APPENDIX 3: The Genetic Basis of Disease

INTRODUCTION

The average human has approximately 4 million germline variants, that is, positions in their genome that differ from the reference genome (see Appendix 2 for a discussion of the reference genome). For the most part, this variation is not harmful and makes us distinct individuals, but some variants have been linked to diseases and health-related traits. The presence of some disease-causing variants means a high degree of certainty that the person with the variant will develop the disease. This is true for so-called Mendelian diseases. For other disease-causing variants, the relationship with disease is complex and the variants not very predictive of disease. Establishing the relationship between a genetic variant and disease is accomplished through different research methods. Mendelian and complex diseases differ in terms of the location of disease-causing variants, their frequency in the population, their penetrance, and their utility as diagnostic/predisposition tests.

MENDELIAN GENETIC DISORDERS

Mendelian genetic disorders are individually rare, but in aggregate are present in about 2–3% of all newborns, although disease may not be manifest for years to decades, or ever. They are characterized by a clear pattern of inheritance, typically autosomal dominant, autosomal recessive, or X-linked.

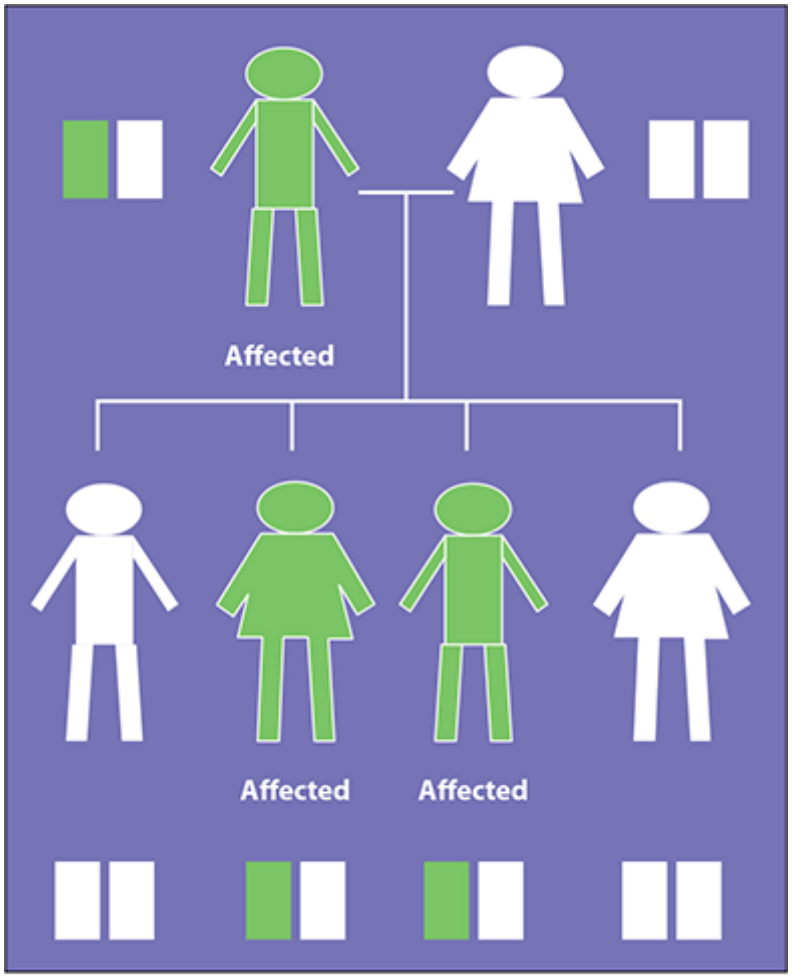
Modes of inheritance

Autosomal dominant

A disease-causing variant in one of the two copies of a gene is sufficient to cause disease. Typical inheritance is from one affected parent who passes the variant to 50% of his or her offspring (Fig. A3-1). These tend to be diseases that occur in adulthood. Both sexes are affected equally. Some common examples include familial hypercholesterolemia (1 in 200), hereditary breast and ovarian cancer (1 in 400), and Lynch syndrome (1 in 400).

Figure A3-1.

Pedigree illustrating autosomal dominant inheritance. The most common situation is one affected parent carrying the disease mutation and passing it on to half of their offspring, who also become affected by the disease.



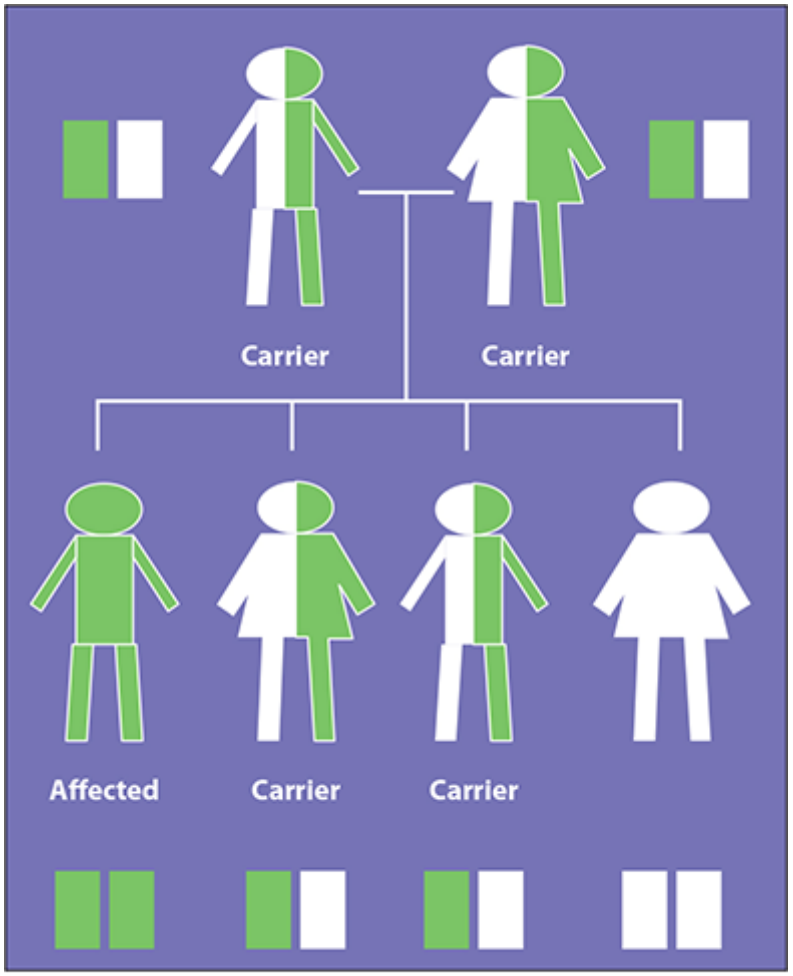
Autosomal recessive

Autosomal recessive diseases are the result of two unaffected parents who are each carriers of a mutation in one copy of the gene, and disease results when their child inherits both abnormal copies (Fig. A3-2). These may be two copies of the same variant (homozygote) or two different

variants (compound heterozygote). Individuals with only one copy of a disease-causing variant are carriers, but will not develop the disease. When two carriers reproduce, one quarter of their offspring are expected to inherit both disease-causing variants and will manifest the disease, one quarter of offspring inherit two normal copies of the gene, and half will inherit just one and be carriers. Recessive diseases affect both sexes equally and are more prevalent in consanguineous populations. Carrier rates are highest in the United States for sickle cell anemia (1 in 12 African-Americans)[1], cystic fibrosis (1 in 29 Caucasian Americans)[2], and Tay–Sachs (1 in 27 Ashkenazi Jews)[3], although the incidence of these diseases has dropped due to effective genetic screening programs.

Figure A3-2.

Pedigree illustrating autosomal recessive inheritance. The most common situation is two unaffected parents, each carrying a single disease mutation that is passed on to half of their offspring. By chance, 25% of those offspring will inherit both disease mutations and become affected.



Source: Jeanette J. McCarthy, Bryce A. Mendelsohn: *Precision Medicine: A Guide to Genomics in Clinical Practice*
www.accessmedicine.com

Contrary to conventional wisdom, carriers of one copy of a disease-causing variant for a recessive trait are not always unaffected. Some examples: carriers of a Gaucher disease variant are at increased risk of Parkinson disease[4]; carriers of a cystic fibrosis variant are at increased risk of pancreatitis and other disorders[5]; Carriers of a sickle cell anemia variant are at increased risk of exertional rhabdomyolysis and other disorders[6].

X-linked

X-linked inheritance is a special case. Males have only one X chromosome, while females have two. A variant on the X chromosome in a male affects his only copy, which is sufficient to cause disease (it behaves in a dominant fashion). In females, who have two X chromosomes, it behaves more like a recessive where carriers of a single mutation are unaffected. This explains why X-linked traits are more common in males than females. This situation can become even more complicated since one X chromosome, chosen at random, is inactivated, in females. Sometimes because of different patterns of X inactivation in different tissues, female carriers may be symptomatic. Examples of diseases where female carriers of X-linked mutations can frequently be affected (though not necessarily to the same degree as males) include fragile X, Fabry disease, and X-linked adrenoleukodystrophy.

Oddly, some X-linked diseases present only in females. This is usually because the disorder is lethal at an embryonic stage in males. An example is Rett syndrome, caused by mutations in *MECP2*.

Mitochondrial

The cell's powerhouses, the mitochondria, contain their own very small genome consisting of 37 genes, mostly involved in the respiratory chain. Mutations in mitochondrial genes mostly affect cells and organs that have the highest energy consumption, leading to a wide range of disorders, primarily affecting the neuromuscular system. Mitochondrial DNA (mtDNA) is inherited independently of nuclear DNA and only from the mother (the mitochondrial genome is carried in the ova but not the sperm). Mitochondrial disorders due to mutations in the mtDNA can affect males and females equally, but will be transmitted only from the mother. Importantly, most of the proteins present in the mitochondria are encoded in the nuclear genome, meaning that most mitochondrial diseases are not inherited from the mother only and are often recessive.

Because each mitochondrion contains several copies of its genome, and each cell contains many mitochondria, meaning a cell, tissue, or individual can be a mixture of different mitochondrial genomes, a concept termed *heteroplasmy*. Heteroplasmy can cause two individuals with the same change in the mitochondrial genome to have different symptoms determined by which tissues have the abnormal mitochondria, and can cause a woman to pass different amounts of abnormal mitochondrial to her offspring, leading to different symptoms.

De novo

While not inherited, per se, de novo mutations that occur in the sperm or the egg are emerging as an important cause of Mendelian disease in patients with healthy parents. De novo disease-causing mutations can cause dominant or X-linked diseases and can contribute to recessive diseases as well. In all cases, the de novo variant will not be found in either parent, by definition.

Most de novo mutations occur in the egg or sperm that gives rise to a new person and appear in every cell in that person. De novo mutations can also occur after fertilization, in which case the individual will be a mosaic of cells that do and do not harbor a mutation. Mosaicism can range from involving most of the body to only a small fraction of cells.

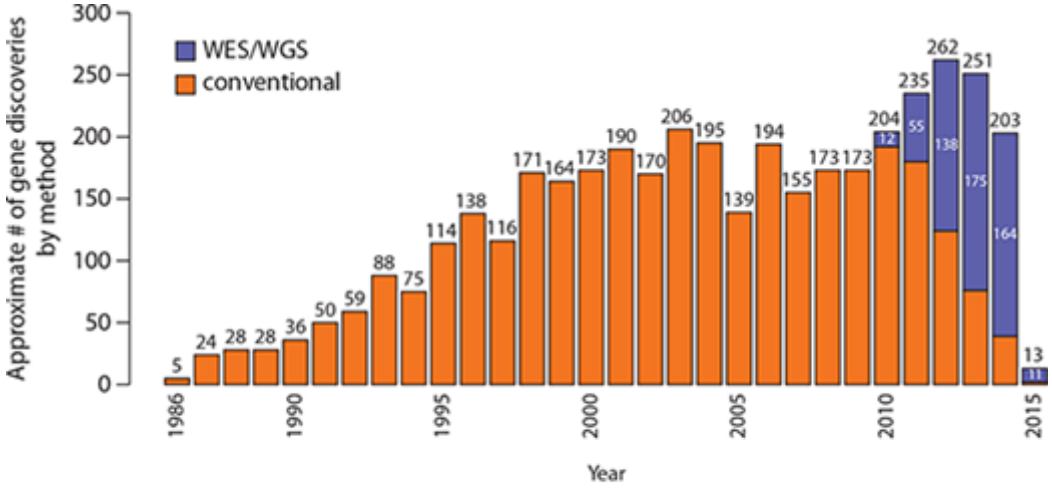
Methods to discover Mendelian disease genes

Linkage analysis and positional cloning

Prior to the sequencing of the human genome, Mendelian genetic diseases were typically discovered using a laborious method called positional cloning. This method saw its first success in 1986 with the cloning of the gene for X-linked chronic granulomatous disease[7]. Positional cloning takes advantage of multigenerational pedigrees with multiple affected and unaffected individuals analyzed across their genomes for a standard panel of highly variable genetic variants (*markers*). These genetic markers are not thought to be disease-causing variants themselves, but rather, serve to mark, or indicate, specific locations in the genome. Linkage analysis involves examining each of these genetic markers to see how well presence of specific marker alleles cosegregate with (i.e., travel with) the disease in the family. Complete cosegregation occurs when the marker allele is present in all affected individuals and absent in all unaffected individuals. The degree to which cosegregation occurs can be statistically measured with a statistic called an LOD score. Genetic markers that exhibit high LOD scores are said to be linked to the disease (more technically, they are linked to the gene underlying the disease). Linkage analysis narrows down the location of a disease gene to a chromosomal region, but further genetic analysis is required to narrow the region down even further, and then to identify candidate genes in that region and confirm their role in disease.

Figure A3-3.

Number of Mendelian disease genes discovered by year. The number of Mendelian disease genes discovered per year has grown, as has the relative proportion of those discoveries made using next-generation sequencing. (Reproduced with permission from Chong JX, Buckingham KJ, Jhangiani SN, et al: The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities, Am J Hum Genet. 2015 Aug 6;97(2):199-215).



Source: Jeanette J. McCarthy, Bryce A. Mendelsohn: *Precision Medicine: A Guide to Genomics in Clinical Practice*
www.accessmedicine.com

Whole-genome/exome sequencing

The ability to sequence human genomes relatively quickly and cheaply has revolutionized the discovery of Mendelian disease genes. Since 2009, next-generation sequencing has led to a dramatic rise in the number of novel disease-causing genes discovered (Fig. A3-3). The most common approach involves sequencing the whole genomes or exomes of a trio, including an affected individual and his or her parents. A series of data filtering steps based on mutation type, allele frequency, and mode of inheritance narrows down the list of possible candidate genes. These genes are reviewed in depth using information from public databases and the scientific literature to identify the most plausible candidate.

Characteristics of genetic variants associated with Mendelian diseases

Of the approximately 8000 Mendelian traits reported in the Online Mendelian Inheritance in Man (OMIM) database (http://www.ncbi.nlm.nih.gov/omim), the underlying gene is known for over half of them. Sixteen percent of genes in the human genome (3038) are known to underlie a Mendelian trait.[1]

Key Point

What about the other 84% of genes not linked to a Mendelian trait? Some may cause conditions so mild they are not recognized. Many are likely essential to life and no individual can be born without the function of these genes. Some genes may only cause disease under very specific

environmental conditions (e.g., infection with a particular pathogen) and have therefore eluded discovery. What is clear is that much remains to be learned.

Rare, protein-coding variants

Most genetic variants associated with Mendelian diseases are very rare (<0.1% population allele frequency) protein-altering changes in the coding region of the gene (e.g., truncating mutations, splice site variants, or missense changes), and most (but certainly not all) of these changes cause the gene to not function or function poorly. Less commonly, rare protein-coding variants cause a gene to have extra or new function(s) and these are referred to as gain of function mutations.

[1] Stats and figure from Centers for Mendelian Genomics: http://www.mendelian.org/about-mendelian-conditions.

Allelic heterogeneity

For a given Mendelian disease gene, there may be more than one variant that can lead to the disease. In some cases, there is one predominant disease-causing variant found in most individuals, and a long tail of rare variants in the gene that round out the population of pathogenic variants. For example, the cystic fibrosis gene, *CFTR*, has one predominant allele, delta F508, found in 85% of cases, but hundreds of additional rare pathogenic variants that also cause cystic fibrosis have been described in the human population.

High penetrance

Mendelian diseases are characterized by having a high penetrance, or probability of developing disease in the presence of a pathogenic variant. However, exceptions to this rule are becoming more evident as more healthy individuals are having their genomes sequenced.

COMPLEX DISEASES

Most diseases are complex, the result of the interplay between genes and environment. Complexity arises from the way that disease is defined, typically based on clinical signs and symptoms or biochemical features, crude measures that might not capture the root molecular cause. Any one clinically defined disease may have forms that are primarily sporadic (environmental in cause), and other forms that are primarily due to inherited gene variants and still others that are due to the interplay of genes and environment. Take leukemia for example, where there are rare Mendelian forms due to inherited germline mutations in *TP53* (e.g., Li–Fraumeni syndrome.) There are also sporadic or environmental forms, like the leukemia that developed in atomic bomb survivors exposed to radiation. However, most leukemia is probably due to the cumulative effects of

common genetic variants, each individually conferring only a modest increased risk, along with environmental factors. On the clinical level, nongenetic forms of the disease are indistinguishable from genetic forms, making it difficult to discern clear Mendelian inheritance patterns.

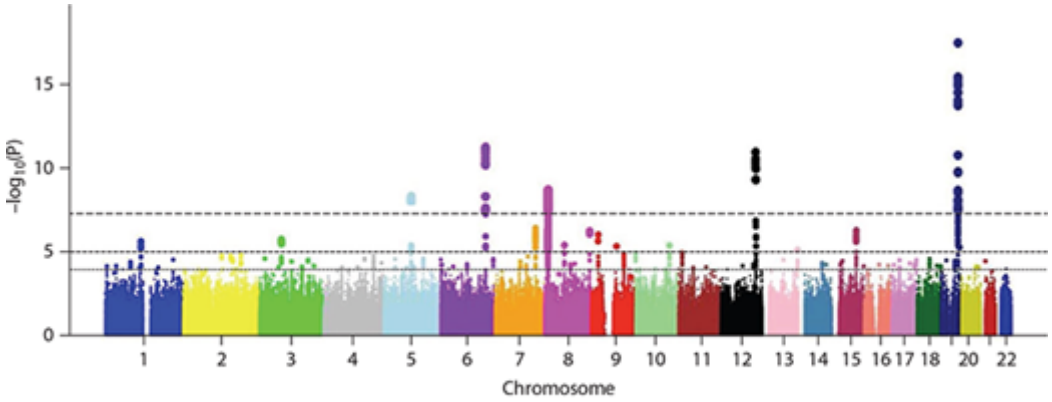
Methods to discover genes underlying common, complex diseases

Complex disease gene discovery is typically carried out among large cohorts of subjects that have been characterized with respect to their disease status (affected or unaffected). Subjects with the disease are compared to those without, with respect to the presence of genetic variation. Statistical tests of association are performed to see if the frequency of a genetic variant is different between the two groups. The significance of that association is expressed as a *p* value, and the strength measured in terms of an odds ratio or relative risk.

How does one select which genetic variants to test in an association study? Earlier attempts to identify the genetic basis of complex diseases relied on candidate gene studies, where individual genes likely to play a role in the disease were queried for genetic variation that might be associated with disease. In the early 2000s, genome-wide association studies (GWAS) were introduced as an unbiased way to query the entire genome. By testing hundreds of thousands of variants located throughout the genome, and relying on inherent correlation between variants (linkage disequilibrium), researchers could pinpoint a much smaller region of the genome that might harbor a disease gene. Results of GWAS are often visualized using a Manhattan plot depicting the most statistically significant findings as peaks in a skyline (Fig. A3-4). The type of variation analyzed in these studies is common single nucleotide variants, also called SNPs (single nucleotide polymorphisms). GWAS have dominated the research landscape, identifying dozens of robust associations across myriad diseases and traits. A catalogue of SNPs significantly associated with diseases from GWAS studies is maintained by the NHGRI-EBI (http://www.ebi.ac.uk/gwas/).

Figure A3-4.

Manhattan plot depicting several strongly associated variants from a genome-wide association study (Reproduced with permission from Ikram MK, Sim X, Jensen RA, et al: Four novel Loci (19q13, 6q24, 12q24, and 5q14) influence the microcirculation in vivo, *PLoS Genet*. 2010 Oct 28;6(10):e1001184).



Characteristics of complex disease genes

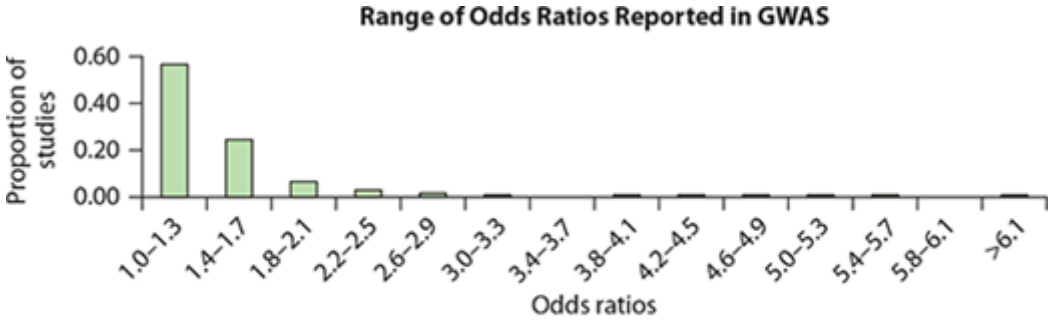
Common noncoding variants

The vast majority (about 90%) of the genetic variants underlying complex diseases are common variants (population allele frequency >5%) located in the noncoding region of the genome (introns and intergenic regions). These variants are concentrated in regulatory regions, suggesting that they impact disease via altering the expression of genes, rather than altering the protein sequence[8].

Low penetrance

For most common complex diseases, the penetrance of variants associated with the disease is low. The average effect size for a common variant is about a 1.18-fold increased odds of disease (Fig. A3-5). This is a very modest effect, making these variants individually uninformative as diagnostic/predisposition tests.

Figure A3-5.
Average effect size (in terms of odds ratio) of variants associated with diseases/traits in genome-wide association studies.



CHARACTERISTICS COMMON TO BOTH MENDELIAN AND COMMON COMPLEX DISEASES

Genetic heterogeneity

Both Mendelian and common complex diseases can exhibit genetic heterogeneity, where deleterious mutations in any one of a number of genes may lead to the same disease. For example, dozens of genes have been implicated in Mendelian forms of nonsyndromic hearing impairment (http://www.ncbi.nlm.nih.gov/books/NBK1434/). For common diseases, GWAS studies typically reveal dozens of genomic regions associated with the disease (https://www.ebi.ac.uk/gwas/).

Pleiotropy

Pleiotropy refers to the situation where a single genetic factor influences multiple seemingly distinct traits. Pleiotropy can occur at the level of the gene (i.e., different variants in the same gene are associated with more than one trait) or at the level of the variant (i.e., same variant is associated with more than one trait). About a quarter of known Mendelian disease genes cause two or more different traits (http://www.mendelian.org/about-mendelian-conditions). Inherited cancer genes are increasingly shown to predispose individuals to more than one type of cancer (see Chapter 7). Complex diseases also show evidence of pleiotropic effects, especially across different psychiatric diseases and autoimmune diseases[9]. Pleiotropy is something health care providers should be aware of as genetic testing for one disease may inform risk of another.

EVALUATING DISEASE-CAUSING VARIANTS FOR DIAGNOSTIC/PREDISPOSITION TESTING

The finding of linkage or association of a genetic variant with a disease does not always mean that the variant will be a clinically valid predictor of disease. In order to make the jump from research to clinical application, the evidence supporting that relationship must be evaluated in terms of the robustness, generalizability, and predictive value of the genetic variant.

How robust is the association?

Robustness of a genetic association in a given research study can be quantified in terms of statistical significance. Generally, the threshold for declaring significance of an association found for a complex disease through GWAS is $p<10^{-8}$, a cutoff that considers multiple testing across a million variants. In addition to a statistically significant finding in a single study, the association should be reproducible in a second, separate population.

For Mendelian disease, the cosegregation of the specific allele with disease in multiple independent families with the same condition provides evidence of robust linkage. The strength of the evidence ranges from supporting to strong, depending on the number of independent families and the strength of linkage of the disease and variant. In addition, genes implicated in disease are usually required to have supporting evidence from functional studies demonstrating an effect of the variant on downstream protein structure and/or function.

How generalizable is the association?

Most GWAS are carried out in European populations and the relationship between these variants and disease in other ethnic groups may not have been studied. Some associations are found for different subsets of disease as well, such as early-onset forms, and may not generalize to later-onset disease. Reviewing the primary studies where the associations were found will give the reader a sense of the generalizability.

How predictive is the genetic variant of disease?

The predictive value, or penetrance, of a genetic association describes the strength of the association. This measure of effect is generally calculated from research studies that may be biased or inappropriately designed for producing such estimates. Ideally, the predictive value of test will be determined from prospective studies representative of the population(s) that the test will be applied to. Short of that, one must recognize that risk estimates derived from retrospective research studies (especially case-control studies) may not be accurate.

Penetrance estimates for Mendelian disease gene variants also tend to be biased. Historically they have been calculated from within families with the disease in question. As more and more ostensibly healthy individuals have been sequenced, we are finding individuals carrying these high

penetrance variants who don't have the disease, suggesting that earlier penetrance estimates are generally overestimated[10]. Still, penetrance for classic Mendelian disease is still much higher than that for common, complex conditions.

RESOURCES

Online Mendelian Inheritance in Man (OMIM) (http://www.ncbi.nlm.nih.gov/omim)

OMIM is a comprehensive, authoritative compendium of human genes and genetic traits that is freely available and updated daily.

GeneReviews (www.genereviews.org)

An online database containing standardized peer-reviewed articles that describe specific heritable diseases.

Catalog of published genome-wide association studies (GWAS) (https://www.ebi.ac.uk/gwas/)

The GWAS catalog is a searchable, manually curated collection of all published GWAS with significant trait associations.

REFERENCES

1. Lorey FW, Arnopp J, Cunningham GC. Distribution of hemoglobinopathy variants by ethnicity in a multiethnic state. *Genet Epidemiol*. 1996;13:501–512. [PubMed: 8905396]

2. Strom CM, Crossley B, Buller-Buerkle A, et al. Cystic fibrosis testing 8 years on: lessons learned from carrier screening and sequencing analysis. *Genet Med*. 2011;13:166–172. [PubMed: 21068670]

3. Scott SA, Edelmann L, Liu L, Luo M, Desnick RJ, Kornreich R. Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases. *Hum Mutat*. 2010;31:1240–1250. [PubMed: 20672374]

4. Mitsui J, Mizuta I, Toyoda A, et al. Mutations for Gaucher disease confer high susceptibility to Parkinson disease. *Arch Neurol*. 2009;66:571–576. [PubMed: 19433656]

5. Noone PG, Zhou Z, Silverman LM, Jowell PS, Knowles MR, Cohn JA. Cystic fibrosis gene mutations and pancreatitis risk: relation to epithelial ion transport and trypsin inhibitor gene mutations. *Gastroenterology*. 2001;121:1310–1319. [PubMed: 11729110]

6. Tsaras G, Owusu-Ansah A, Boateng FO, Amoateng-Adjepong Y. Complications associated with sickle cell trait: a brief narrative review. *Am J Med*. 2009;122:507–512. [PubMed: 19393983]

7. Royer-Pokora B, Kunkel LM, Monaco AP, et al. Cloning the gene for the inherited disorder chronic granulomatous disease on the basis of its chromosomal location. *Cold Spring Harb Symp Quant Biol*. 1986;51(pt 1):177–183. [PubMed: 3472714]

8. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*. 2012;337:1190–1195. [PubMed: 22955828]

9. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet*. 2013;14:483–495. [PubMed: 23752797]

10. Van Driest SL, Wells QS, Stallings S, et al. Association of arrhythmia-related genetic variants with phenotypes documented in electronic medical records. *JAMA*. 2016;315:47–57. [PubMed: 26746457]