2018 William Allan Award: Discovering the Genes for Common Disease: From Families to Populations[1]

Eric S. Lander[2,3,4,*]

Thank you, Mark, for that wonderful and warm introduction. I have had the enormous pleasure to be a close friend and colleague of Mark Daly's since 1986, when he joined my lab as an MIT freshman, and to have watched him become one of the leading intellects in human genetics today. So, it is a special privilege to be introduced by Mark on this occasion.

The Allan Award is deeply meaningful to me because, while I have worked on many things, I am first and foremost a human geneticist, and the Allan Award is the highest honor in human genetics. In these brief remarks, there's just enough time to share a bit of personal history and to highlight a few of the amazing people I've been lucky enough to work with over my career. (This is personal history, not a comprehensive review. I'll thus focus on work by my own trainees, close colleagues, and collaborators, noting parallel work in a few places.)

First, a warning to students. Students often ask me how to plan their careers. It turns out I'm a terrible person to ask—my career has depended far more on luck than planning. In fact, my foray into human genetics was a complete accident. I've therefore subtitled my remarks: *How a chance meeting at MIT and a dinner in Finland shaped my entire scientific career.*

A Chance Meeting at MIT

In the spring of 1985, I met David Botstein after the weekly biology colloquium at MIT. In 1980, he had published a landmark paper,[1] proposing how to use DNA polymorphisms to systematically map the genes responsible for simple Mendelian diseases. (Botstein had discovered in 1977 the idea of using DNA sequence polymorphism—specifically, restriction fragment length polymorphisms [RFLPs] detected on Southern blots—as markers to recombinationally map the location of the ribosomal gene cluster in yeast.[2] At a human genetics department retreat in 1978, he realized the approach could be adapted to enable genetic mapping in humans.) The 1980 paper proposed creating a comprehensive human genetic linkage map and using it to perform linkage analysis in families to discover the chromosomal location of human disease genes. Notably, it was the first paper to propose a "human genome project," and it explicitly laid out how to do it.

When a colleague introduced me to Botstein after the seminar, he was in a ruminating mood. He was frustrated that his mapping paradigm applied only to monogenic Mendelian disorders, but not to common human diseases (for example, heart disease, diabetes, or schizophrenia), which don't follow the simple rules of Mendelian inheritance. He'd been looking for someone who spoke both mathematics and genetics to discuss the problem with. When he heard that I was a mathematician who had been learning genetics (moonlighting for the prior few years in the labs of fly geneticists Peter Cherbas and Bill Gelbart at Harvard and nematode geneticist Bob Horvitz at MIT, while teaching at Harvard), he pounced—lobbing a bombastic pronouncement: "There's no way to map the genes for common human diseases, due to locus heterogeneity and polygenicity." David is from the Bronx; I'm from Brooklyn. So, we immediately fell into the form

[1]This article is based on the address given by the author at the meeting of the American Society of Human Genetics (ASHG) on October 18, 2018, in San Diego, California. The audio of the original address can be found at the ASHG website

[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA; [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; [4]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
*Correspondence: lander@broadinstitute.org

https://doi.org/10.1016/j.ajhg.2019.01.016.

of discourse employed by natives of those boroughs: we began arguing. I said, "Nonsense!" (or rather stronger words, as I recall) and insisted that it should be possible. We had such a good time arguing together that we agreed to meet again the next day to argue some more.

That first conversation with Botstein provoked me to think deeply about how to tackle polygenic disorders. We worked together madly for a couple of months on various ideas for going beyond monogenic traits in families. As it turns out, David was under a bit of time pressure. Based on his famous paper in 1980, he had been invited as a keynote speaker at the Human Genome Mapping 5 meeting in Helsinki in July 1985. But, he hadn't published anything more in human genetics since the 1980 paper. So, David was very pleased to have new results to describe. He invited me to come with him to Helsinki.

Dinner in Helsinki

In Helsinki, I had what was surely the most influential dinner of my scientific career. At the conference dinner at the restaurant *Kappeli*, on the elegant esplanade in central Helsinki, we sat opposite Albert de la Chapelle, one of the two great Finnish geneticists at the time (the other being Leena Peltonen, who I later had the pleasure to come to know well). Albert regaled us with the story of Finnish population history, of which at the time I knew nothing. He described how modern Finland traced back to a small founding population about 2,000 years ago. As a result, certain "Finnish genetic diseases" had higher frequency in Finland due to founder effects resulting from the population bottleneck. Captivated by this, I went back that evening and started thinking about what it meant for genetic mapping. Of course, what it meant was that if a single founder had introduced a single copy of a disease allele, you should able to map the region carrying the disease allele—without needing to trace meioses in families—because the vast majority of affected individuals would carry an ancestral segment (or haplotype) around the locus. With a sufficiently dense genetic map, the ancestral haplotype should stand out by virtue of being "associated" with the disease—that is, overrepresented in disease cases. In short, while family-based linkage mapping was wonderful, you might not need families at all. At least under some circumstances, linkage mapping in families might be replaced by linkage-disequilibrium mapping in entire populations.

When I returned from Finland, I started working on the question of how feasible it would be in practice to detect the presence of a common ancestral haplotype that arose a hundred generations ago. The approach turned out to be eminently feasible—provided one had a sufficiently dense genetic map.

Methods for Mapping Complex Traits

I gave a talk about these ideas in the opening session of the annual Cold Spring Harbor Symposium in June 1986, and Botstein and I wrote up a paper entitled *Mapping Complex Genetic Traits in Humans: New Methods Using a Complete RFLP Linkage Map*.[3] We then set out to apply the ideas in practice. With Johanna Hastbacka, who was a graduate student with de la Chapelle and then a postdoctoral fellow with me, we set out to localize the gene for a Finnish disease, diastrophic dysplasia (DTD). As predicted, we were able to use linkage disequilibrium mapping in the Finnish population to pinpoint the DTD gene to within a few tens of kilobases and subsequently to clone the gene.[4,5] (At the time, it was only the second gene cloned without the benefit of a frank chromosomal deletion.) In the ensuing years, an impressive cadre of Finnish trainees, working with de la Chapelle and Peltonen, proceeded to use linkage disequilibrium mapping to clone the genes for all of the Finnish genetic diseases.

At the same time, again provoked by that first conversation with Botstein, I became interested in polygenic traits—phenotypes shaped by the combined action of many genes. Back then, the idea of studying polygenic traits in humans was crazy, but I thought it would be possible in experimental organisms. With Botstein, we developed the framework for mapping quantitative trait loci (QTLs) in crosses.[6] While we would later apply the approach to map disease modifiers in mouse, plants offered the best place to test the approach. So, although I grew up in New York City, I spent time tromping through corn fields and talking with agricultural geneticists studying maize, tomatoes, and other crops. With Steve Tanksley at Cornell, we set out to create a genetic map of the tomato and to use it to map polygenic factors contributing to such traits as fruit size, pH, and soluble-solids content. Remarkably, it worked: we found multiple genes controlling each trait.[7]

In principle, one might combine these ideas—linkage-disequilibrium mapping and polygenic trait mapping—to tackle *any* complex human disease in *any* human population. But to do this, one would need extremely dense genetic maps—or, as we dubbed them, "infinitely dense genetic maps."

Initial Human Genetic Map

At the time, however, it was challenging to build even a rudimentary human genetic map. Early linkage mapping studies, such as Jim Gusella's mapping of Huntington disease,[8] used random unmapped polymorphisms in the hope of getting lucky. But, Botstein insisted on the importance of being systematic. He had co-founded a biotech company, Collaborative Genetics, and persuaded it to launch a genetic mapping group in 1984 to create a genome-wide map. Directed by Helen Donis-Keller and advised by Botstein, the group undertook the first human genome mapping project—hybridizing thousands of DNA fragments to Southern blots to find those that revealed RFLPs across a set of individuals and then tracing the inheritance of these variants in large pedigrees. Ray White, one of Botstein's collaborators on his 1980 paper, also launched an RFLP mapping effort at the University of Utah. (Most RFLPs are caused by single-nucleotide

polymorphism [SNPs] in the human genome, which today can be genotyped much more easily.)

Turning these data into a genetic map required overcoming one key hurdle: the available algorithms for human genetic mapping could study only a few genetic markers at a time (because the computation grew exponentially with the number of loci). New multi-locus mapping approaches were needed. I took up the challenge, together with Phil Green, and developed an EM-based algorithm that scaled linearly in the number of markers.[9] I recruited a few amazing MIT undergraduates (including the remarkable Mark Daly) to help create a software package and then to analyze the data. In October 1987, the teams led by Donis-Keller, Botstein, and me published a human genetic map with 403 DNA polymorphisms spanning the human chromosomes with detectable linkage between consecutive loci.[10]

Toward an Infinitely Dense Genetic Map

A vastly stronger foundation would be required, however, to fulfill the promise of human genomics—including achieving the power of an "infinitely dense" genetic map to discover genes for common diseases by population-based association with variants and haplotypes. So, out of sheer scientific necessity, I decided to devote much of the next phase of my life to the Human Genome Project (HGP). As it happens, the 1986 Cold Spring Harbor symposium at which I'd given a talk about mapping complex traits was also the occasion of the first public debate about the idea of a human genome project. I spoke twice during the debate and, perhaps as a result of these interventions, was invited to participate in subsequent discussions about the Human Genome Project at the National Institutes of Health and the National Academy of Sciences.

The Human Genome Project was launched in October 1990, with my group serving as one of four initial US genome centers and eventually becoming one of the leading contributors to the project. There's so much I could say about the Human Genome Project, but this is not the time or place. Suffice it to say that, by discovering and making freely available the near-complete sequence of the human genome,[11,12] the project laid a foundation for biomedicine in the 21st century. I also want to say a huge thank you to the many amazing people with whom I got to work. In particular, I want to call out the late John Sulston, Bob Waterston, and Francis Collins. The opportunity to work with these extraordinary scientists on a mission of true public importance and public good was one of the greatest privileges of my life.

Returning to the Genetics of Common Diseases

About five years into the Human Genome Project, I began to have enough breathing room to return to the idea of discovering the genes underlying common diseases by using an infinitely dense genetic map to detect ancestral haplotypes. In 1996, I sketched out the idea of genome-wide association studies, based on comprehensive population-genetic information about both coding variants and linkage disequilibrium mapping across the genome.[13] (At the same time, Neil Risch and Kathleen Merikangas published an important statistical analysis comparing association analysis and linkage analysis.[14]) Bringing these ideas to feasibility would take tremendous work over the next decade.

A number of brilliant young scientists contributed to laying the foundation for the genetics of common disease. I want to cite five in particular who played central roles and would grow into leading human geneticists. David Altshuler, a physician-scientist specializing in endocrinology, had joined the lab with the aim of understanding the genetics of type 2 diabetes, but soon realized that the scientific foundations were missing and committed himself to building them. Mark Daly had remained in the lab after graduating from MIT and was steadily building his remarkable career as the leading analyst of human genetic information. Stacey Gabriel brought extensive expertise in human genetics from her doctoral studies on Hirschsprung disease with Aravinda Chakravarti and would later grow to become the director of the Broad Institute's Genomics Platform, playing a role in countless genomics projects. David Reich brought a deep interest in human populations, which he would later apply to studies of Neanderthal and ancient *Homo sapiens*. Joel Hirschhorn, another endocrinologist, brought a distinguished personal pedigree in the field (both his mother and father were distinguished medical geneticists[15]); he would become a leader in the genetics of height and other anthropometric traits.

Together, we focused on five issues.

Common Diseases and Common Variants

The first issue concerned the genetic architecture of common complex diseases. For rare Mendelian diseases, the genetic architecture was straightforward: because these diseases are highly deleterious, disease alleles are rapidly eliminated by purifying selection—typically resulting in an allelic spectrum with a large number of rare alleles. These diseases are perfectly suited for family-based linkage mapping, because any given family will be segregating only a single highly penetrant disease-causing allele. Botstein's paradigm was wildly successful: by the late 1990s, more than 1,000 rare Mendelian diseases had been mapped by linkage analysis and many had been molecularly cloned.

By contrast, common diseases seemed likely to be very different. Reich and I realized that the genetic architecture was likely to be strongly influenced by the unusual history of the human population. Specifically, it was known (based on the extent of genetic polymorphism in the population) that humans had had a small effective population size ($\sim 10^4$) for a long time, before undergoing a rapid exponential expansion. By studying the speed at which the allelic spectrum underlying phenotypes equilibrated to the larger population size, we realized that common variants should play a major role in most common phenotypes. We called this prediction the common disease-common variant

(CDCV) hypothesis—and it propelled much of our subsequent work.[16] The CDCV hypothesis was consistent with the observation that applying the linkage-mapping approaches used for Mendelian diseases (whether in large families or large collections of subpairs) had yielded few results. Moreover, at these few loci, the disease-related alleles had much higher frequency (e.g., $>10^{-2}$ versus $<10^{-4}$) and increased risk by much less (e.g., 3- to 5-fold versus several hundred-fold) than seen for Mendelian diseases. Despite the theoretical and experimental evidence, the CDCV hypothesis was controversial at the time (some favored the idea that common disease would be largely due to rare variants[17]) but would come to be confirmed by copious data and is now widely accepted (see below).

Rigor and Power

The second issue concerned the ramshackle state of association studies. Association studies, in which investigators tested their favorite candidate genes for allelic association with diseases, were plagued by irreproducibility. Most papers were based on small sample sizes with little power to detect true associations, and they were largely filled with false associations because they often ignored multiple hypothesis testing and publication bias—reporting results that reached nominal significance ($p = 0.05$) and ignoring the rest.

I'd had a long-standing interest in understanding the appropriate standards for genome-wide significance, dating back to our work on linkage mapping of both quantitative and disease traits.[6,18] With the aim of bringing rigor to association studies, Altshuler and Hirschhorn tested 16 previously published associations related to type 2 diabetes.[19] Using large sample sizes and special designs to control for effects of population stratification, they found that only one of the published associations was reproducible: a p.Pro12Ala variant in PPARG. With 3,000 samples, the variant had a clear if modest effect, increasing risk by 25%. Curiously, the association had been reported in a first paper and then seemingly refuted by four later papers, which each found no association. Altshuler and Hirschhorn found that pooling data from those four (underpowered) "refutations" yielded a "confirmation"—that is, the combined data made the association significant. Their paper made clear that large sample sizes and rigorous thresholds would be essential going forward.

Cataloging Human Genetic Variation

The third issue concerned how to generate a near-complete catalog of common human genetic variation (e.g., down initially to 1% and ultimately to much lower frequencies) and how to genotype huge numbers of polymorphisms simultaneously. The genetic maps being created by the Human Genome Project focused on a specialized set of highly polymorphic loci (tandem repeats that could be genotyped by performing PCR and measuring the sizes of the resulting fragments); the effort had expanded the initial human genetic map by an order of magnitude, to include ~5,000 genetic markers. While adequate for mapping simple Mendelian diseases, the approach would never generate the "infinitely dense" genetic map that we required for the genetic analysis of common diseases.

The vast majority of genetic variants carried in a human genome are SNPs and small insertion-deletion polymorphisms. Altshuler and I focused on ways to discover such variants at massive scale across the genome. To jumpstart the approach even before we had a human reference sequence, we developed various experimental and computational approaches to analyzing shotgun sequence, including focusing on "reduced representations" of the genome.[20] We reached out to our colleagues at the Sanger Centre in Cambridge, England and Washington University in St. Louis to create a collaborative effort, called The SNP Consortium, aimed at discovering human genetic variation. By the time of the publication of the Human Genome Project's draft sequence in 2001, the SNP Consortium published a companion paper reporting 1.42 million common SNPs across the genome.[21] (Since then, follow-on projects have yielded more than 26.7 million common variants at frequency above 1% in one or more global populations, according to the GnomAD database [see Web Resources]. In European populations alone, there are more than 15.5 million SNPs with frequency variants at frequencies above 0.1%.)

Large-Scale Genotyping

The fourth issue concerned how to genotype an "infinitely dense" genetic map. Since the mid-1980s, genetic markers had been assayed either one-at-a-time (RFLPs) or a few-at-a-time (combining PCR-based length polymorphisms on sequencing gels). Instead, we needed to develop approaches for massively parallel genotyping. For this, we worked with Affymetrix, a California-based start-up, to develop genotyping arrays capable of assaying many SNPs in parallel.[22] Over time, this approach made it possible to simultaneously genotype 100,000 and even 1 million SNPs in a human sample.

Human Haplotype Structure

The fifth issue concerned the haplotype structure of the human genome. In 1986, we had described the approach of linkage disequilibrium mapping in Finland. Given Finland's recent founding (~100 generations) and small bottleneck (~100 chromosomes), the ancestral haplotypes carrying any disease-predisposing allele were very large (typically, >1 megabase). Linkage disequilibrium mapping would, in principle, apply to mapping *any* alleles that affected disease risk in *any* population—provided that one had a sufficiently dense genetic map to reliably detect the ancestral haplotypes in the population. Little was known about the overall haplotype structure across the human genome in large continental populations. There had been pioneering studies of linkage disequilibrium within individual disease genes going back to the 1980s,[23] but the scale and resolution were too limited.

Mark Daly cracked open this problem in a seminal paper in 2001.[24] With the aim of pinpointing a gene in chromosome 5q21 that affected risk for inflammatory bowel disease in European-ancestry individuals, we had undertaken extremely dense genotyping of the region, with variants at an average spacing of 5 kilobases across 0.5 megabases. By carefully analyzing the data, Daly discerned a pattern: clear ancestral haplotypes separated by recombinational hotspots. It was clear that linkage disequilibrium spanned many tens of kilobases, even in the general European population. Although Daly had analyzed only a single locus, we were sure the pattern was general and entitled the paper *High Resolution Haplotype Structure in the Human Genome*. Our confidence turned out to be justified. Stacey Gabriel published a compelling study the following year reporting detailed analysis of 51 regions of the human genome in samples from Africa, Europe, and Asia; the results confirmed the general properties of human haplotypes and the extent of linkage disequilibrium.[25]

Why was the haplotype structure of the human population so important? Because it held the key to creating the "infinitely dense" genetic map needed to locate the common variants underlying common disease. With knowledge of the haplotype structure of the human genome, one could use the genotypes for a *subset* of genetic variants in an individual to *impute* the genotypes for essentially all other genetic variants, based on their local correlation structure.

Recognizing the power of the approach, Daly's paper proposed the goal of "creating a comprehensive haplotype map of the human genome." Within a year, an International Haplotype Map Project was launched—resulting in detailed maps defining the correlation of genetic variants across the genome. Today, for example, one can use information for 500,000 genetic variants to reliably impute genotypes for ~15 million common variants, down to frequencies of ~0.1% in European-derived samples.

Positive Selection in Recent Human History

I also want to mention a slightly different but related topic. Around this time, another brilliant young scientist in the lab, Pardis Sabeti, had a powerful insight about how to use ancestral haplotypes to detect the fingerprints of positive selection in recent human evolution. She reasoned that genetic variants that are strongly advantageous will rise rapidly to high frequency, whereas other genetic variants that happen to reach high frequency will do so only by slow genetic drift. A rapid ascent will thus be correlated with a long haplotype around the variant, because there will have been too little time for genetic recombination to "break down" the haplotype. Thus, high-frequency long-range haplotypes should be smoking guns for positive selection.[26] Using this approach, Pardis went on to discover hundreds of regions in the human genome that have been under strong positive selection and, in a number of cases, to discover the underlying genes and likely selective advantages—for example, related to resistance to infectious disease.[27,28] (Driven by her interest in some of these loci, Sabeti established deep collaborations in West Africa and played an important role in the scientific response to the Ebola outbreak in 2014. She's become an important force in the genomics of outbreak response.)

GWAS/CVAS

The late 1990s and early 2000s was an exhilarating time of ferment in human genetics—filled with remarkable colleagues creating ideas, experimental approaches, comprehensive catalogs, and computational methods to enable a systematic attack on understanding the genetic basis of human disease. Importantly, the science moved forward through international collaborations premised on open and rapid data sharing—in the spirit that had been established by the Human Genome Project.

The result of the ideas, data, and tools was a robust methodology for discovering common genetic variants underlying common disease via linkage disequilibrium mapping in populations. The methodology is generally referred to as genome-wide association studies (GWASs)—although I prefer the terminology common variant association studies (CVASs), to distinguish it from rare variant association studies (RVASs), which is the other important type of genome-wide association study (see below).

Within a few years, the efforts began to bear fruit. As noted above, the genetic dissection of complex diseases had been proceeding at a snail's pace—with the total number of loci discovered being roughly one per year across all common diseases combined through 2004. The pace then picked up—with four discoveries in 2005, eight in 2006, and more than 65 in 2007. The floodgates had burst, and the total kept going up and up.

Missing Heritability

But all was not happiness. Some observers were not impressed by the discoveries of common variants associated with common diseases. They voiced a litany of valid concerns: the early results had yielded only a handful of loci for any given disease, which tended to have only modest effects on disease risk, to occur largely in non-coding regions, and to explain only a small portion of the heritability. This led to the famous "crisis of missing heritability"[29] with various competing explanations—including that the common-variant discoveries might be artifacts and that the rare variants with large effects might be the true drivers of common disease genetics.

It took some years for the answer to become clear: the primary issue was insufficient sample size to detect many of the common-variant loci. Schizophrenia provides a good illustration, based on work by Mark Daly, Ben Neale, and their many colleagues in the Psychiatric Genomics Consortium. An initial study, led by my colleagues Shaun Purcell and Pamela Sklar (who sadly passed away from cancer in 2017), involving ~6,000 people found *no* loci that reached genome-wide significance. The result seemed discouraging but, when the investigators closely examined

the data, they realized it revealed many more loci *approaching* significance than expected by chance—suggesting that robust discoveries would emerge with larger sample sizes.[30] Sure enough, the number of genome-wide discoveries reached 5 loci with ~20,000 samples, 62 loci with ~50,000 samples, 108 loci with ~110,000 samples, and by today 245 loci with ~163,000 samples analyzed.[31] Similar results were found for other common diseases, including type 2 diabetes, inflammatory bowel disease, and coronary artery disease with 403, 273, and 166 loci, respectively, so far. Across all diseases and traits, the total number of genome-wide significant discoveries reported in the literature currently exceeds 30,000 (see Web Resources). With unpublished reports based on sources such as the UK BioBank, the current total likely approaches 100,000 loci.

Recognizing that there are many additional loci for each disease that currently fall below the threshold for genome-wide significance, Peter Visscher and colleagues developed an elegant statistical analysis that makes it possible to estimate their contribution to heritability.[32] Current estimates suggest that perhaps half of the heritability is attributable to the additive effects of these common variants. It's likely that genetic interactions (that is, non-additive effects) explain a significant additional portion of heritability, but current methods lack statistical power to detect such interactions.[33]

Role of Rare Variants

Of course, the success with CVASs does not mean that rare variants play no role in common disease. Even if they explain only a minority of the heritability, rare variants with large effects—especially those in coding regions—are very informative biologically. RVAS efforts have thus been growing in recent years.

For common disorders associated with extremely strong loss of reproductive fitness (such as severe developmental disorders, severe intellectual disability, and autism with low IQ), RVAS has been quite successful—identifying disruptive coding mutations in many different genes.[34] Indeed, these disorders might be appropriately considered as sets of many nearly-Mendelian disorders in which the rare variants dramatically increase disease risk.

For most common diseases, however, the jury is still out with respect to the role of rare variants with large effects. Despite much optimism, early RVAS efforts with several thousands of samples found few loci. With a few RVAS reaching larger samples sizes, a handful of genome-wide significant loci have begun to emerge. Just as for early CVAS, these studies are still likely to be underpowered. We're beginning to get a clear picture of the necessary sample sizes based on (1) careful power analysis (carried out by Or Zuk, when he was a postdoctoral fellow in the lab) that calculates the number of cases based on the combined frequency of loss-of-function (LoF) variants in a gene and their effect on disease risk[35] and (2) empirical measurements of the frequency of LoF alleles for each gene, from the Exome Aggregation Consortium (ExAC) launched by Daniel MacArthur.[36] Given the median frequency of LoF alleles across human genes (~1.5 LoFs per 10,000 chromosomes), well-powered RVASs will ideally need 100,000 or more cases to detect most genes in which rare variants substantially increase disease risk. Fortunately, such sample sizes will be feasible in the years ahead. In the meantime, smaller sample sizes can provide insights into the overall contribution of rare variants, based on the aggregate burden of LoF alleles.

Biological Insights and Clinical Utility

With robust paradigms for mapping genes underlying complex traits now established, the key issue has increasingly become how to translate the results into biological insights and clinical utility.

There's already been considerable progress by many groups around the world. I'll give a few examples from my colleagues in the Broad Institute community. For inflammatory bowel disease, Ramnik Xavier, Mark Daly, and their collaborators have used the genetic mapping results to implicate many new pathways—including a crucial role of autophagy, which had not previously been connected to the disease.[37] This work has directly led to animal studies, human physiological studies, and several drug development projects by pharmaceutical companies. For schizophrenia, Steve McCarroll, Beth Stevens, and their collaborators showed that the strongest CVAS signal in the genome is caused by variation in the C4 gene that modulates the pruning of synapses by microglia—implicating the brain's immune-like cells in the etiology of the disease.[38] Interestingly, CVASs in Alzheimer disease also point to a key role for microglial functions.[39] For cardiovascular disease, Sekar Kathiresan has elucidated multiple mechanisms with implications for drug development. He's also applied the elegant approach of Mendelian randomization to the genes associated with lipid levels to show that the epidemiological observation that high HDL levels are associated with protection from heart attack is likely not to be a causal effect, but rather a correlation; instead, the causal culprit appears to be triglyceride levels, which are inversely associated with HDL levels.[40]

In the past year, it has also become clear that the information from massive CVASs—with millions of SNPs assayed or imputed in hundreds of thousands of people—can yield clinically meaningful risk predictors. As recently reported by Kathiresan and colleagues, "genome-wide polygenic scores" can already identify significant fractions of the population (in the range of 5%) at substantially higher risk for a number of diseases, including heart disease, atrial fibrillation, inflammatory bowel disease, and breast cancer.[41] These polygenic risk scores will increasingly be used in clinical practice, especially where relevant interventions or screening are available, and also in clinical trials, to substantially increase statistical power and decrease sample size.

The Next Phase of Common-Disease Genetics

Yet, even with all this progress, we need a new revolution in the genetics of common disease. With the number of

robust genetic associations with human diseases and traits pushing 100,000, the current process of moving from genetic variants to biological function is too slow—often requiring heroic efforts to connect an individual locus with molecular mechanisms and disease physiology.

The first revolution—which enabled the mapping of genes for common diseases—required clear scientific vision, careful planning, and tremendous collaboration. The next revolution—to dramatically accelerate progress from maps to mechanisms—will require similar efforts. There will be at least three components.

Using the Full Power of Human Populations

The recent explosion in human genetic studies—through case-control studies for specific diseases, national biobanks with diverse phenotypes in longitudinal cohorts, and large health systems with electronic medical records—will only accelerate as the cost of whole-genome sequencing falls to $100. In years ahead, the total number of participants with extensive genotypic and phenotypic information will surely exceed 50 million.

Such information offers tremendous power. Larger CVASs will enable genetic fine-mapping to identify causal variants within disease-associated loci, as well as discovery of many new loci for many traits. In addition, RVASs will reach the scale required to identify genes carrying rare coding variants with large effect. Most importantly, the ability to perform phenome-wide association studies (PheWASs)—that is, to annotate individual variants and sets of variants according to their association with a vast range of phenotypes—will help elucidate their biological effects. For example, a common amino acid substitution in SLC39A8, a zinc and manganese transporter, is associated with higher risk of schizophrenia, Crohn disease, and obesity—but protection from hypertension and Parkinson disease. Eventually, we want the complete "genotype-by-phenotype matrix" for the human population.

Achieving this vision, however, will require overcoming many important challenges in data sharing (including cloud-based platforms for effective storage and analysis, methods for joint analysis that protect personal information, incentives for data sharing, and compliance with national laws) and global equity (ensuring that non-European-ancestry populations are well represented).

Systematically Connecting Variants to Function

We need to develop rapid and reliable ways to connect disease-associated non-coding variants with (1) the genes they regulate, (2) the cell types in which they act, and (3) the molecular processes they perturb to increase disease risk. This will require generating comprehensive information resources and new analytical methods. Various approaches are emerging, with most building on the foundation being laid by the recently launched Human Cell Atlas project, led by my colleague Aviv Regev and Sarah Teichmann at the Sanger Institute.[42]

One class of approach uses "bulk genetic signals" to understand the cell types and molecular processes involved in a disease, by analyzing the genome-wide pattern of SNP associations for one or more common diseases. Hilary Finucane and her colleagues have pioneered ways to connect a phenotype with specific organs, tissues, or cell types, by assessing the genome-wide correlation between the association statistics and gene expression levels.[43] It will be exciting to see how far this approach can go as the Human Cell Atlas delivers comprehensive information on all human cell types. In principle, similar approaches could be used to connect phenotypes with specific molecular processes, by correlating association statistics with a complete catalog of signatures of all cellular processes, which might be derived from the Human Cell Atlas. Separately, one might also exploit polygenic risk scores to reveal molecular processes that play a causal role in a disease, by comparing gene expression patterns in relevant cell types in *pre-symptomatic* individuals at opposite ends of the polygenic risk spectrum. (After disease onset, it is much harder to distinguish whether differences reflect cause or effect.)

Another class of approach focuses on understanding individual disease-associated loci. Ultimately, we want to be able to predict the immediate molecular consequence of each of the millions of common variants in the human population—for example, for non-coding variants, how they affect the binding of regulators and the expression of target genes in all cell types.

While the goal seems daunting, a wide variety of creative ideas are being pursued. These include: systematic perturbation studies of enhancer-promoter connections (by Jesse Engreitz, a former trainee) that suggest that reasonably good predictions may be derived from cell type-specific information about chromatin state and three-dimensional folding; ways to generate a comprehensive catalog of all common expression QTLs (eQTLs) based on human cell atlases (that is, comprehensive single-cell RNA sequencing) from several hundred people; machine learning to predict the regulatory consequence of variants from diverse data about molecular readouts, evolutionary conservation and clinical consequences; and much more efficient ways to perform and systematically assess the effects of isogenic genome editing of disease-associated variants.

Devising Disease-Specific Approaches to Identify Mechanisms

While large-scale human genetic studies and systematic functional catalogs will be essential, understanding the genetics of each common disease and developing therapeutic hypotheses will require disease-specific approaches. These approaches will include assays for specific cellular processes and appropriate model systems, such as human biopsies, organoids, cell lines, and genetically engineered animals. Although the specifics will vary across disease, there is great potential to develop efficient general methods to develop such assays and models.

Conclusion

Human genetics continues to advance at a stunning pace. Goals go from inconceivable to routine within a decade. As the genetics of common disease comes into sharper focus,

it is increasingly informing biological mechanism and therapeutic development. I am more excited about the field today than ever.

From my earliest discussions with David Botstein to today, I have had the enormous pleasure of getting to work with truly amazing scientists—extraordinary trainees in my lab, colleagues at the Broad Institute, partners in the international consortia, and collaborators around the globe. I could not have been luckier.

Web Resources

gnomAD, http://gnomad.broadinstitute.org
GWAS Catalog, https://www.ebi.ac.uk/gwas/

References

1. Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. *32*, 314–331.
2. Petes, T.D., and Botstein, D. (1977). Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. Proc. Natl. Acad. Sci. USA *74*, 5091–5095.
3. Lander, E.S., and Botstein, D. (1986). Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. Cold Spring Harb. Symp. Quant. Biol. *51*, 49–62.
4. Hästbacka, J., de la Chapelle, A., Kaitila, I., Sistonen, P., Weaver, A., and Lander, E. (1992). Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat. Genet. *2*, 204–211.
5. Hästbacka, J., de la Chapelle, A., Mahtani, M.M., Clines, G., Reeve-Daly, M.P., Daly, M., Hamilton, B.A., Kusumi, K., Trivedi, B., Weaver, A., et al. (1994). The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell *78*, 1073–1087.
6. Lander, E.S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics *121*, 185–199.
7. Paterson, A.H., Lander, E.S., Hewitt, J.D., Peterson, S., Lincoln, S.E., and Tanksley, S.D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature *335*, 721–726.
8. Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234–238.
9. Lander, E.S., and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA *84*, 2363–2367.
10. Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., et al. (1987). A genetic linkage map of the human genome. Cell *51*, 319–337.
11. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.
12. International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.
13. Lander, E.S. (1996). The new genomics: global views of biology. Science *274*, 536–539.
14. Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.
15. Hirschhorn, K., Hirschhorn, R., and Hirschhorn, J.N. (2017). A Conversation with Kurt and Rochelle Hirschhorn. Annu. Rev. Genomics Hum. Genet. *18*, 31–44.
16. Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.
17. Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.
18. Lander, E., and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. *11*, 241–247.
19. Altshuler, D., Hirschhorn, J.N., Klannemark, M., Lindgren, C.M., Vohl, M.-C., Nemesh, J., Lane, C.R., Schaffner, S.F., Bolk, S., Brewer, C., et al. (2000). The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. Nat. Genet. *26*, 76–80.
20. Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., and Lander, E.S. (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature *407*, 513–516.
21. Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al.; International SNP Map Working Group (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.
22. Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., et al. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science *280*, 1077–1082.
23. Orkin, S.H., Kazazian, H.H., Jr., Antonarakis, S.E., Goff, S.C., Boehm, C.D., Sexton, J.P., Waber, P.G., and Giardina, P.J.V. (1982). Linkage of β-thalassaemia mutations and β-globin gene polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature *296*, 627–631.
24. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., and Lander, E.S. (2001). High-resolution haplotype structure in the human genome. Nat. Genet. *29*, 229–232.
25. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225–2229.
26. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.
27. Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.
28. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007).

Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.
29. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
30. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., Sklar, P.; and International Schizophrenia Consortium (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.
31. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.
32. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
33. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.
34. Fitzgerald, T.W., Gerety, S.S., Jones, W.D., van Kogelenberg, M., King, D.A., McRae, J., Morley, K.I., Parthiban, V., Al-Turki, S., Ambridge, K., et al.; Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature *519*, 223–228.
35. Zuk, O., Schaffner, S.F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M.J., Neale, B.M., Sunyaev, S.R., and Lander, E.S. (2014). Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA *111*, E455–E464.
36. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al.; Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
37. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W., et al. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat. Genet. *39*, 596–604.
38. Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183.
39. Efthymiou, A.G., and Goate, A.M. (2017). Late onset Alzheimer's disease genetics implicates microglial pathways in disease risk. Mol. Neurodegener. *12*, 43.
40. Voight, B.F., Peloso, G.M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M.K., Hindy, G., Hólm, H., Ding, E.L., Johnson, T., et al. (2012). Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet *380*, 572–580.
41. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
42. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al.; Human Cell Atlas Meeting Participants (2017). The Human Cell Atlas. eLife *6*, e27041.
43. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C., Shoresh, N., et al.; Brainstorm Consortium (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat. Genet. *50*, 621–629.