Leading Edge
Primer

Cell

Mapping Rare and Common Causal Alleles for Complex Human Diseases

Soumya Raychaudhuri[1,2,3,4,*]
[1]Division of Genetics
[2]Division of Rheumatology
Brigham & Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[3]Partners HealthCare Center for Personalized Genetic Medicine, Boston, MA 02115, USA
[4]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
*Correspondence: soumya@broadinstitute.org
DOI 10.1016/j.cell.2011.09.011

Advances in genotyping and sequencing technologies have revolutionized the genetics of complex disease by locating rare and common variants that influence an individual's risk for diseases, such as diabetes, cancers, and psychiatric disorders. However, to capitalize on these data for prevention and therapies requires the identification of causal alleles and a mechanistic understanding for how these variants contribute to the disease. After discussing the strategies currently used to map variants for complex diseases, this Primer explores how variants may be prioritized for follow-up functional studies and the challenges and approaches for assessing the contributions of rare and common variants to disease phenotypes.

Most common diseases are complex: many genetic and environmental factors mediate the risk for developing the disease, and each individual factor explains only a small proportion of population risk (Cardon and Abecasis, 2003). Genome-wide genotyping with high-throughput approaches has led to the identification of >2,600 associated common risk alleles, with convincing associations in >350 different complex traits (most with modest effect size of odds ratio <1.5) (Hindorff et al., 2009). More recently, low-cost, high-throughput sequencing of exomes and whole genomes is giving investigators access to the spectrum of rare inherited variants and de novo mutations. Once an associated allele is discovered, a critical step to characterizing pathogenesis is the definition of the causal allele, that is the functional allele that influences disease susceptibility and explains the observed association. However, for the vast majority of associated alleles, the identities of causal genes and variants, as well as the function of these variants, remain uncertain. This Primer discusses the population genetics features of rare and common alleles, strategies for connecting these alleles to disease, and strategies to prioritize them for functional follow-up studies.

Population Genetics of Rare and Common Alleles

Geneticists have long debated the extent to which rare and common alleles contribute to complex disease (Pritchard, 2001; Pritchard and Cox, 2002; Reich and Lander, 2001). Although there is evidence of susceptibility alleles across the frequency spectrum in many complex diseases, it is important to realize that rare alleles and common alleles have different population characteristics that are relevant to medical genetics.

The exact distinction between rare and common alleles is to an extent an arbitrary one. We define common alleles as those with frequencies >1%; these alleles are frequent enough that they can be queried by genotyping in standard marker panels. Rare alleles are polymorphic alleles with <1% frequency that might be most effectively studied with sequencing technologies. The rarest alleles are seen in only a handful of individuals or are private to a single individual and can only be observed by sequencing.

The Origin of Polymorphic Alleles

De novo mutations occurring spontaneously in individuals are constantly and rapidly introduced into any population. These mutations are initially "private" to the individual that they occurred in but might then be passed on to progeny. Most of these mutations are quickly filtered out or lost by genetic drift and will never achieve appreciable allele frequencies. I illustrate this concept by a simulation in which de novo neutral mutations (conferring no effect on fitness) are introduced into a population of 2,000 diploid individuals. In 31 generations, 95% of these mutations disappear from the general population, and not one of these mutations achieves an allele frequency of >1% in 200 generations (see Figure S1 available online). Mutations that are deleterious are even more rapidly purged from populations. Although any de novo mutation is very unlikely to become a common allele, even a somewhat deleterious mutation may persist for a few subsequent generations as a rare allele before disappearing.

Thus populations harbor many rare alleles, most of which have been derived recently, but relatively few common ones. In fact, there is only about one common variant on average per ~500 bp in European populations (1000 Genomes Project Consortium, 2010). On the other hand, recent and rapid expansion of human populations has resulted in the presence of many rare alleles. At the extreme of the allele frequency spectrum are de novo mutations; each individual harbors ~40

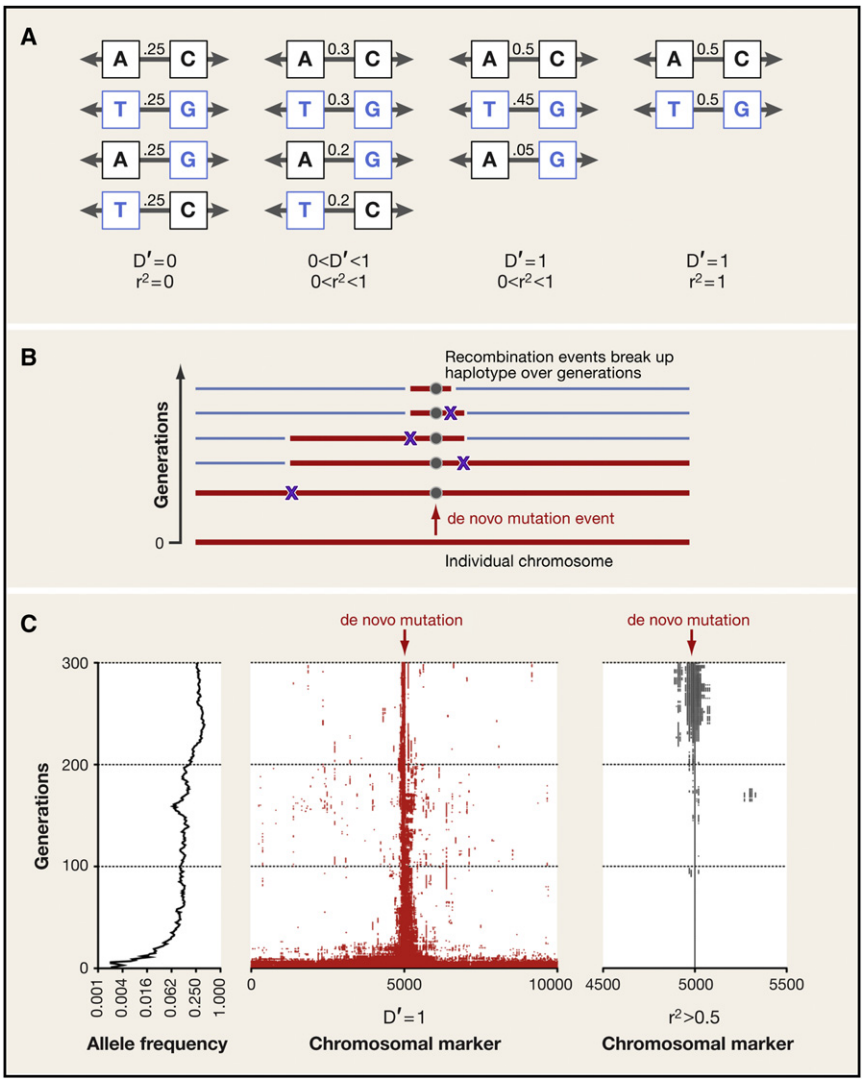

Figure 1. Linkage Disequilibrium and Haplotype Lengths

(A) Linkage disequilibrium metrics. Left: For two markers that are random with respect to each other, each with a 0.5 allele frequency, there is no linkage between them; each resulting haplotype has a frequency of 0.25. Middle left: Here the two markers are not entirely random, and alleles at one marker correlate partially with alleles at the other marker. The A allele on the left is observed more frequently with the C allele on the right, and the T allele on the left is observed more frequently with the G allele on the right. Middle right: Here the two alleles are more tightly linked or have tighter LD than in the previous case. In this instance, the presence of the T allele on the left predicts with certainty the G allele on the right. This could be the case if the T allele arose de novo on a haplotype with the G allele on the right Right: For instances of tight LD, an allele at one marker predicts perfectly the allele at the other marker; in this case, these two markers form only two haplotypes.

(B) Changing LD properties of a persistent de novo mutation. A de novo event (circle), when it first occurs on a chromosome (bottom), is on one haplotypic background defined by the chromosomal markers on which it forms (red). As generations pass (moving upward), the event propagates through the population. Recombination events (Xs) occur, reducing the common haplotype (red) on which a variant is present and decoupling it from distal markers (blue).

(C) Simulating LD structure of a de novo event as it becomes a common variant. Here a computer simulation depicts a chromosome with 10,000 common markers with 1,000 randomly assigned hot spots. Random mating occurs here with an average of one recombination event per generation. A single rare variant is introduced in the middle of the chromosome on one individual (bottom) and allowed to propagate through the population. The left panel depicts the allele frequency as it increases through the generations (upwards). In the middle panel, all markers in LD with that variant (with $D' = 1$) are indicated with a red dot. Initially that variant is in LD with every common marker that it is in phase with on that chromosome, revealed by the red band stretching across the bottom of the plot. As random recombination events occur and the allele becomes more frequent, the number of markers in phase decreases, revealed by the shrinking red band in the middle. On the right panel, a gray dot indicates markers for which the genotypes correlate with the rare variant ($r^2 > 0.5$). For the first few generations, there are no other variants that correlate with the de novo mutation as it becomes a rare allele. As time progresses and the allele becomes more common, it begins to develop genotypic correlations with nearby variants that remain on the same haplotype.

de novo point mutations that may not be present in any other individuals (Conrad et al., 2011).

Common alleles tend to be more ancient than rare ones as it takes many generations for a rare allele to rise to a reasonable allele frequency. There are important exceptions to these generalizations. An ancient allele may be rare because it is being depleted from the population. A common allele may be recent if it confers a critical survival advantage or has emerged after a rapid population expansion from a small founder population.

Linkage Disequilibrium and Haplotypes

Genetic linkage is the tendency of alleles at nearby loci to be transmitted together; two nearby loci are in linkage disequilibrium (LD) when recombination events occur between them very infrequently. Two common metrics quantify pairwise LD between biallelic markers (see Figure 1A). The R-squared (r^2) between two markers is their correlation across chromosomes within a population. If two markers have $r^2 = 1$, then alleles are always in phase (or in *cis*) with each other; in a genetic study, their association statistics will be identical. The D-prime (D') between two markers is inversely related to the fraction of chromosomes that have had historical recombination between them. If $D' = 1$, two biallelic variants constitute only two or three haplotypes, whereas if $D' < 1$, all four possible haplotypes are present in the population. If $D' = 0$ or $r^2 = 0$, then the two markers are unlinked and statistically independent of each other.

Recombination events break down pairwise linkage between markers over time and reduce the lengths of haplotypes in a population. Recombination events are much more likely to occur in hot spot regions in the genome than in other regions (Myers et al., 2005). As a result, markers without a recombination hot

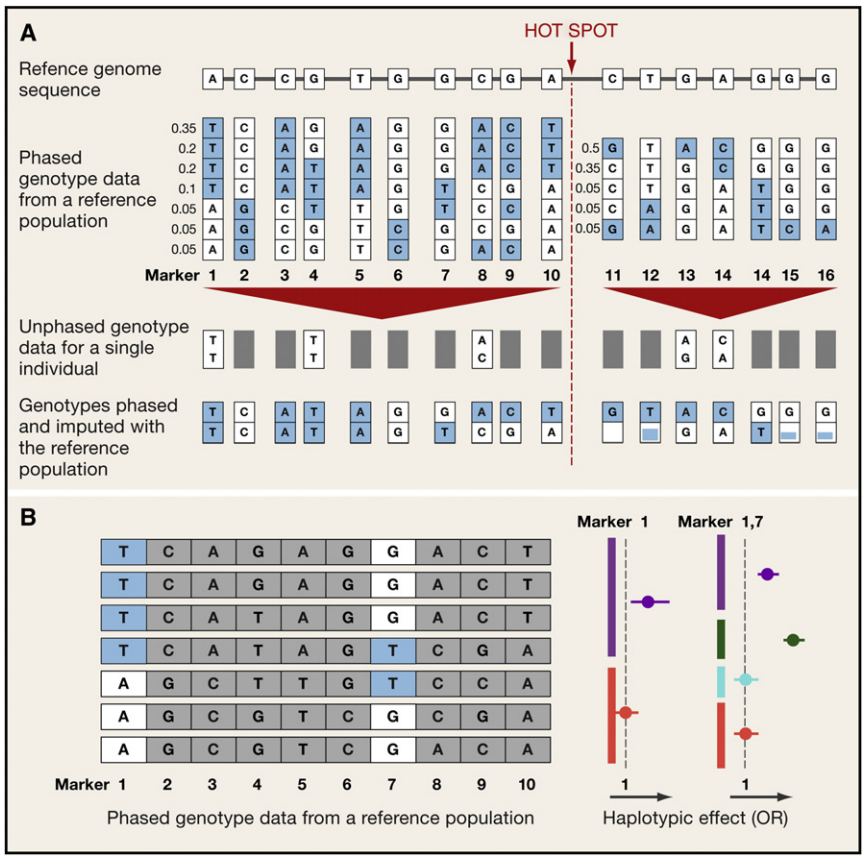

Figure 2. Common Variants and Fine-Mapping with Conditional Haplotype Analysis

(A) Common variants. This image illustrates the structure of common variants and LD blocks. The top lists a reference genome spanning ~10 kb and the reference genotypes of the polymorphic variants. The haplotype structure is broken up into two blocks by a recombination hot spot. Each block contains a set of markers in tight LD, which can be phased into a small number of haplotypes. Below that, a limited number of genotypes are depicted for a hypothetical individual because a commercial array would assay only a limited collection of all of the common variants in a region. The bottom row demonstrates how data for those genotypes can be phased using reference population data and how missing genotypes can be imputed if the haplotype can be inferred accurately. In some instances, imputed genotypes may be uncertain.

(B) Fine-mapping with conditional haplotype analysis. The left-hand side lists genotypes at ten variant sites (numbered) that define seven common haplotypes. Each row represents a haplotype, and genotypes at variant sites are listed in each column. Assuming that a common variant association is observed at marker 1, identical associations will be observed at the markers 2, 3, and 5 because their genotypes are correlated across haplotypes. In the first step, haplotypes are grouped by marker 1. The result is that the seven haplotypes form two subgroups (indicated by purple and red bars on the right). The purple group demonstrates association with disease (right). Including marker 7 breaks the groups up further into four haplotypes (indicated by purple, green, blue, and red bars on the far right). By adding marker 7, differential risk association between haplotypes is apparent. Whereas the T/G haplotype confers risk, the T/T haplotype confers even more risk. Thus marker 1 alone does not parsimoniously explain all of the risk at that locus.

spot between them are often linked over long periods of time and have high pairwise D′. Those markers can often be grouped into a set of limited number of common haplotypes (see Figure 2A). Phasing algorithms can be applied to determine markers in *cis* and to define the most likely haplotypes.

Rare alleles generally sit on long haplotypes whereas common alleles sit on shorter ones. When a mutation first occurs de novo on a chromosome, it occurs on the background of a single rare haplotype defined by all markers on that chromosome (see Figure 1B). Because the de novo mutation appeared as a random event, it initially has no correlation with other markers on that chromosome ($r^2 = 0$). In initial generations, prior to a recombination event, the mutation has $D' = 1$ with other markers across the chromosome. But, if the mutation survives generations and becomes a common allele, repeated recombination events fragment that haplotype and reduce its length. The allele retains high D′ to only proximate markers that are not separated from it by a recombination hot spot. As the variant becomes more frequent, so does the haplotype that it occurs on; over time the emerging variant develops correlation ($r^2 > > 0$) with the markers on that short haplotype (see Figure 1C).

Finding Pathogenic Variants, Both Rare and Common

Common variant associations to phenotype are often facile to find. Their high frequencies allow case-control studies to be adequately powered to detect even modest effects. Their high r^2 to other proximate common variants allows for association signals to be discovered by genotyping the marker itself or other nearby correlated markers. But mapping those associated variants to the specific causal variant that functionally influences disease risk can be challenging because the statistical signals invoked by intercorrelated variants are difficult to disentangle.

On the other hand, individual rare variant associations are challenging to find. Their low frequency renders current cohorts underpowered to detect all but the strongest effects, and lack of correlation to other markers often prevents them from being picked up by standard genotyping marker panels. But, once a rare associated variant is observed, mapping the causal rare variants is relatively facile because recent ancestry is likely to limit the number of intercorrelated markers.

Functional Properties of Pathogenic Variants, Both Rare and Common

Because common alleles tend to be ancient, they have weathered the influences of purifying negative selection. Therefore, common variants that influence disease risk are likely to have functionally modest effects that are compatible with their high population frequency. There are two possibilities outlined by Kryukov et al. that might allow for this (Kryukov et al., 2007). First, common variants that are medically detrimental act subtly or specifically to confer disease without altering evolutionary fitness. As an example, consider a variant that confers risk of

Box 1. Glossary

Associated allele: An allele that, in a genetic study, is observed to have differential allele frequencies in cases compared to controls. The presence of an association suggests that it, or some other variant in LD, is influencing disease susceptibility.
Causal allele: The functional allele that influences disease susceptibility and explains the observed associated allele.
Common alleles: Alleles with a high population frequency, typically defined as >1%. Standard marker panels can often be used to identify common allele associations.
Rare alleles: Alleles with a lower allele frequency of <1%. These alleles can be polymorphic in the population being seen in multiple distantly related individuals; alternately they might be alleles that are private to an individual or seen in a small number of closely related individuals.
De novo mutations: A mutation that has occurred in an individual and that was not inherited from a parent. These mutations are initially private. If a de novo mutation is passed on and persists through generations, it can become a polymorphic allele.
Linkage disequilibrium (LD): Two polymorphic loci are in LD when they are colocated, and alleles at those loci are distributed nonrandomly with respect to each other on chromosomes in the population. Linkage disequilibrium is present when recombination events between two loci occur infrequently. Two metrics for LD are r^2 and D′ (see Figure 1A).
Recombination hot spots: Individual regions within the genome that have frequent recombination events.
Negative selection: Selection acting to remove new deleterious mutations that reduces evolutionary fitness of an individual. Also known as purifying selection.
Positive selection: Selection acting to propagate new advantageous mutations that increase evolutionary fitness of an individual.
Balancing selection: Selection acting to increase allelic variability at a locus.
Genotype imputation: A statistical technique to infer missing genotypes in a set of individuals using a reference panel of genotyped individuals. Imputation exploits LD between genotyped and ungenotyped variants.
Genome-wide significance: A level of statistical significance typically used to establish association for a common variant in genome-wide association studies ($p = 5 \times 10^{-8}$), which assumes that there are ~1,000,000 effective independent tests genome-wide.
Stratification: A genetic confounder if there are differences in the ancestral origin of cases and controls. The resulting systematic allele frequency differences can result in false-positive associations.
Genomic inflation factor (λ): The ratio of the median of the observed chi-square statistics for an association study and the expected median chi-square statistic. If there is stratification, the test statistic is inflated, causing the genomic inflation factor to be substantially greater than 1, resulting in inappropriately significant p values.
Fine-mapping: The use of dense genotyping data around an associated allele to identify the causal allele(s) to account for the observed statistical signal in the region.
Second-generation sequencing: Recent sequencing technologies not using Sanger chemistry that characteristically generate many short read sequences.
Targeted region: The region of the genome selected for a sequencing experiment.
Whole-genome sequencing: A sequencing experiment where the full ~3 GBp of whole genome is sequenced. Does not require DNA capture. For most medical genetic studies, the sequencing data are not reassembled but mapped to a reference genome sequence.
Whole-exome sequencing: A sequencing experiment where the protein-coding sequences of all known genes are targeted, captured, and sequenced (~30 Mbp).
Coverage: In a sequencing experiment, coverage at a genomic position is the total number of reads mapped to that position.

addiction to tobacco (Thorgeirsson et al., 2008). Such a variant might have little impact on survival historically but might have specific neuropsychiatric effects that mediate the risk of 21st century diseases such as lung cancer or coronary artery disease that play a role later in life after reproduction. Second, forces that select specifically for these common variants counteract their medically detrimental qualities; the variant, although causing disease, also offers evolutionary benefit simultaneously. For example, common *ApoL1* variants that confer high risk of chronic kidney disease in African Americans protect from Trypanosoma brucei rhodesiense infection at the same time (Genovese et al., 2010).

Because rare alleles are typically more recent, they may not have been subjected to the same negative selective pressures yet and may include among them more relatively deleterious mutations. Rare alleles therefore often are enriched for those variants more likely to have more dramatic functional consequences. This is supported by data indicating that rare deletions are more likely than more common deletions to remove entire genes, exons, promotors, or stop codons (Conrad et al., 2010). Similarly, rare variants are twice as likely as common ones to be nonsynonymous (1000 Genomes Project Consortium, 2010). Because rare variants are relatively unrestricted in terms of their functional impact in general, a subset of rare pathogenic variants with large effect might offer more obvious insight about disease mechanism.

Common Variants

Detecting Common Variants with High-Throughput SNP Arrays

High-throughput genotyping of standard marker panels of common single-nucleotide polymorphisms (SNPs) has become possible with microarrays (Gunderson et al., 2005). Their application to large case-control sample collections has facilitated detection of even the most modest risk alleles, with odds ratios of 1.1 or less. There are a finite number of common variants present in the general population, i.e., <6 million are estimated in European populations (1000 Genomes Project Consortium, 2010). But nearby common SNPs are in LD with one another and define a limited number of haplotypes (see Figure 2A), so the effective number of independent variants is much fewer. Thus, genotyping a limited number of common variants genome-wide has the effect of covering many more common variants. In European populations, the Affymetrix 5.0 array with 440K SNPs has $r^2 > 0.8$ for 57% of common variants, and the Affymetrix 6.0 array with roughly double the number of SNPs (900K) has $r^2 > 0.8$ for 66% of common variants (Bhangale et al., 2008).

Genome-wide genotyping also allows investigators to use imputation to estimate genotypes of markers not directly genotyped; in doing so, it becomes possible to combine samples genotyped on different platforms. Probabilistic multipoint imputation algorithms, using a limited number of genotyped common variants, can determine the genotypes of ungenotyped common variants by comparing to a reference panel of comprehensively genotyped individuals (see Figure 2A). Most of these methods currently use probabilistic Hidden Markov Model approaches to infer the local LD structure (Browning, 2008; de Bakker et al., 2008).

Selecting Populations for Study

Initial efforts to map complex traits emphasized selected isolated populations, for example the Finish populations (Peltonen et al., 2000). These populations can offer the advantage of increased inbreeding, more uniform genetic and environmental backgrounds, detailed genealogical records, availability of intact extended families, and longer LD intervals. Populations that have undergone rapid population expansion may be of particular use because LD intervals are longer. The most successful validation of this approach is represented by deCODE genetics and their study of a wide-range of complex diseases in Iceland.

Now, investigators are increasingly focused on the inclusion of individuals from multiple ethnic backgrounds in order to enhance the ability of studies to discover risk alleles with variable allele frequencies across different backgrounds (Rosenberg et al., 2010). Different ethnic backgrounds might highlight different mechanisms of disease pathogenesis, including differences in environmental exposures, as well as reflect different degrees of genetic diversity and LD patterns. A striking example of this is the discovery of an *IL18B* variant that predicts response to hepatitis C treatment with equivalent effect in European, African, and Hispanic American patients; allele frequency differences of the variant explain about half of the differences in treatment response across populations (Ge et al., 2009).

Genome-wide Association Studies

In a case-control genome-wide association study (GWAS), samples are genotyped for a set of 100,000–2,000,000 markers; case and control allele frequencies are compared directly to each other. Statistical significance is assessed with a simple 2 × 2 chi-square test or with logistic regression when genotypes are probabilistic (e.g., from imputation).

Critical to the success of GWAS has been the application of stringent statistical significance thresholds that result in reproducible associations that account for the large number of simultaneous tests (Risch and Merikangas, 1996). Testing for common variant associations throughout the genome represents ~1 million independent tests (Hoggart et al., 2008). Thus investigators routinely use a genome-wide significance threshold representing a Bonferoni correction for multiple tests ($p = 0.05/10^6 = 5 \times 10^{-8}$).

Because effect sizes for most common variants are modest, large sample sizes and careful adjustment for subtle technical artifacts that can easily obscure results or produce false-positive associations are of paramount importance (Balding, 2006; Clayton et al., 2005; McCarthy et al., 2008). The genomic inflation factor is an important metric that indicates the extent of inflation due to stratification and other technical confounders. Fortuitously, the strength of genome-wide genotyping goes beyond simply measuring case-control allele frequency differences throughout the genome. It also allows investigators to look at patterns in the genotyping data to identify key technical confounders. For instance, patterns of excessive "missing" genotype data for an individual indicate that intensity data could not be clustered into genotype, likely as a function of low DNA quality or concentration. Another key confounder is population stratification, that is the presence of the systematic allele frequency differences observed in a population as a consequence of ancestry rather than case-control status. As a dramatic example, Campbell et al. showed, even in studies using only European populations, that not carefully adjusting for an individual's country of origin results in a highly statistically significant false-positive association for height at a lactase SNP (Campbell et al., 2005). Genome-wide genotype data allow investigators to identify and correct for case-control population stratification.

Once markers are identified as having statistically significant allele frequency differences in cases and controls, they are ideally replicated in independent populations. Replicating in an independent population not only adds statistical confidence to the results but also adds confidence that the results of the initial study are not the consequence of technical confounding or stratification.

Identifying an associated marker rarely clarifies whether the marker itself is the functional allele that causes altered disease susceptibility. The observed association at a marker might be the result of an underlying causal allele with high r^2 with the associated variant, a rare functional allele on a haplotypic background shared with the associated variant, or multiple functional alleles that cause an apparent association. Nevertheless, the causal alleles must closely correlate and be in LD with associated variants.

Fine-Mapping Common Variant Loci

Dense genotyping of markers in the region, followed by fine-mapping, can identify the causal allele, or at least reduce the number of potential candidates. The underlying assumption is that the causal allele will most parsimoniously explain the entirety of the evidence of association. In many instances, however, fine-mapping is complicated if the association is not being driven by a marker that has been genotyped; in those instances, it might be possible to identify a risk haplotype defined by genotyped markers and to then sequence selected individuals to identify the causal allele. Thus in order to fine-map effectively, dense genotyping to include all known markers in the region is key. Additionally, in many instances there might be multiple causal alleles, and in order to be powered to detect multiple effects, it is often necessary to densely genotype a large number of samples, perhaps more than those used to discover the association.

After densely genotyping a large number of samples, there are two major statistical tools utilized in fine-mapping common variants. The first is conditional regression. If a single lead marker (or another marker in perfect LD with it) is causal, then applying conditional regression adjusting for that lead marker should obviate all other association in the region. The second statistical tool is conditional haplotype analysis (Figure 2B). With conditional haplotype analyses, investigators start with data from a subset of the genotyped markers and phase genotypes to define

haplotypes. If the selected markers are causal, then the defined haplotypes should parsimoniously explain the risk at that locus. That is, the addition of additional markers (and thus creation of more haplotypes) should not explain risk better, and removal of any marker (and thus removal of haplotypes) should reduce the explained risk. With both approaches, if the causal allele is in perfect LD (r^2 = 1) with other markers, then distinguishing between statistically identical associations may not be possible.

One striking example of fine-mapping was an effort by Pereyra et al. where they used GWAS to demonstrate that multiple *HLA-B* classical alleles are associated with long-term viral load control in HIV-infected individuals (Pereyra et al., 2010). Then, with conditional haplotype analysis, they demonstrated that allelic risk was best defined by amino acid variation at a few sites along the binding groove of HLA-B.

Data from multiple ethnic populations may be particularly useful to fine-map associations (Rosenberg et al., 2010). Ideally a single allele might explain risk across multiple ethnic groups. This approach is effective only if the same causal allele is present with a high allele frequency in both, and there are ethnic differences in local LD structure. The inclusion of African populations might be particularly useful because LD patterns are generally shorter. This approach might be complicated if multiple different alleles in populations influence disease susceptibility within the same locus. Adrianto et al. looked at SNPs associated with systemic lupus erythematosus (SLE) spanning the *TNFAIP3* gene (Adrianto et al., 2011). When they looked at markers associated in Asian and European populations, they were able to fine-map the associated region from a span of ~100 kb to ~50 kb. Subsequent sequencing identified a novel AA > T single base pair deletion polymorphism that acts to disrupt an NF-κb binding site. This single variant explained the associated risk of the locus.

Rare Variants

It is possible that associated rare variants for complex diseases will be more facile to fine-map and to evaluate for functional impact. The discovery of a rare variant near a common variant might be particularly informative. A rare variant that clearly impacts one of the candidate genes implicated by a common variant might clarify which of the candidate genes is pathogenic. Furthermore, the rare variant's function might offer clues about the mechanism of the common variant. There have been several examples of this phenomena reported in the literature already. Common alleles associated with type II diabetes are near five genes, *PPARG*, *HNF1A*, *KCNJ11*, *WFS1*, and *HNF1B*, that have rare mutations that cause familial forms of diabetes (Voight et al., 2010). Similarly, 18 of the 95 known common variants associated with serum lipid levels are near genes that have been implicated in monogenic lipid disorders (Teslovich et al., 2010). Indeed studies to find rare coding variants near common risk loci have already shown success in type I diabetes (Nejentsev et al., 2009), age-related macular degeneration (S.R. and J. Seddon, unpublished data), and Crohn's disease (Momozawa et al., 2011).

The extent to which rare variants explain complex disease susceptibility in general remains an open question. It has been speculated that the gap between the heritability explained by known common variants and that which might be predicted from family studies might be explained by rare variants (Bansal et al., 2010), and that even many observed common variant associations might be the consequence of functional undiscovered rare variants (Anderson et al., 2011; Dickson et al., 2010). Other investigators have suggested that undiscovered common variants themselves might explain much of that missing heritability (Purcell et al., 2009; Yang et al., 2010).

Identifying Rare Variants with High-Throughput Sequencing

Advances in DNA capture and sequencing technology have greatly facilitated targeted, exome, and whole-genome sequencing (Maxmen, 2011; Ng et al., 2010) and have in the process enhanced the search for rare variants. Whereas the cost of sequencing is rapidly dropping, the computational and statistical challenges to rapidly aligning sequences to reference sequences, separating variant calls (SNPs, indels, and structural variants) from sequencing artifact, data storage, and establishing associations are mounting (McKenna et al., 2010).

Second-generation sequencing technologies have now come online and are distinct from prior approaches in that they do not use Sanger chemistry and are characterized by high sequencing yield with shorter reads (Shendure and Ji, 2008). The Illumina HiSeq 2000 system, for example, generates >1 billion 100 bp paired-end useable reads per run. Efficiently mapping a large volume of short reads to the reference genome accurately has been an important area of methodological progress (Li and Homer, 2010). Look-up (or hash-table) based methods map reads quickly but are not as accurate as less-efficient alignment-based methods. Accurate alignment is especially important in regions with short insertions or deletions (indels); poor alignment in such regions can result in false-positive SNP calls and false-negative indel calls. Repetitive genomic regions and regions with homology can be challenging to map and, in some instances, may not be possible to query effectively. Paired-end sequencing generates two sequence reads from opposite ends of the same contiguous genomic fragment and helps overcome some of these alignment issues.

To sensitively and accurately call a heterozygote nonreference base, a minimum of ~20× coverage is necessary to overcome the uncertainty resulting from sampling short sequence reads across a diploid genome. Additional coverage may be necessary to compensate for random and nonrandom sequencing error, which may vary across technologies. Even with a high-coverage sequencing experiment, the coverage is typically nonuniform across the targeted region. Nonuniform coverage can be related to biases in DNA capture technologies, in unequal pooling of amplicon products from different genomic regions or individuals, and in intrinsic sequence properties (Harismendy et al., 2009). Careful experimental technique and sample normalization can minimize some biases in coverage. Average coverage of an experiment is thus not as useful of a metric as is the percentage of target genomic region achieving more than a prespecified coverage threshold. A set of independently genotyped SNPs to verify sequence-based genotype calls and assess the accuracy of sequencing studies is useful to confirm accuracy.

Sequencing can be applied to a set of samples to discover variants or to genotype variants. For variant discovery, sequence data can be pooled across multiple samples to boost power to detect a nonreference base. After application of sequencing to

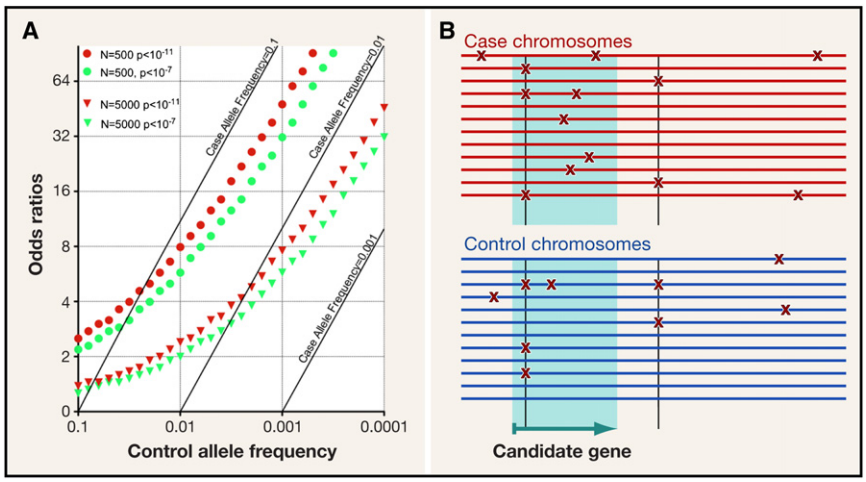

Figure 3. Power to Find Rare Variants and Burden Testing

(A) Power to find rare variants. Here is a plot of 80% power to discover rare associated alleles at $p < 10^{-7}$ and $p < 10^{-11}$ for cohorts of both 500 and 5000 cases and controls. The control allele frequency and odds ratio (OR) are plotted along the x axis and the y axis, respectively. Diagonal lines indicate corresponding case allele frequencies.

(B) Burden testing. Here data from sequenced cases (top) and controls (bottom) are depicted around a gene of interest. Each horizontal line represents an individual. Variants are shown as red Xs. Certain variants are rare (i.e., seen once), and others are more common (vertical line). In this example, the case variants within the candidate gene (arrow at bottom and blue shading) are seen more frequently than in controls. If common variants are excluded, there are five case chromosomes with a rare variant compared to one control chromosome. This pattern of enrichment is not evident outside the gene. A burden test of association for rare variants within the gene might be statistically significant.

discover rare variants, confirming the presence of the variant in discovery samples with TaqMan or capillary electrophoresis sequencing is useful before exploring in independent samples to establish disease association.

Power Considerations and Significance Testing

One of the challenges to establishing a rare variant disease association is that in any given study, few variants are observed. Therefore, genetic studies are more poorly powered to detect a rare SNP association than they are to detect more common association with the same effect size (see Figure 3). Thus to detect associations at the same statistical threshold, sample collections larger than those currently used might be necessary. Establishing association of de novo or private mutations may not be possible at all because they may be seen only once in an entire study.

For rare variant associations, the field has not yet defined accepted standards for statistical significance that account for the burden of multiple hypothesis testing. Because there are many more rare variants than common ones, and they are not typically intercorrelated with each other, a more stringent threshold may be necessary than applied for common variants. One conservative approach is to correct for the total number of bases genome-wide, i.e., $p = 0.05/3 \times 10^9 \sim 10^{-11}$ as a significance threshold. Most recent studies have limited themselves to exomes or to a subset of targeted genes; in these instances the multiple-hypothesis testing burden might be significantly less. But with spectre of genetic studies with genome sequencing in the very near future, this conservative threshold may ultimately turn out to be appropriate.

Despite limitations in power and the need for achieving greater significance, rare variant associations with strong effects might be imminently detectable. For instance, as part of a genome-wide study, Holm et al. were able to identify a rare variant for sick sinus syndrome (Holm et al., 2011); the coding variant that explained the association was highly statistically significant in a modestly sized cohort as it had such a large effect size (odds ratio [OR] > 12). One strategy to further enhance the prospects of discovery is to identify those individuals most likely to have highly penetrant rare mutations. For example, individuals with younger onset or more severe disease, those with familial forms of disease, or those individuals that have disease despite a lack of other clinical or genetic risk factors might be promising candidates for rare variant association studies.

Burden Testing

If a genomic region is critical to disease pathogenesis, rare mutations may modulate disease susceptibility. Then, many affected individuals may have rare mutations more frequently in that region, though the mutations may be different from and unrelated to one another. This concept has sparked interest in the genetics community, and workers in statistical genetics have devised strategies to examine rare variants in aggregate across a target region (Bansal et al., 2010). These "burden" tests assess whether rare variants within a specific region are distributed in a nonrandom way, suggesting that they might be playing a role in disease pathogenesis (see Figure 3B). For example, a simple burden test might assess whether cases are enriched for rare variants compared to controls. More sophisticated tests account for the possibility that the region contains both protective and risk-conferring mutations. The target region might be a specific subregion of a gene, an entire gene transcript, or the entire genome.

This approach is an important alternative to the challenging task of establishing the association of individual rare variants; using these approaches to test multiple variants simultaneously might enhance power over testing individual variants. For instance, a burden test might be able to identify nonrandom distributions even of *private* mutations.

In an early application of rare variant burden testing, Cohen et al. examined individuals from the general population with high and low HDL levels and assessed the burden of rare variation in three candidate genes known to harbor Mendelian mutations that cause familial low serum high-density lipoprotein (HDL) levels (Cohen et al., 2004). They found that individuals with low HDL levels were significantly more likely to contain rare nonsynonymous mutations than those with high HDL levels; of the low HDL individuals, 16% had at least one rare mutation, compared to 2% of high HDL individuals. This suggested

strongly that for individuals with low HDL levels, ~14% of them may have mutations in these three genes mediating phenotype. The idea of comparing the proportion of case individuals with rare alleles to control individuals with rare alleles was formalized into a statistical test, the "Cohort Allelic Sums Test" (CAST) (Morgenthaler and Thilly, 2007). Subsequently, more sophisticated tests have been proposed that allow investigators to combine association testing of rare and common alleles by either testing for association together in multivariate tests (Li and Leal, 2008) or combining rare and common alleles weighted inversely to their allele frequency (Madsen and Browning, 2009).

One very powerful way of enhancing burden testing is to filter variants that are more likely to be causal from those that are likely not to be causal. For example, investigators may focus their studies on nonsynonymous alleles. Alternative approaches might include filtering variants based on sequence conservation properties or other bioinformatics approaches (Adzhubei et al., 2010; Ng and Henikoff, 2003).

A successful test, where statistical significance is obtained, can be used to argue that (1) the tested rare variants play a role in a specific disease, and (2) the target region tested plays an important role in disease pathology. But, it fails to implicate specific variants, and ambiguity about the causal variants might remain. For example, if rare variants are enriched in a gene 2-fold in cases compared to controls, then roughly half the variants seen in cases might be pathogenic, but the other half are part of the background distribution of rare variation in that gene and may not influence disease risk.

Structural Variants

Rare structural variants have gained recent interest; the frequency and size of structural variants have repeatedly shown enrichment in schizophrenia and other neuropscychiatric disease (International Schizophrenia Consortium, 2008; Sebat et al., 2007; Walsh et al., 2008). However, except for a few specific regions such as 22q11 and 16p11, most rare events have uncertain pathogenecity. For instance, although the rates of >100 kb deletion events are significantly increased in cases compared to controls, there is great uncertainty as to which individual events are pathogenic and which ones are nonpathogenic events that might occur in the general healthy population. This is analogous to the circumstance that might occur with a statistically significant burden test for point mutations, described above.

Extended Haplotypes

As previously discussed, many rare variants are recent and occur on extended haplotypes that can be identified using common variant markers. Thus GWAS datasets may be used to identify long-range haplotypes based on common markers and to then assess whether they are associated with phenotype. If this is the case, the phenotypic association might be driven by a highly penetrant rare variant. We used this approach to find an extended haplotype in the *CFH* gene that conferred high risk of age-related macular degeneration; subsequent sequencing identified the causal mutation to be an argenine to cysteine change in the C terminus of the protein (S.R. and J. Seddon, unpublished data).

This approach might be most effective in isolated populations where reduced genetic diversity and founder effects make it possible to identify long-range haplotypes (Kong et al., 2008). One recently published method to identify long and rare haplotypes, and to then test for association to phenotype, has been successfully applied to multiple phenotypes in out-bred populations (Gusev et al., 2011).

From Variants to Function

Translating rare and common variants to function can be challenging. In many instances the presence of an association does not clarify which variants are functionally causing disease susceptibility. For common variants, fine-mapping might be stymied by local LD. For rare variants, burden testing might be able to identify a genomic region enriched for rare variants but may not be able to specifically distinguish the individual causal rare variants from spurious nonpathogenic variants. Here we describe broad approaches that might be pursued to clarify pathogenic functions and causality, in the absence of genetic mapping that has clearly identified a single causal variant.

Evaluating Nonsynonymous Coding Variants

About 1% of the genome consists of protein-coding sequences. Variants in this portion of the genome are potentially the most amenable to follow up by biochemical characterization of the protein product in vitro, characterization in cell lines, or evaluation in transgenic model organisms. Only a minority of associated common variants can be explained by a nonsynonymous coding variant (~10%) (Hindorff et al., 2009). Currently, most studies of rare variation emphasize nonsynonymous coding variants; in many cases, noncoding variants are altogether ignored even if they are sequenced. An important challenge in the field is to prioritize discovered coding variants for potentially time-consuming functional follow up.

Computational approaches can be effective at assessing the degree to which a specific amino acid substitution in a protein, induced by a variant, might disrupt function. The functional impact of a substitution can often be estimated by using information about sequence conservation at the mutated site from comparative sequence analysis of a gene with orthologs and paralogs. If an amino acid site in a protein sequence is functionally critical, then most de novo mutations are deleterious and are subject to purifying selection; these sites then are expected to show little variation. Thus, a nonsynonymous allele from a study in a highly conserved site is likely to be deleterious. Sequence conservation in organisms more closely related to human is particularly informative because more distantly related organisms may have divergent biology and protein function. Many software tools using these principles to assess coding variants have now been devised (Cooper and Shendure, 2011). One example of such a program is Polymorphism Phenotyping 2 (or PolyPhen 2) (Adzhubei et al., 2010). The most predictive features in this method are the estimated likelihood that the mutant allele fits the substitution pattern observed in the multiple-sequence alignment; the evolutionary distance to the organism with a protein harboring a similar nonsynonymous substitution; and whether the mutant allele occurs at a site that is hypermutable. The method uses these features and others, including information from the three-dimensional protein structure, to define a statistical model that includes the probability of disease based on a catalog of known pathogenic Mendelian mutations. The functional importance of an amino acid replacement is predicted

from these features based on a naive Bayes classifier. PolyPhen 2 and other related methods demonstrate similar performance in their ability to predict pathogenic mutations achieving an area under the curve (AUC) of 75%–80% (Hicks et al., 2011).

Experimental approaches to individually interrogate rare variants with functional assays can also be very powerful. But, for an approach to be effective, it is critical that the functional assay is high throughput, and that it has an assayed function that is relevant to the phenotype. Otherwise, mutations that affect the assayed gene function might not in fact be pathogenic. In one application of this approach, Davis et al. used it to look at individual mutations with the *TTC21B* gene and to show that they cause human ciliopathies (Davis et al., 2011). First they demonstrated that a translation-blocking morpholino specific for *TTC21B* resulted in gastrulation defects in zebrafish that were consistent with cilliary dysfunction. Then, when they resequenced *TTC21B* in a large, clinically diverse ciliopathy cohort and matched controls, they observed a similar frequency of rare variants. But, when they tested those rare alleles to identify those that caused gastrulation defects in zebrafish, they observed a significant enrichment of functional alleles in cases compared to controls.

Evaluating Noncoding Variants

Noncoding variants pose a particular challenge to the field at the moment. The noncoding genome represents 99% of the genome and at present is poorly annotated (Alexander et al., 2010). About 10% of the noncoding genome is under purifying selection, suggesting that they harbor critical processes that if disrupted could be pathogenic (Davydov et al., 2010). Many common variants, if they contribute to disease, likely act by impacting the noncoding genome. As one example, an associated Crohn's disease SNP in LD with polymorphic deletion overlaps the *IRGM* gene promotor and modulates gene expression (McCarroll et al., 2008). In the last several years, however, several promising approaches have emerged to evaluate noncoding variants that might point the way to causality, such as analyzing sequence conservation, gene expression, and chromatin state.

Sequence Conservation

A computational approach to prioritizing noncoding variants is to identify those that are at sites with a high degree of sequence conservation across mammalian organisms and are thus under purifying negative selection (Cooper et al., 2005; Miller et al., 2007). These approaches differ from those approaches used to prioritize coding substitutions, as they can only use nucleotide sequence similarity. Indeed, investigators have argued that the conservation information from nucleotide sequences is as predictive as the information gained by peptide sequence similarity and protein structural features (Cooper et al., 2010). The value of assessing common variants with sequence conservation approaches is uncertain, as common variants are presumably not under purifying negative selection. But, rare noncoding variants that have dramatic effects on disease susceptibility might be effectively prioritized with this approach.

eQTL Data Can Suggest Causal Genes and Mechanism

Expression quantitative loci (eQTL) are genetic variants that correlate with the transcript level of a gene (Jansen and Nap, 2001). To date, most reported eQTLs are *cis*-effects, acting on nearby genes by encoding variants that modulate promotor activity, enhancer activity, or mRNA stability. Expression QTL acting in *trans* have been largely unexplored thus far. Although most recently discovered eQTL have been common variants, there is evidence of rare eQTL also (Montgomery et al., 2011). Identifying rare eQTL might be challenging given the limited power of currently sized cohorts. In the future, burden tests previously described might be able to effectively identify small genomic regions where rare variants dramatically impact transcript levels.

It has been shown that common trait-associated variants have a significant overlap with eQTL, suggesting the possibility that many common disease variants act by altering transcript levels (Nicolae et al., 2010). Thus, it might be insightful to assess whether a specific disease-associated common variant is itself an eQTL. If it is, then the gene whose transcript is influenced by the risk allele might be the causal gene. Furthermore, if the risk allele is increasing the transcript level, then the gene may increase disease risk by magnifying gene function; alternatively, if the risk allele reduces transcript level, then the gene may cause disease by mitigating gene function. A convincing eQTL effect can be isolated by transfecting constructs with risk haplotype fragments, as was done to identify the causal variant in the *SORT1* lipid locus (Musunuru et al., 2010). Another compelling example of an eQTL that influences disease susceptibility is a type II diabetes-associated variant upstream of the *KLF14* transcription factor. Investigators showed that this variant acts not only as a *cis*-eQTL influencing *KLF14* levels in adipose tissue but also as a *trans*-eQTL for many genes regulated by *KLF14* that are important in metabolic traits (Small et al., 2011).

There are a few important caveats about this seemingly straightforward approach.

First, because eQTL are spread throughout the genome, spurious overlap between disease-associated variants and eQTL is possible (Nica et al., 2010). If a risk variant confers risk by modulating transcript levels, and it is itself causal (or in LD with the causal variant), then it should also be consistent with the strongest eQTL effect in the region. Checking to ensure that the disease-associated variant is consistent with the strongest eQTL effect itself mitigates the risk of spurious overlap. However, it is still possible that the causal allele and the strongest eQTL effect are strongly correlated by chance, and that eQTL association is unrelated to disease risk.

Second, although many eQTL act generically, most are tissue specific (Dimas et al., 2009; Price et al., 2011). In fact, certain eQTL may not be detectable unless the cell has responded to a specific stimulus or stress. In order to understand the transcriptional impact of disease alleles most effectively, identifying eQTL in the pathogenic tissues is key. Current eQTL databases are based on a small number of resting cell types, for example lymphoblastoid cell lines (Stranger et al., 2007). Many important pathogenic tissues are not easily accessible for eQTL studies. In the near future, the catalog of available tissues profiled will expand dramatically with the NIH-sponsored Genotype Tissue Expression (GTEx) project, aiming to profile >60 separate tissues (https://commonfund.nih.gov/GTEx/).

Finally, although eQTL data can offer potential in identifying the likely causal gene and provide hints about mechanism for common variants, they may not clarify ambiguity about the

causal variant if there are multiple variants in LD. Certain variants may seem more promising, for example structural variants or SNPs overlapping a regulatory variant. As with disease-associated common variants, eQTL datasets often face challenges in fine-mapping signals.

Chromatin Modifications

Identifying regions of the genome that act as regulatory elements can offer important complementary information to eQTL data in evaluating noncoding variants. Specific functional regulatory elements can be identified from genome-wide profiles of key histone modifications: H3K4me3 marks active promoters; H3K4me1 marks enhancers; H3K4me2 and most histone acetylation mark both promoters and enhancers (Barski et al., 2007; Heintzman et al., 2007; Wang et al., 2008). Similarly, DNase I hypersensitive sites also flag open chromatin regions harboring promoters and enhancers (Sabo et al., 2006). With the advancement of high-throughput sequencing technologies and development of techniques such as ChIP-seq (Park, 2009) and DNase-seq (John et al., 2011), there are mounting public data on genome-wide chromatin profiles. For instance, histone mark ChIP-seq and DNase-seq data on over 100 cell lines and tissues have now been generated through the ENCODE and Roadmap Epigenomics projects (Bernstein et al., 2010; Birney et al., 2007).

Although computational approaches to identify putative binding sites based on sequence data alone are nonspecific, recent reports suggest that the prediction of active regulatory sites within assayed tissues is possible by including ChIP-seq and DNase-seq data (Ernst and Kellis, 2010; He et al., 2010; Pique-Regi et al., 2011; Song et al., 2011). One potential approach then to prioritize noncoding variants for follow up is to identify those that are in regions that have been predicted to be regulatory elements. These variants might, for example, disrupt or enhance a transcription factor binding at an enhancer or a promoter. Particularly promising variants might be those that have eQTL activity in the same cell type. Histone mark locations and DNase hypersensitive sites have been shown to be enriched near associated variants (Ernst et al., 2011; McDaniell et al., 2010). A key limitation of this approach is that, like eQTL data, it requires genome-wide chromatin data from the same or similar cell types as those that are pathogenic.

Identifying Causal Processes with Integrative Analyses

In many instances where the specific causal variant within a locus cannot be identified, examination of the genes implicated may still help to suggest the key underlying functional networks and pathways that might be active in a disease. For instance, age-related macular degeneration associations have implicated the complement pathway without necessarily identifying causal variants. This task can be challenging in general because for any given associated allele, 20 or more genes might be implicated by LD, and any of them may harbor the causal mutation.

But despite that, statistically significant connectivity between genes in different associated loci can often be identified. We and others have devised strategies to look for functional connections or similarity between genes across implicated loci. These networks can predict novel gene loci and offer insight about disease mechanism. Gene Relationships Across Implicated Loci (GRAIL) uses >400,000 published scientific PubMed texts to assess pairwise gene similarity between genes across loci (Raychaudhuri et al., 2009a). In addition to repeatedly showing highly statistically significant connectivity between genes across loci in multiple diseases, GRAIL has been used to prospectively predict and prioritize associated variants (Raychaudhuri et al., 2009b) and prioritize disease genes within a locus (Beroukhim et al., 2010). Investigators used a similar approach, Disease Association Protein-Protein Link Evaluator (DAPPLE) algorithm, to demonstrate that protein-protein interactions are enriched among genes within disease loci more than by chance alone (Rossin et al., 2011). They demonstrated enrichment most convincingly in autoimmune diseases and furthermore demonstrated that the enrichment of interactions was often between genes within the same immune cell types. These networks offer insight as to how protein products of genes across many loci might be interacting together to initiate disease. We note importantly that pathway analyses can be easily confounded, in particular in neuropsychiatric diseases because there is a correlation between the sizes of transcripts and the likelihood that they will have brain function (Raychaudhuri et al., 2010).

Conclusions

The advances in genotyping and sequencing technologies over the last few years have revolutionized genetics. Only a few years ago, researchers were still tackling the challenges of gene mapping and discovery of complex diseases. Now we face an embarrassment of riches in which the ability to map loci has become quick and reproducible. The next important challenge is streamlining functional validation, which in most cases is still a critical bottleneck. Rare variant discovery has the potential to yield more obviously functional variants with larger effect sizes because they are less constrained by purifying selection. The discovery of rare variant associations might shed light on those loci discovered by common variant mapping. However, strategies to prioritize functional follow-up studies will be key at those loci where common variants cannot be effectively fine-mapped or individual rare variants (beyond the presence of case enrichment) cannot be identified. Strategies to use regulatory variants, chromatin state data, and sequence conservation offer a potential path forward to prioritize candidate variants.

SUPPLEMENTAL INFORMATION

Supplemental Information includes one figure and can be found with this article online at doi:10.1016/j.cell.2011.09.011.

ACKNOWLEDGMENTS

The author would like to acknowledge helpful discussions and feedback from colleagues including Drs. Mark Daly, Paul I.W. de Bakker, X. Shirley Liu, Cynthia Sandor, Eli A. Stahl, Barbara E. Stranger, and Shamil Sunyaev.

REFERENCES

1000 Genomes Project Consortium. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

Adrianto, I., Wen, F., Templeton, A., Wiley, G., King, J.B., Lessard, C.J., Bates, J.S., Hu, Y., Kelly, J.A., Kaufman, K.M., et al; BIOLUPUS and GENLES Networks. (2011). Association of a functional variant downstream of TNFAIP3 with systemic lupus erythematosus. Nat. Genet. *43*, 253–258.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

Alexander, R.P., Fang, G., Rozowsky, J., Snyder, M., and Gerstein, M.B. (2010). Annotating non-coding regions of the genome. Nat. Rev. Genet. *11*, 559–571.

Anderson, C.A., Soranzo, N., Zeggini, E., and Barrett, J.C. (2011). Synthetic associations are unlikely to account for many common disease genome-wide association signals. PLoS Biol. *9*, e1000580.

Balding, D.J. (2006). A tutorial on statistical methods for population association studies. Nat. Rev. Genet. *7*, 781–791.

Bansal, V., Libiger, O., Torkamani, A., and Schork, N.J. (2010). Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet. *11*, 773–785.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., et al. (2010). The landscape of somatic copy-number alteration across human cancers. Nature *463*, 899–905.

Bhangale, T.R., Rieder, M.J., and Nickerson, D.A. (2008). Estimating coverage and power for genetic association studies using near-complete variation data. Nat. Genet. *40*, 841–843.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Browning, S.R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. Hum. Genet. *124*, 439–450.

Campbell, C.D., Ogburn, E.L., Lunetta, K.L., Lyon, H.N., Freedman, M.L., Groop, L.C., Altshuler, D., Ardlie, K.G., and Hirschhorn, J.N. (2005). Demonstrating stratification in a European American population. Nat. Genet. *37*, 868–872.

Cardon, L.R., and Abecasis, G.R. (2003). Using haplotype blocks to map human complex trait loci. Trends Genet. *19*, 135–140.

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., et al. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat. Genet. *37*, 1243–1246.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869–872.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al; Wellcome Trust Case Control Consortium. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.

Conrad, D.F., Keebler, J.E., DePristo, M.A., Lindsay, S.J., Zhang, Y., Casals, F., Idaghdour, Y., Hartl, C.L., Torroja, C., Garimella, K.V., et al; 1000 Genomes Project. (2011). Variation in genome-wide mutation rates within and between human families. Nat. Genet. *43*, 712–714.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J., and Nickerson, D.A. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat. Methods *7*, 250–251.

Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A.; NISC Comparative Sequencing Program. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. *15*, 901–913.

Davis, E.E., Zhang, Q., Liu, Q., Diplas, B.H., Davey, L.M., Hartley, J., Stoetzel, C., Szymanska, K., Ramaswami, G., Logan, C.V., et al; NISC Comparative Sequencing Program. (2011). TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. Nat. Genet. *43*, 189–196.

Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025.

de Bakker, P.I., Ferreira, M.A., Jia, X., Neale, B.M., Raychaudhuri, S., and Voight, B.F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum. Mol. Genet. *17*(R2), R122–R128.

Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H., and Goldstein, D.B. (2010). Rare variants create synthetic genome-wide associations. PLoS Biol. *8*, e1000294.

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science *325*, 1246–1250.

Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Ge, D., Fellay, J., Thompson, A.J., Simon, J.S., Shianna, K.V., Urban, T.J., Heinzen, E.L., Qiu, P., Bertelsen, A.H., Muir, A.J., et al. (2009). Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. Nature *461*, 399–401.

Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., et al. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. Science *329*, 841–845.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., and Chee, M.S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. *37*, 549–554.

Gusev, A., Kenny, E.E., Lowe, J.K., Salit, J., Saxena, R., Kathiresan, S., Altshuler, D.M., Friedman, J.M., Breslow, J.L., and Pe'er, I. (2011). DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. *88*, 706–717.

Harismendy, O., Ng, P.C., Strausberg, R.L., Wang, X., Stockwell, T.B., Beeson, K.Y., Schork, N.J., Murray, S.S., Topol, E.J., Levy, S., and Frazer, K.A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. *10*, R32.

He, H.H., Meyer, C.A., Shin, H., Bailey, S.T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., et al. (2010). Nucleosome dynamics define transcriptional enhancers. Nat. Genet. *42*, 343–347.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.

Hicks, S., Wheeler, D.A., Plon, S.E., and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum. Mutat. *32*, 661–668.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional

implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. USA *106*, 9362–9367.

Hoggart, C.J., Clark, T.G., De Iorio, M., Whittaker, J.C., and Balding, D.J. (2008). Genome-wide significance for dense SNP and resequencing data. Genet. Epidemiol. *32*, 179–185.

Holm, H., Gudbjartsson, D.F., Sulem, P., Masson, G., Helgadottir, H.T., Zanon, C., Magnusson, O.T., Helgason, A., Saemundsdottir, J., Gylfason, A., et al. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. Nat. Genet. *43*, 316–320.

International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature *455*, 237–241.

Jansen, R.C., and Nap, J.P. (2001). Genetical genomics: the added value from segregation. Trends Genet. *17*, 388–391.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat. Genet. *43*, 264–268.

Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., et al. (2008). Detection of sharing by descent, long-range phasing and haplotype imputation. Nat. Genet. *40*, 1068–1075.

Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am. J. Hum. Genet. *80*, 727–739.

Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. *11*, 473–483.

Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.

Maxmen, A. (2011). Exome sequencing deciphers rare diseases. Cell *144*, 635–637.

McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H., et al. (2008). Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat. Genet. *40*, 1107–1112.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science *328*, 235–239.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Miller, W., Rosenbloom, K., Hardison, R.C., Hou, M., Taylor, J., Raney, B., Burhans, R., King, D.C., Baertsch, R., Blankenberg, D., et al. (2007). 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. Genome Res. *17*, 1797–1808.

Momozawa, Y., Mni, M., Nakamura, K., Coppieters, W., Almer, S., Amininejad, L., Cleynen, I., Colombel, J.F., de Rijk, P., Dewit, O., et al. (2011). Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. Nat. Genet. *43*, 43–47.

Montgomery, S.B., Lappalainen, T., Gutierrez-Arcelus, M., and Dermitzakis, E.T. (2011). Rare and common regulatory variation in population-scale sequenced human genomes. PLoS Genet. *7*, e1002144.

Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature *466*, 714–719.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J.A. (2009). Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science *324*, 387–389.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814.

Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D., Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. *42*, 30–35.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. *6*, e1000895.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet. *6*, e1000888.

Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Peltonen, L., Palotie, A., and Lange, K. (2000). Use of population isolates for mapping complex traits. Nat. Rev. Genet. *1*, 182–190.

Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I., Walker, B.D., Ripke, S., Brumme, C.J., Pulit, S.L., Carrington, M., et al; International HIV Controllers Study. (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science *330*, 1551–1557.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447–455.

Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet. *7*, e1001317.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? Am. J. Hum. Genet. *69*, 124–137.

Pritchard, J.K., and Cox, N.J. (2002). The allelic architecture of human disease genes: common disease-common variant…or not? Hum. Mol. Genet. *11*, 2417–2423.

Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P.; International Schizophrenia Consortium. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature *460*, 748–752.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C.Y., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D., and Daly, M.J.; International Schizophrenia Consortium. (2009a). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. PLoS Genet. *5*, e1000534.

Raychaudhuri, S., Thomson, B.P., Remmers, E.F., Eyre, S., Hinks, A., Guiducci, C., Catanese, J.J., Xie, G., Stahl, E.A., Chen, R., et al; BIRAC Consortium; YEAR Consortium. (2009b). Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. Nat. Genet. *41*, 1313–1318.

Raychaudhuri, S., Korn, J.M., McCarroll, S.A., Altshuler, D., Sklar, P., Purcell, S., and Daly, M.J.; International Schizophrenia Consortium. (2010). Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet. *6*, e1001097.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends Genet. *17*, 502–510.

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

Rosenberg, N.A., Huang, L., Jewett, E.M., Szpiech, Z.A., Jankovic, I., and Boehnke, M. (2010). Genome-wide association studies in diverse populations. Nat. Rev. Genet. *11*, 356–366.

Rossin, E.J., Lage, K., Raychaudhuri, S., Xavier, R.J., Tatar, D., Benita, Y., Cotsapas, C., Daly, M.J., and Daly, M.J.; International Inflammatory Bowel Disease Genetics Constortium. (2011). Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet. *7*, e1001273.

Sabo, P.J., Kuehn, M.S., Thurman, R., Johnson, B.E., Johnson, E.M., Cao, H., Yu, M., Rosenzweig, E., Goldy, J., Haydock, A., et al. (2006). Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. Nat. Methods *3*, 511–518.

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., et al. (2007). Strong association of de novo copy number mutations with autism. Science *316*, 445–449.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nat. Biotechnol. *26*, 1135–1145.

Small, K.S., Hedman, A.K., Grundberg, E., Nica, A.C., Thorleifsson, G., Kong, A., Thorsteindottir, U., Shin, S.Y., Richards, H.B., Soranzo, N., et al; GIANT Consortium; MAGIC Investigators; DIAGRAM Consortium; MuTHER Consortium. (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat. Genet. *43*, 561–564.

Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., Lee, B.K., Sheffield, N.C., Gräf, S., Huss, M., Keefe, D., et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. Published online August 19, 2011. 10.1101/gr.121541.111.

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224.

Teslovich, T.M., Musunuru, K., Smith, A.V., Edmondson, A.C., Stylianou, I.M., Koseki, M., Pirruccello, J.P., Ripatti, S., Chasman, D.I., Willer, C.J., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. Nature *466*, 707–713.

Thorgeirsson, T.E., Geller, F., Sulem, P., Rafnar, T., Wiste, A., Magnusson, K.P., Manolescu, A., Thorleifsson, G., Stefansson, H., Ingason, A., et al. (2008). A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. Nature *452*, 638–642.

Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., Thorleifsson, G., et al; MAGIC investigators; GIANT Consortium. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat. Genet. *42*, 579–589.

Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A., et al. (2008). Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science *320*, 539–543.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

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.