of HapMap discussions from the first planning meeting in 2001. At that meeting, some participants expressed concern that the HapMap data would permit scientists to compare patterns of variation among both individuals and populations, which might lead to stigmatization of certain groups. For example, if researchers found a higher frequency of disease-associated genetic variants within a particular group, and scientists then overgeneralized that information to all or most of its members, that error could lead to socially harmful stereotypes of individuals belonging to that particular group. In response to those ethical concerns, the HapMap scientists established two separate planning groups, one devoted to technical and methodological issues, and one devoted to ethical and sampling issues. The latter group included population geneticists as well as experts in the ethical, legal, and social implications of genetics and genomics research. The HapMap scientists tasked those individuals with tackling two main questions, how to sample human

genetic variation to identify common haplotypes, and whether to name the populations from which the donors came.

Lurking in the background of those ethics discussions were memories of the Human Genome Diversity Project, or HGDP, an attempt by US researchers to collect DNA from what they referred to as isolated Indigenous human populations. After its launch in 1991, the HGDP became embroiled in controversy for two main reasons. One concerned the choice and definition of those populations from whom they planned to collect DNA. Some anthropologists criticized the implicit logic of the project, which seemed to equate those populations with evolutionary relics, living echoes of once-pure human groups that typified an earlier stage of human history. Another concern, voiced by representatives of indigenous groups themselves, was that the project appeared to be a form of biopiracy, in which Western scientists were extracting valuable resources in the form of DNA from vulnerable groups and then profiting off them.

To avoid those previous mistakes and controversies, HapMap researchers debated the best way to identify and sample populations. Their solution was, in part, to avoid collecting blood samples from vulnerable or Indigenous groups altogether. In addition, the researchers attempted to involve the community members in both the process of deciding on appropriate population descriptors as well as the informed consent process. HapMap researchers also sought to determine population membership in culturally specific ways, for example, by asking a donor whether all four grandparents were part of the Yoruba people, whether at least three or four grandparents were Han Chinese, and how Japanese individuals self-identified.

The main tool the ethics group proposed to avoid other controversies that befell the HGDP was a process of community engagement. The goal of community engagement was to allow people in the various localities to have input on the informed consent process, the sample collection processes, and the process of naming the samples. The recognition of the need for community engagement was also why the HapMap scientists did not use previously collected blood samples or cell lines in their project and instead collected new blood samples. In addition to community engagement, the HapMap scientists also formed a Community Advisory Group at each collection site to deal with concerns that might arise in the future. Through the Community Advisory Group, individuals from the four sample locations could opt to have their samples removed, although SNP data that was already in the database could not undergo removal as it would already be in the public domain.

Though HapMap leaders put significant effort into considering how and where to obtain samples ethically, their ultimate choice of focusing on one population each from Europe and Africa and two from Asia engendered criticism from some observers. To those observers, that sampling strategy seemed to suggest a racial division of world, with a few individuals from

each continent seemingly representing the genetic diversity of the continent as a whole. HapMap scientists stated explicitly in their published papers that the sampling strategy should not be seen that way, but such stipulations did not prevent some downstream users from viewing the samples in that way.

Use and Impact of the HapMap

The International HapMap Consortium announced their initial haplotype map in 2005 in the journal *Nature*. There, they describe identifying more than one million SNPs from 269 DNA samples from four geographically dispersed populations. The HapMap shows the location of SNPs in the genome and identifies their associated haplotypes. The scientists state that the SNPs and haplotypes that make up the HapMap will make it possible for researchers to conduct genome-wide association studies to identify the genetic contributions to common diseases such as cardiovascular disease, cancer, obesity, diabetes, psychiatric illnesses, and autoimmune diseases. A genome-wide association study, or GWAS, involves scanning markers across the genomes of many people to find genetic variations associated with a particular disease.

There are both direct and indirect ways of conducting genome-wide association studies. In the direct way, researchers ask whether a particular SNP is more common in one group versus the other. However, the cost of looking for associations for millions of SNPs across the entire genome to see which SNPs correlate with a specific disease was, at the time of the HapMap's creation, prohibitively expensive. The HapMap enabled an indirect approach. Researchers can use a set of markers—the tag SNPs that identify unique haplotypes—to detect an association between a particular genomic region and a disease. Once they find an association between a particular haplotype and the disease, they can focus on just that region, searching for the genes that may play a role.

Since 2005, scientists have used the HapMap to conduct several hundred GWAS looking for disease associations. Scientists published the first GWAS using HapMap data in March 2005, identifying a SNP linked to a disease affecting the eye, called age-related macular degeneration. By August 2008, more than 170 GWAS had identified more than 150 genetic loci associated with more than sixty complex diseases and traits. Due to a security breach, the NIH retired the HapMap website in 2016 and it is no longer accessible, although as of 2025, researchers may still order the biobanked samples from Coriell.

Since the completion of the HapMap, NIH scientists have pursued other genome variation studies looking at even more populations and even more genetic variants. Among those are the 1,000 Genomes Project and the *All of Us* Project.

Sources

1. Broad Institute. “HapMap 3.” Broad Institute. https://www.broadinstitute.org/medical-and-population-genetics/hapmap-3 (Accessed July 9, 2024).
2. Brown, T. A. “Chapter 1 The Human Genome.” In *Genomes 2nd Edition*. Oxford: Wiley-Liss, https://www.ncbi.nlm.nih.gov/books/NBK21134/ (Accessed July 9, 2024).
3. Cavalli-Sforza, L. Luca. “The Human Genome Diversity Project: Past, Present, and Future.” *Nature Reviews Genetics* 6 (2005): 333–40.
4. Collins, Francis S., Mark S. Guyer, and Aravinda Chakravarti. "Variations on a Theme: Cataloging Human DNA Sequence Variation." *Science* 278 (1997): 1580–1.
5. Coriell Institute for Medicine. “International HapMap Project.” Coriell Institue for Medicine. https://www.coriell.org/1/NHGRI/Collections/HapMap-Collections/HapMap-Project (Accessed July 9, 2024).
6. Fujimura, Joan H., and Ramya Rajagopalan. "Different differences: The use of ‘genetic ancestry’ versus race in biomedical human genetic research." *Social studies of science* 41, no. 1 (2011): 5-30.
7. Gabriel, Stacey B., Stephen F. Schaffner, Huy Nguyen, Jamie M. Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, Shau Neen Liu-Cordero, Charles Rotimi, Adebowale Adeyemo, Richard Cooper, Ryk Ward, Eric S Lander, Mark J Daly, and David Altshuler "The Structure of Haplotype Blocks in the Human Genome." *Science* 296, no. 5576 (2002): 2225–9.
8. International HapMap 3 Consortium. “Integrating Common and Rare Genetic Variation in Diverse Human Populations.” *Nature* 467 (2010): 52–8.
9. International Human Genome Sequencing Consortium. “Initial Sequencing and Analysis of the Human Genome.” *Nature* 409 (2001): 860–921. https://www.genome.gov/Pages/Newsroom/Webcasts/2010ScienceReportersWorkshop (Accessed July 9, 2024).
10. Ku, Chee Seng, En Yun Loy, Agus Salim, Yugi Pawitan, and Kee Seng Chia. “The Discovery of Human Genetic Variations and Their Use as Disease Markers: Past, Present, and Future.” *Journal of Human Genetics* 55 (2010): 403–15. https://www.nature.com/articles/jhg201055 (Accesed July 9, 2024).
11. Marks, Jonathan. *Human Biodiversity: Genes, Race, and History*. New York City: Walter de Grutyer, Inc, 1995.
12. National Human Genome Research Institute. “Developing a Haplotype Map of the Human Genome for Finding Genes Related to Health and Disease.” National Human Genome Research Institute.

https://www.genome.gov/10001665/developing-a-haplotype-map-of-the-human-genome-to-find-genes-related-to-health-and-disease-meeting-summary (Accessed July 9, 2024).
13. National Human Genome Research Institute. “International Consortium Launches Genetic Variation Mapping Project.” National Human Research Institute. https://www.genome.gov/10005336/2002-release-genetic-variation-mapping-launch (Accessed July 9, 2024).
14. National Human Genome Research Institute. “International HapMap Project.” National Human Genome Research Institute. https://www.genome.gov/10001688/international-hapmap-project (Accessed July 9, 2024).
15. National Human Genome Research Institute. “International HapMap Project – Participants: Sample Collection.” National Human Genome Research Institute. https://www.genome.gov/10005340/hapmap-sample-collection (Accessed July 9, 2024).
16. National Library of Medicine. “NCBI Retiring HapMap Resource.” National Library of Medicine. https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/ (Accessed July 9, 2024).
17. Ossorio, Pilar N. "Race, genetic variation, and the haplotype mapping project." *La. L. Rev.* 66 (2005): 131.
18. Reardon, Jenny. *Race to the Finish: Identity and Governance in an Age of Genomics*. Princeton: Princeton University Press, 2004.
19. Reardon, Jenny. *The Postgenomic Condition: Ethics, Justice & Knowledge after the Genome.* Chicago: Chicago University Press, 2017.
20. Sachidanandam, S., D. Weissman, S. C. Schmidt, J. M. Kakol, L. D. Stein, G. Marth, S. Sherry, C. Mullikin, B. J. Mortimore, D. L. Willey, S. E. Hunt, C. G. Cole, P. C. Coggill, C. M. Rice, Z. Ning, J. Rogers, D. R. Bentley, P. Y. Kwok, E. R. Mardis, R. T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R. H. Waterston, J. D. McPherson, B. Gilman, S. Schaffner, W. J. Van Etten, D. Reich, J. Higgins, M. J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M. C. Zody, L. Linton, E. S. Lander, D. Altshuler; International SNP Map Working Group. “A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms.” *Nature* 409 (2001): 928–33.
21. The International HapMap Consortium. “A Second Generation Human Haplotype Map of over 3.1 Million SNPs.” *Nature* 449 (2007): 851–61. https://www.nature.com/articles/nature06258#MOESM27 (Accessed July 14, 2024).

22. The International HapMap Consortium. “A Haplotype Map of the Human Genome.” *Nature* 437 (2005): 1299–1320. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1880871/ (Accessed July 9, 2024).
23. The International HapMap Consortium. “Integrating Ethics and Science in the International HapMap Project.” *Nature* 5 (2004): 467–75. https://www.nature.com/articles/nrg1351 (Accessed July 9, 2024).
24. The International HapMap Consortium. “The International HapMap Project.” *Nature* 426 (2003): 789–96. https://www.nature.com/articles/nature02168 (Accessed July 9, 2024).
25. Thorisson, Gudmundur A., Albert V. Smith, Lalitha Krishnan, and Lincoln D. Stein. “The International HapMap Project Web Site.” *Genome Research* 15 (2005): 1592–3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1310647/ (Accessed July 9, 2024).

Organization