Deep characterization of molecular function of genetic variants in the human genome is becoming increasingly important for understanding genetic associations to disease and for learning to read the regulatory code of the genome. In this paper, I discuss how recent advances in both quantitative genetics and molecular biology have contributed to understanding functional effects of genetic variants, lessons learned from eQTL studies, and future challenges in this field.

Most of human genetics research falls under two main questions: What are the genetic origins of variation in human disease and other traits? How does the blueprint of the human genome function to give rise to a living individual? These questions have different historical roots—in quantitative or medical genetics and molecular biology, respectively—as well as different molecular and statistical methods, and thus for decades they have been largely distinct areas of research. However, a question of increasing importance for understanding the human genome lies at their intersection: What are the functional effects of genetic variants across the human genome?

The study of the evolutionary origins of human genetic variation and its contribution to human disease and traits has its origins in quantitative, statistical, and population genetics. Advances in high-throughput genotyping and sequencing technologies during the past 10 years have led to tremendous progress in this field, with the HapMap and 1000 Genomes projects (The International HapMap Consortium et al. 2007; The 1000 Genomes Project Consortium 2012) creating the foundation for hundreds of genome-wide association studies (GWAS) and now also rare variant analyses in the context of both common and rare diseases (Bamshad et al. 2011; Lee et al. 2014). However, these maps of genetic associations to disease do not give us direct information of the function of these variants: how they perturb the biology of the genome, the cell, and eventually the organism to affect disease—or from a population genetics perspective, to affect different traits? How does the blueprint of the human genome function to give rise to a living individual? These questions have different historical roots—in quantitative or medical genetics and molecular biology, respectively—as well as different molecular and statistical methods, and thus for decades they have been largely distinct areas of research. However, a question of increasing importance for understanding the human genome lies at their intersection: What are the functional effects of genetic variants across the human genome?

On the other side, understanding the mechanistic function of the human genome—as well as genomes of other species—has always been one of the fundamental questions of molecular biology. During the past five years, the approach has become genome-wide via the development of diverse high-throughput sequencing assays, applied to multiple cell types. Projects such as ENCODE (The ENCODE Project Consortium 2012), the Epigenomics Roadmap (Roadmap Epigenomics Consortium 2015), and FANTOM (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) 2014) have produced large catalogs of functional elements in the genome—or more accurately, some genomes, since naturally, there is no archetype of the human genome. These studies do not typically capture variation in genome function among individuals, and the contribution of genetic differences in variation between samples is often ignored in study design. Thus, while these resources are used to annotate the putative regulatory function of genetic variants, this is done via indirect inference rather than direct measurement of genetic contribution to human phenotype diversity at the cellular level.

The need to bridge conventional quantitative genetics and functional or molecular genetics has now become widely acknowledged (Fig. 1). The concept is not new—medical genetics has a long history of characterizing cellular effects of disease-causing mutations. However, the development of genome-wide methods now allows systematic high-throughput analysis, which is eventually more cost-efficient and informative of generalizable patterns than laborious locus-specific characterization. High-throughput analysis, with scalable and robust molecular assays, careful statistical analysis, and deep biological interpretation, are essential to achieve the future goal of being able to accurately read the genetic code, i.e., predict functional and phenotypic effects of genetic variants.

Mapping regulatory variation by QTL approaches

Expression quantitative trait locus (eQTL) analysis has been the trailblazer in genome-wide functional population genomics. First applied in humans in the mid-2000s (Cheung et al. 2005; Stranger et al. 2007), associating genotypes to gene expression levels in population samples has become a mainstream approach to map variants that affect gene expression levels in cis (e.g., in Emilsson et al. 2008; Montgomery et al. 2010; Pickrell et al. 2010; Grundberg et al. 2012; Lappalainen et al. 2013; Battle et al. 2014; GTEx Consortium 2015); for a recent review, see Albert and Kruglyak (2015). Results from GWAS studies have been a major motivator for this work: 80% of genetic associations to common diseases are outside coding regions, which highlights the necessity of understanding regulatory variation (Farh et al. 2015). By now, eQTLs studies have uncovered >10,000 genes with eQTLs, demonstrating that common regulatory variants are extremely widespread in the genome (Lappalainen et al. 2013; Battle et al. 2014; GTEx Consortium 2015). This has allowed us...
to learn properties of proximal regulatory variants affecting gene expression in cis: They show widespread sharing across populations (Stranger et al. 2012), they are enriched for targets of positive natural selection (Fraser 2013; Grossman et al. 2013), and they are often located in promoter and enhancer regions but also, e.g., in 3' UTRs (Lappalainen et al. 2010; Gaffney et al. 2012; Battle et al. 2014; GTEx Consortium 2015).

Disease-associated variants are expected to impact cellular phenotypes that ultimately underlie the change in disease risk; indeed, several studies have shown an overrepresentation of eQTLs among GWAS loci (Nica et al. 2010; Nicolae et al. 2010). In hundreds of GWAS loci, eQTL associations have allowed the pinpointing of the GWAS variant to the likely target gene; annotation analyses sometimes indicate specific regulatory mechanisms, and tissue- or cell-type-specific eQTL data can point to tissue-specific mechanisms of disease etiology. For example, a genetic variant rs633185 with association to a QT interval is in high linkage disequilibrium with an eQTL that is particularly active in the heart but not in most other tissues (GTEx Consortium 2015).

However, showing that the eQTL and GWAS association signals in the same locus are driven by the same causal variant, rather than randomly overlapping, is not trivial despite several proposed statistical methods (Nica et al. 2010; Giambartolomei et al. 2014). Even when such statistical evidence is solid, real proof of shared causality cannot be obtained without experimental perturbations in cell lines and/or model organisms. Furthermore, most GWAS hits in noncoding regions still remain unexplained by current eQTL catalogs, motivating further research—both more comprehensive eQTL analysis and other approaches (Farh et al. 2015).

A key feature of regulatory genetic effects is its context-specificity, i.e., varying effects of a given variant due to differences in the surrounding cellular or genomic environment. This is an area of intensive study, as many key questions are currently unknown: how widespread such variable effects are, what the key mechanisms are, and what consequences are at the level of the organism. Several studies have provided insight into how the effects of cis-regulatory variants can be modified by tissue-specificity, systemic effects such as sex, and cellular stimuli mimicking environmental effects (Dimas et al. 2012; Ye et al. 2014; GTEx Consortium 2015).

The largest ongoing project in this domain is the Genotype Tissue Expression (GTEx) project (GTEx Consortium 2013, 2015), with analysis of genotype and RNA sequencing data, as well as other assays, eventually from over 30 tissues from 900 individuals. This project is building a foundation of gene expression and eQTL variation across human tissues in the normal population and provides an unparalleled resource for the scientific community. However, the primary tissue samples in this and many other projects consist of multiple cell types, and further characterization of the architecture of regulatory variation in diverse, specific cell types will be important to capture the full biological complexity and avoid averaging out effects from rare cell types.

While samples from a few hundred individuals are sufficient for well-powered standard cis-eQTL analysis, further increase of sample sizes is essential for capturing other, more subtle genetic effects on gene expression. The most important gap in the current literature concerns trans-eQTL associations to distal genes in the human genome. They are likely to explain a large proportion of heritable variation in gene expression and also act as modifiers of cis-eQTLs (Price et al. 2011; Grundberg et al. 2012; Buil et al. 2015). However, few studies have been large enough to capture them well (Westra et al. 2013; Battle et al. 2014), and characterizing their properties and mechanisms is an important topic for future research. Another controversial question in human genetics is epistasis or interaction between genetic variants in which combinations of variants either in cis and in trans may affect the trait outcome, and gene expression has been used as a model trait to detect such interactions (Brown et al. 2014; Hemani et al. 2014). However, pinpointing specific interactions has been challenging with the existing sample sizes and statistical methods, and the prevalence, mechanisms, and phenotypic importance of genetic epistasis remains currently unsolved. Finally, as larger and larger studies capture more of hereditary variation in gene expression, predictive imputation of gene expression levels in individuals, based on genotype data, is becoming possible, allowing association studies between disease phenotypes and predicted gene expression levels (Gamazon et al. 2015).

In addition to ongoing efforts in eQTL mapping, the same approach is increasingly being applied to other types of quantitative phenotypes of the cell, for example, to characterize genetic effects on chromatin state (Degner et al. 2012), methylation (Bell et al. 2011; Gutierrez-Arcelus et al. 2015), and transcript stability (Pai et al. 2012), as well as translation and protein levels (Battle et al. 2015). These cellular QTL studies are enabled by continuing development of scalable and affordable molecular assays that can be applied to hundreds of samples, ideally from multiple cell types and conditions. Other cellular QTLs have uncovered regulatory mechanisms of GWAS loci that are not captured by eQTL analysis, and thus QTL analysis of various cellular phenotypes is likely to continue to be one of the primary approaches for uncovering functional mechanisms of GWAS associations. Furthermore, integration of different QTL data provides extremely valuable information of causal mechanisms of genome function and gene regulation—for example, when and how genetic variants affecting epigenetic state lead to change in gene expression (Gutierrez-Arcelus et al. 2013; Pai et al. 2015).

**Functional effects of rare variants**

One of the major caveats of the QTL approach is that, as an association analysis, it lacks statistical power to pinpoint effects of rare variants, which have become a major target in human genetics research. Currently, analysis methods for high-throughput analysis of cellular effects of rare variants are still under development (Li et al. 2014). Prior knowledge of predicted functional effects can help, derived from annotation of the variants—such as whether a variant...
Breaking the regulatory code with genetic perturbations

The importance of cellular QTL approaches is not only in filling in the functional gaps of the GWAS catalog. One of the ultimate goals of genomics is to learn to read the regulatory code and eventually predict regulatory changes caused by any genetic variants. Naturally occurring variation is still the world’s largest mutagenesis “experiment,” and systematic analysis of how genetic variants affect the cell is an important source of information for understanding the basic biology of the genome. Given how common eQTLs are, it is clear that the vast majority of them have no effect on organism-level phenotype, but yet they are informative of how eQTLs are, it is clear that the vast majority of them have no effect standing the basic biology of the genome. Given how common aperturbationsofgenetic variants affect the cell is an important source of information for understanding the basic biology of the genome. Given how common eQTLs are, it is clear that the vast majority of them have no effect on organism-level phenotype, but yet they are informative of how eQTLs are, it is clear that the vast majority of them have no effect

The functional effects of genetic variants in the human genome allow precise perturbation of the genetic code, and their high throughput yields comprehensive data for computational analysis of sequence motifs and their function. However, these approaches rely on artificial systems, with the elements being outside their native genomic context, and analysis in vitro may not always sufficiently recapitulate the complexity of the cellular environment in vivo.

The novel genome editing technology by CRISPR/Cas is opening a vast universe of new possibilities for analyzing how genetic variants affect phenotypes (Doudna and Charpentier 2014). Introducing variants in human cell lines and measuring the resulting cellular phenotypes in a high-throughput manner provides the possibility for experimental validation of cellular QTLs in their native genomic context, as well as testing cellular consequences of systematic high-throughput mutagenesis (Findlay et al. 2014). In addition to genome editing, CRISPR assays that allow targeted transcription regulation will be valuable tools for understanding causal networks of genome regulation (Konermann et al. 2015). Truly high-throughput applications of the CRISPR technology are currently limited to gene knock-out screens (Shalem et al. 2014), but this will likely change during the next few years as both molecular assays and analytical approaches develop. However, genome editing can be used to manipulate the human genome only in cell lines, and extrapolating that information to understand a complex living organism and its phenotypes is unlikely to be straightforward. Thus, observational data from human tissue samples as well as modified model organisms will continue to be important for interpreting and applying results from CRISPR assays.

Functional genomics and human health

How has a decade of research into the cellular effects of genetic variants across the genome contributed to improving human health? Information of regulatory mechanisms behind genetic associations to disease can be informative of novel drug targets and other interventions, and this will hopefully be an increasingly fruitful approach in the near future. In the rare variant domain, knowing the functional effects of a variant causing (or protecting against) disease is important for knowing whether the treatment should, for example, block a truncated protein or boost gene expression or protein levels. Furthermore, understanding the range of functional variation observed in healthy individuals can be a powerful tool for understanding what type of manipulations of the functional landscape of the cell are likely to be well tolerated.

While exome and genome sequencing are rapidly becoming part of standard clinical practice, the same is not yet true for high-throughput assays in functional genomics such as RNA sequencing, epigenome analysis, and protein quantification. Yet, these assays can have significant clinical value. In addition to studies aiming to profile patients based on the transcriptome or the epigenome (Michels et al. 2013), these assays are also being pursued as personal biomarkers allowing a longitudinal monitoring of cellular state (Chen et al. 2012). Furthermore, geneticists are now painfully aware of how difficult the interpretation of an individual’s genome is, but the epigenome and the transcriptome can provide a layer of information close to the genome that enables better interpretation of phenotypic effects of genetic variants. It is only now that the assays, analytical approaches, and general understanding of the spectrum of epigenome and transcriptome are starting to be advanced enough for clinical analysis. Although extensive benchmarking and standardization of bedside functional
genomics is still lacking, functional genomics assays hold substantial clinical potential for the future.

Summary

Analysis of functional effects of genetic variants has become one of the fastest growing areas of human genetics—and rightly so, as it addresses some of the most burning questions in the quest toward understanding genome function as well as genetic background of phenotypic variation in humans. It brings together the formerly largely distinct fields of molecular biology and quantitative genetics, contributing to the development of both (cf. Fig. 1). Quantitative genetics is now reaching beyond disease associations as statistical constructs, toward real biological understanding. On the other hand, molecular biology has a lot to gain in understanding the range of population variation in genome function and using genetic effects as a causality anchor in cellular networks and disease etiology. The GWAS community has been exemplary in establishing commonly accepted gold standards for statistical analysis. While functional genomics data is more diverse in nature, the development toward similarly high standards must continue.

The future of this field looks bright: Increasing sample sizes allow deeper interrogation of more and more complex effects of genetic variants, and characterization of additional cell types and conditions with diverse assays will provide not only more comprehensive catalogs but also deeper mechanistic understanding beyond incremental increases. Genome editing technology will redefine the toolkit in unprecedented ways. Massive data sets often produced by consortium projects will continue to fuel research and provide the accessible, carefully curated data resources for discovery, both at the level of individual loci and, in particular, in genome-wide systems-level approaches. However, many novel biological phenomena, technologies, and statistical approaches will still be discovered and developed in individual laboratories in the future, via analysis of both humans and model organisms. Both detailed dissection of specific mechanistic components of genome function and systems-level approaches to link all the components back together are necessary. Finally, the application of technologies and results from functional genomics to improve drug development and interpretation of personal genomes has substantial potential to improve human health.

Acknowledgments

I thank Ana Vasileva, Stephane Castel, Pejman Mohammadi, and Margot Bradt for helpful comments. The author is funded by National Institutes of Health (NIH) grant R01MH101814.

References


1430 Genome Research www.genome.org


