Medical and Population Genetics at The Broad Institute

INTRODUCTION

Our DNA is a vast repository of biological information. Written in a deceptively simple code of four chemical "letters," it contains instructions for a stunning array of molecular workhorses and capabilities that power the human body — for instance, how to make hemoglobin, the oxygen-carrying protein in the blood; what color our eyes and hair are; and how to digest cow's milk, an ability that dramatically shaped early human history. The basis of all of these properties and many others lies embedded within the billions of DNA letters that make up each individual's genome.

Our DNA also holds valuable clues about the root causes of many diseases. What makes one person develop heart disease while another person remains healthy? What is the basis for a young child's devastating and mysterious illness? Answers to many of these questions lurk within the genome, and unearthing them is important not just for understanding personal health, but also for propelling the scientific community in its relentless effort to understand the biological causes of diseases — fueling the development of new, more effective diagnostic tests and therapies that can benefit people around the world.

Broadly speaking, these are the goals of medical and population genetics — the study of genes that influence disease and how they vary within a

population. The Broad Institute of MIT and Harvard has a longstanding interest and commitment to pursuing these goals, and since its founding over a decade ago, it has been an international leader in the study of human genetic variation and its application to disease. Broad scientists have pioneered the development of diverse tools and methods to enable the large-scale, systematic studies necessary to reveal the genetic causes of common diseases; organized and led sweeping international efforts to probe the genetics of many common conditions, including metabolic, heart, inflammatory, psychiatric, and other diseases; and identified hundreds of genetic risk factors for common conditions including type 2 diabetes, heart disease, high cholesterol levels, inflammatory bowel disease, rheumatoid arthritis, and multiple sclerosis.

THE BROAD'S PROGRAM IN MEDICAL AND POPULATION GENETICS

A major hub of the Broad Institute's efforts to unravel the genetics of disease is its Program in Medical and Population Genetics (MPG). The MPG program, one of the largest of the institute's programs, brings together a diverse scientific community drawn from MIT, Harvard, and the Harvard-affiliated hospitals. Experts in population genetics, statistics, bioinformatics, molecular biology, genomics, disease biology, and medicine collaborate to discover the genes underlying a wide range of diseases and decipher how these genes influence the biological processes that either keep us healthy or make us prone to illness.

Historically, MPG has focused largely on understanding common, "complex" diseases — such as type 2 diabetes, heart disease, and inflammatory bowel disease — that stem from the actions of many genes,

sometimes hundreds or more, each of which may affect a person's risk of disease only slightly. Efforts to study complex diseases continue today, but MPG scientists also are deeply interested in understanding rare, so-called "Mendelian" diseases that are caused by mutations in single genes that prevent them from functioning properly. Studies of such rare diseases can yield insights that help unlock the basis of more common disorders — and vice versa — underscoring the power of this combined approach to illuminate key aspects of human biology.

Regardless of how common or rare a disease is, researchers within MPG are committed to pursuing work that will ultimately have a significant impact on patients. "There's really an arc spanning basic research — which gives us the genes and the genetic architecture underlying these different diseases — all the way through to translation," says Daniel MacArthur, associate director of the MPG Program. MacArthur is also an assistant professor at Harvard Medical School and a group leader within the Analytic and Translational Genetics Unit (ATGU) at Massachusetts General Hospital.

This bench-to-bedside framework is echoed in the diverse community of researchers that comprises the MPG program. The MPG community includes Broad Institute scientists and staff, as well as faculty, postdoctoral researchers, and graduate students from the Broad's partner universities and hospitals. Working closely with other groups across the Broad as well as with many other researchers in the Harvard/MIT community, the Harvard hospitals, and beyond, program members share data and ideas freely, and launch collaborative projects to tackle key challenges in human genetics together.

THE MPG TOOLKIT

Over the past decade, as the cost of DNA sequencing has dropped a million-fold, the amount of genomic information collected from patients has grown (and

> "There's really an arc spanning basic research — which gives us the genes and the genetic architecture underlying these different diseases — all the way through to translation."
>
> — Daniel MacArthur, Ph.D.

continues to grow) exponentially. That means many biological problems that once seemed impossible or impractical to solve — typically because the necessary data could not be generated on a large scale — are now tractable. MPG program researchers now rely heavily on DNA sequencing as a research tool for understanding disease, including whole-genome sequencing (which reads every letter of DNA in the genome) and whole-exome sequencing (which reads only the small but critical portion of the genome that contains instructions for making proteins, known collectively as the "exome").

The power of DNA sequencing-based approaches lies in their ability to reveal the full spectrum of variation that exists within an individual's genome — from the most common variants, shared with a majority of the human population, to the rarest DNA changes, carried by just one person. But this power comes with a cost. Because genome-scale sequencing generates such a vast amount of information, it opens the door for false-positive findings as well as real ones — and therefore, extreme care must be taken to properly sort the genomic wheat from the chaff.

Broad MPG researchers recognized, perhaps sooner than any others, the level of rigor required to achieve robust results from genomic studies. That includes the

numbers of samples often needed for analysis (from tens of thousands to 100,000 or more) as well as the proper study design and application of innovative statistical methods to the resulting data. Many of the field's cutting-edge methods were developed at the Broad Institute by MPG members and are now bearing fruit (see following section).

"These kinds of questions are absolutely critical — how do we do unbiased discovery and genuinely find the genes that we can hang our hat on and say, 'This is a real thing we can trust', before we start doing functional studies and clinical translation," says MacArthur.

MPG researchers also continue to use a study design that gained popularity before large-scale genome sequencing became feasible. This approach, known as a genome-wide association study (GWAS), uses DNA microarrays, or "chips," to rapidly skim the genome, looking for common genetic variations tied to disease. Over the past several years, genome-wide association studies have revealed hundreds of genetic risk factors for a range of common diseases, including type 2 diabetes, heart disease, inflammatory bowel disease, and many others.

GWAS and genome sequencing can help researchers zero in on specific genes or regions of DNA, but they do not provide enough information to directly illustrate how those genes or regions influence human biology and disease. For example, does a variant completely cripple the function of a gene's protein and affect key cellular processes, or does it reduce the protein's activity only slightly? Or, is the variant functionally silent, having no effect at all on human cells and tissues? Without methods that can systematically probe gene activity, scientists cannot distinguish these effects with certainty.

To explore these questions, MPG scientists are wielding a variety of tools, including large-scale methods for measuring gene activity. These include RNA

sequencing (RNA-Seq), which decodes the full suite of RNA present in a cell (RNA is the chemical messenger that carries the information encoded within genes, allowing it to be translated into proteins), and ChIP-Seq (chromatin immunoprecipitation sequencing), which reads the information contained within the epigenome (a collection of chemical tags embedded within the genome that determines how genes get turned on or off). Broad researchers are also developing unique, high-throughput, cell-based assays that are customized to particular genes of interest and can systematically test how variations in those genes affect specific processes within the cells.

KEY FINDINGS

In recent years, MPG researchers have disseminated many important genetic discoveries and resources to the scientific community, spanning the full bench-to-bedside spectrum from basic discoveries of the genes and genetic architecture of disease to findings that have direct and immediate applications for patients. These include:

I.   The discovery of genetic mutations that naturally protect their carriers from common conditions such as type 2 diabetes and heart disease, yielding insights into human biology that are already being leveraged to develop novel therapies;

II. Genetic analyses showing that high levels of HDL, the so-called "good" cholesterol, do not necessarily protect patients from heart disease — a finding with direct implications for drug development;

III. The discovery of an easily detectable "premalignant" state in the blood that could serve as a basis for cancer screening;

IV. The development of innovative resources to accelerate genetic discoveries, such as a public database that combines whole-exome sequencing data from nearly 100,000 people — the largest such dataset ever assembled; and

V.   The development of systematic, high-throughput methods to test genetic variants for their biological effects.

We further highlight and describe each of these seminal findings below.

I. Uncovering clues to disease protection

Typically, scientists have thought of inherited genetic variants as raising risk of disease — but in recent years, MPG scientists and others have found several variants that prevent people from getting sick. These so-called "protective mutations" have been found in people who carry major risk factors for disease yet remain healthy. They are a particularly valuable source of information for drug discovery; therapies that mimic them may also protect people from disease.

A well-known example of this phenomenon is the gene *CCR5.* Mutations in this gene were found to protect against infection with HIV, the virus that causes AIDS. Drugs have been developed that block the *CCR5* protein. A similar protective association for heart disease set off a race to discover new cholesterol-lowering drugs when mutations in the gene *PCSK9* were found to lower cholesterol levels and heart disease risk.

In the last year, MPG researchers have uncovered important new examples of mutations that protect against disease. In two separate studies, MPG member Sekar Kathiresan and his team uncovered loss-of-function mutations in the *APOC3* and *NPC1L1* genes that protect against coronary artery disease. The *APOC3* mutations lower levels of triglycerides, a type of fat in the blood, and significantly reduce a person's risk of coronary heart disease, dropping it by 40 percent. The work sheds light on the biological role of triglycerides and contributes to a growing body of knowledge that suggests that high triglyceride levels — rather than low HDL — are a major culprit in heart disease. The mutations in *NPC1L1* lower the levels of

LDL cholesterol — the so-called "bad" cholesterol — in the blood. Individuals who carry these inactivating mutations also have a lower risk of coronary heart disease — roughly half the risk compared to those individuals without the mutations.

"Protective mutations like the ones we've identified for heart disease are a treasure trove for understanding human biology," says Kathiresan, director of preventive cardiology at MGH, associate professor of medicine at Harvard Medical School, and an associate member of the Broad Institute. "They can teach us about the underlying causes of disease and point to important drug targets."

"Protective mutations like the ones we've identified for heart disease are a treasure trove for understanding human biology."

— Sekar Kathiresan, M.D.

In another study, former MPG Program Director David Altshuler, now Executive Vice President of Global Research and Chief Scientific Officer at Vertex Pharmaceuticals, revealed that loss-of-function mutations in the gene *SLC30A8* can reduce the risk of developing type 2 diabetes, even in people who have risk factors such as obesity and old age.

"It's been recognized for some time now that naturally occurring human mutations can suggest potential disease treatments that are more likely to succeed than those based on intuition or model systems," says Jason Flannick, first author of the study. "Our *SLC30A8* discovery represents the first protective, loss-of-function mutations for type 2 diabetes — and highlights a new therapeutic hypothesis."

In addition to underscoring the power of genetics to inform drug discovery, these studies also highlight the large number of samples required — on the order of 150,000 samples — as well as the close collaborations among researchers needed to enable sample collection at such a massive scale.

II. Casting doubt on HDL's "good" side

In a landmark study published in 2012, a large-scale genetic study led by Kathiresan and his colleagues found no causal association between the so-called "good cholesterol," HDL, and heart disease, challenging the long-held view that increasing HDL levels would lower the risk of heart disease. The team explored naturally occurring genetic variations in humans to test the connection between HDL levels and heart attack. By studying the genes of roughly 170,000 individuals, the team discovered that, when examined together, the 15 HDL-raising variants they tested do not reduce the risk of heart attack. These genetic studies are consistent with the recent negative results from multiple clinical trials of agents that raised HDL substantially, but failed to lower risk of heart disease.

"It's been assumed that if a patient, or group of patients, did something to cause their HDL levels to go up, then you can safely assume that their risk of heart attack will go down," says Kathiresan. "This work fundamentally questions that."

III. Finding harbingers of disease

In recent back-to-back papers (found here and here) in the *New England Journal of Medicine,* two research teams affiliated with the Broad and its partner institutions uncovered an easily detectable, "pre-malignant" state in the blood that significantly increases the likelihood that an individual will go on to develop blood cancers such as leukemia, lymphoma, or myelodysplastic syndrome. The discovery opens

new avenues for research aimed at early detection and prevention of blood cancer.

Most genetic research on cancer to date has focused on studying the genomes of advanced cancers, to identify the genes that are mutated in various cancer types. These two new studies instead looked at somatic mutations (those that cells acquire over time as they replicate and regenerate within the body) in DNA samples from people not known to have cancer or blood disorders.

Taking two very different approaches, the teams found that a surprising percentage of those sampled had acquired a subset — some but not all — of the somatic mutations that are typically present in blood cancers. These individuals were more than ten times more likely to go on to develop blood cancer in subsequent years than those in whom such mutations were not detected.

The "pre-malignant" state identified by the studies becomes more common with age; it is rare in those under the age of 40, but appears with increasing frequency with each decade of life that passes, ultimately appearing in more than 10% of those over the age of 70. Carriers of the mutations are at an overall 5% risk of developing some form of blood cancer within five years. This "pre-malignant" stage can be detected simply by sequencing DNA from blood.

"People often think about disease in black and white — that there's 'healthy' and there's 'disease' — but in reality most disease develops gradually over months or years. These findings give us a window on these early stages in the development of blood cancer," said Steven McCarroll, senior author of one of the papers. McCarroll is an assistant professor of genetics at Harvard Medical School and director of genetics at the Broad's Stanley Center for Psychiatric Research. Benjamin Ebert, an institute member of the Broad and associate professor at Harvard Medical School and

Brigham and Women's Hospital, is the senior author of the other paper.

In addition, the findings show just how important it is to collect and share large datasets of genetic information: both studies relied on a total of roughly 30,000 DNA samples collected for studies completely unrelated to cancer.

IV. Bringing exomes together

With a view toward making data from whole-exome sequencing more useful and accessible, an international team of investigators, led by MPG Program Associate Director Daniel MacArthur, recently launched the Exome Aggregation Consortium (ExAC). MacArthur and his ExAC colleagues collected exome data from 92,000 individuals and jointly analyzed them using standardized methods across all samples. This consistency in analysis is critical to ensure that any genetic discoveries made using these data are bona fide, and not technical flaws stemming from variability in sequencing technologies or analytical pipelines.

Using this exome data, the consortium has created a publicly accessible data set spanning nearly 61,000 individuals, which can be used as so-called "reference set" — a comparison control group that other researchers can match against their own genetic data to ensure that their variants of interest do not appear. The ExAC reference set, while not drawn exclusively from healthy individuals (many of the samples come from patients with adult-onset diseases), is proving to be a particularly useful resource for researchers who study rare, pediatric-onset diseases. The website has already garnered 720,000 page views since its release in October 2014.

V. Probing gene variants for their functional impact

Over the past few years, genomic methods have unearthed several thousand genetic variants that are associated with human disease. Yet the impact these variants have on disease biology is readily discernible in only a very few cases. To determine with certainty how variations within a gene affect its function, scientists need to rigorously and systematically test those variants using experimental approaches relevant to the disease of interest. Currently, high-throughput tools of this nature are not widely available, representing a key area of scientific need.

Recently, MPG scientists set out to develop such tools to accelerate research in two distinct areas of biology: (1) the role of variations in the low-density lipoprotein receptor (LDLR) gene, associated with the risk of coronary artery disease and (2) the effects of mutations in the PPARG gene, linked with type 2 diabetes risk. As part of the LDLR study, the researchers developed a scalable, cell-based assay to test the biological consequences and disease relevance of 70 distinct gene mutations. Similarly, in the PPARG study, the scientists analyzed 49 gene variants using a novel, high-throughput assay that models the differentiation of human fat cells and subsequent risk of type 2 diabetes. The true power of these studies, in addition to their technological innovation, lies in their ability to help researchers pinpoint which genetic variants have biological effects that influence the course of disease and which do not.

While the multitude of findings described here are just a sampling of the latest advances in human genetics flowing from the Broad, they underscore the fundamental shift taking place within the MPG community and beyond — one that is moving away from pure data generation and towards the application of robust, reliable methods for unraveling the genetics of disease biology and harnessing genome-based knowledge to improve human health.

FURTHER READING

Thormaehlen A, et al. Systematic cell-based pheno-typing of missense alleles empowers rare variant association studies: A case for LDLR and myocardial infarction. *PLOS Genetics.* DOI: 10.1371/journal.pgen.1004855.

Majithia A, et al. Rare variants in PPARG with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *PNAS* DOI: 10.1073/pnas.1410428111.

Crosby J, Peloso G, and Auer P, et al. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *New England Journal of Medicine.* DOI: 10.1056/NEJMoa1307095.

The Myocardial Infarction Genetics Consortium Investigators. Inactivating mutations in *NPC1L1* and protection from coronary heart disease. *New England Journal of Medicine.* DOI: 10.1056/NEJMoa1405386.

Flannick J, et al. Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nature Genetics.* DOI: 10.1038/ng.2915.

Voight B, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet.* DOI: 10.1016/S0140-6736(12)60312-2.

Jaiswal S, et al. Age-related clonal hematopoiesis associated with adverse outcomes. *New England Journal of Medicine.* DOI: 10.1056/NEJMoa1408617.

Genovese, G et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *New England Journal of Medicine.* DOI: 10.1056/NEJMoa1409405.

Last updated April 2015