INTRODUCTION

Most common human diseases, such as coronary heart disease, diabetes, cancers, bipolar affective disorders and schizophrenia, have complex etiology. While they tend to cluster in families, they do not exhibit the characteristic Mendelian segregation ratios of single-gene disorders. These diseases therefore cannot be solely caused by a single genetic mutation, but have a more complex genetic architecture. For any complex disease, many questions about its genetic architecture can be
raised. How many genetic variants are involved in individual differences in the propensity to develop the disease (e.g., just a handful, tens, hundreds, or thousands)? Where are these sequence changes located on the 23 chromosomes that constitute the human genome? What is the nature of the sequence changes in these variants (e.g., single base pair changes, copy number changes, etc.)? What are the functional consequences of these changes (e.g., change of amino acid sequence and therefore protein structure, or changes in the level or regulation of gene expression)? What are the frequencies and effect sizes of these changes? How important are these changes relative to the environmental variation in explaining individual differences in disease susceptibility? And how do the genetic changes interact with each other and with environmental factors?

This chapter reviews the research approaches that have been used to address some of the above questions and summarizes our current state of knowledge and understanding regarding the genetic architecture of complex diseases that have come out of these studies. The general conclusion is that common diseases are highly heterogeneous, with a small proportion of cases having relatively simple etiology dominated by a single genetic mutation, while the vast majority of cases are caused by the combined effect of multiple genetic and environmental factors each contributing a minor influence.

The genetics approach to the study of complex diseases is complementary to other research paradigms such as the use of cell culture or animal models. The advantages of the genetics approach are that: (1) the finite size and regularity of the genome allows a systematic search for sequence-phenotype relationships, which may unveil novel associations that implicate previously unsuspected biological pathways, and (2) the demonstration of sequence-phenotype relationships offers strong direct evidence for the role of a gene or a pathway in human disease, minimizing the need to perform potentially hazardous experiments on humans. On the other hand, the genetics approach is limited in its ability to tease apart detailed molecular mechanisms involved in disease etiology. The genetics approach is therefore a valuable complement, rather than an alternative, to other biological approaches.

**GENETIC MODELING: TWIN, ADOPTION AND FAMILY STUDIES**

One immediate question regarding the genetic architecture of complex traits is the relative importance of genetic versus environmental factors in explaining the individual differences in disease susceptibility in the
population. If genetic factors are relatively unimportant, then further genetic studies may be unwarranted, and research efforts should be directed at environmental factors. On the other hand, if the contribution of genetic factors is substantial, then further genetic studies may help to identify the specific genetic variants involved and elucidate the mechanisms by which these variants influence disease propensity. The proportion of the total variance in disease liability that is explained by genetic (as against environmental) factors is defined as the heritability of the disease. It is important to appreciate that heritability is dependent on the genetic and environmental variations present in a population, so that changes in the variability of genetic or environmental factors can both lead to changes in heritability.

Heritability is typically estimated from twin and adoption studies (see [1] for an overview of methods for estimation of genetic and environmental components of variance). The principle of twin studies is as follows. Identical or monozygotic (MZ) twins share 100% of their genomes, while fraternal or dizygotic (DZ) twins share on average only 50% of their genomes by common descent. On the other hand, for twin pairs who are reared together, their sharing of environmental exposures may be the same regardless of zygosity. Thus, the presence of a greater phenotypic similarity among MZ than DZ twins can be attributed to the greater genetic similarity of MZ than DZ twins. Indeed, if the phenotype is a continuous trait, then the phenotypic similarity within twin pairs can be measured by an intraclass correlation, and the heritability estimated by twice the difference in intraclass correlations in MZ and DZ twins.

The use of twin studies for heritability estimation of disease phenotypes is more complicated because of the dichotomous nature of the phenotype. This is usually done via a liability-threshold model, where the underlying liability is normally distributed in the population, and those individuals with liability above a certain threshold value develop disease. The twin data is then used to estimate the twin correlations for the underlying liability (tetrachoric correlations), and then the heritability can be estimated by twice the difference in these correlations between MZ and DZ twins. As seen in Table 1.1, heritability estimates from twin studies on a number of complex diseases range from 40% to as high as 90%, which are typical of most complex traits.

Heritability can also be estimated by adoption studies, including MZ twins reared apart, whose correlation gives a direct estimate of heritability. In general, the correlations between biological relatives who have been separated by adoption provide estimates of heritability, whereas the correlations between adoptive relatives reared together but are biologically unrelated provide estimates for the influence of the family environment (see Plomin & Loehlin [12] for a discussion of direct estimates of heritability). Arguably the most prominent among such studies,
the Minnesota Study of Twins Reared Apart, confirms that practically all complex traits have a substantial genetic component (e.g., [13–15]).

The modeling of twin, adoption and family data can be used to address other important questions about the genetic architecture of complex disorders. For example, two different diseases can be modeled simultaneously, to detect shared genetic influences on the two diseases. In twin studies, shared genetic influences would be indicated by the presence of cross-trait cross-twin correlation (i.e., correlation between disease 1 in twin 1 and disease 2 in twin 2) for both MZ and DZ twins, but which is greater in MZ than in DZ twins. Such studies have indicated substantial genetic sharing for some complex diseases; for example, schizophrenic with manic symptoms [16], and bipolar disorder with unipolar depression [17]. Differences in the genetic influences on disease liability between males and females, for different ages, or under different environmental conditions, can also be modeled in twin data. For example, Kendler et al. [18] found that twin similarity for social phobia was due primarily to genetic influences in males but a result of shared environmental influences in females.

Another type of genetic modeling is aimed not at estimating the relative importance of genetic against environmental factors, but at whether the genetic component is made up of a single genetic factor of major effect (the single major locus [SML] model), or a large number of genetic factors each of small effect (the polygenic model). These are two extreme scenarios, and other possible models include the presence of a major locus on a polygenic background (the mixed model), or a few loci of major effect (the oligogenic model).

### TABLE 1.1  Heritability of Common Complex Traits and Diseases from Various Twin Studies. Adapted from MacGregor et al. [2]

<table>
<thead>
<tr>
<th>Trait</th>
<th>Heritability</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthma</td>
<td>60</td>
<td>[3]</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>40–70</td>
<td>[4]</td>
</tr>
<tr>
<td>Bone mineral density</td>
<td>60–80</td>
<td>[5]</td>
</tr>
<tr>
<td>Insulin dependent diabetes</td>
<td>70</td>
<td>[7]</td>
</tr>
<tr>
<td>Obesity</td>
<td>50–90</td>
<td>[8]</td>
</tr>
<tr>
<td>Osteoarthritis</td>
<td>50–70</td>
<td>[9]</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>60</td>
<td>[10]</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>50</td>
<td>[11]</td>
</tr>
</tbody>
</table>
One approach to discrimination between different genetic models is to consider the drop-off in recurrence risk of disease as a function of genetic relationship to an affected index case (proband). A polygenic model is predicted to have a steeper drop-off than an SML model; the empirical recurrence risks for schizophrenia appear to be more consistent with a polygenic than an SML model [19]. An alternative, more sophisticated method for discriminating between different genetic models is complex segregation analysis, which uses maximum likelihood on family data to fit model parameters and test different models. Thus, the presence of an SML can be tested by a likelihood ratio test of a mixed model with both SML and polygenic components, against a polygenic model. An alternative test for the presence of an SML considers a generalized transmission model in which genetic transmissions are allowed to deviate from Mendelian proportions, against a model in which genetic transmissions are constrained to Mendelian proportions. Complex segregation analysis has been applied to complex diseases with largely inconclusive results [20, 21]. This is because complex segregation analysis suffers from both low statistical power which throws doubt on negative results, and from numerous possible artifacts which throw doubt on positive results. Thus, a complex segregation analysis showed strong but likely erroneous evidence for an SML effect for medical school attendance. The problem is that the model did not incorporate sibling environment and therefore could account for a higher sibling concordance than parent–offspring concordance for medical school attendance only through a recessive SML [22].

Notable oligogenic models for complex diseases were proposed by Risch [23]. In these models, the effects of the different loci on disease risk can combine in an additive or multiplicative fashion. Risch derived, under each model, how the overall disease prevalence and recurrence risks can be related to the effects of the individual loci. Risch considered that the pattern of recurrence risks in schizophrenia is consistent with an oligogenic model with three to five loci. This would mean that these loci must have quite large effects, providing optimism for studies which aim to identify individual susceptibility loci.

**DISEASE GENE MAPPING: LINKAGE STUDIES**

Linkage is based on the co-segregation of marker variants and disease in families. In humans, the frequency of crossovers in meiosis is such that each gametic genome has on average only around 35 crossover points. Co-segregation should therefore be detectable for marker loci quite far away from the disease-causing variant. Because linkage operates over long
genetic distances, a positional mapping approach based on linkage can cover the entire genome by using a relatively small number of highly polymorphic markers. Standard marker sets for whole-genome linkage scans, based on 200–800 microsatellite polymorphisms, which became available in the 1990s, enabled the successful mapping of hundreds of rare single-gene disorders.

Classical linkage analysis is typically carried out using the lod-score method, which is based on a single major locus parameterized by the disease allele frequency and the penetrances (conditional probability of disease) of the three disease locus genotypes. For Mendelian diseases, the values of these parameters can be easily specified from the results of population prevalence studies and segregation analyses. For complex disease, the SML model is likely to be simplistic, and appropriate values of the model parameters unknown, and indeed may be different for different loci. Nevertheless, classical linkage analysis has been optimistically applied to complex diseases, particularly on pedigrees with an unusually large number of affected individuals. There have been some successes of this approach; for example, the identification of loci responsible for early-onset familial breast cancer [24], maturity-onset diabetes in the young [25] and early-onset Alzheimer's disease [26–28]. However, the patients in these successful linkage studies represent only a very small proportion (<5%) of the overall incidence of each of the complex disease. When successful, the families involved are usually large and have a pattern of inheritance that is very close to autosomal dominant with high penetrance. For collections of smaller families with less clear-cut Mendelian inheritance, the results of classical linkage analysis are much more often unconvincing and difficult to replicate. Examples of non-replicated linkage findings include schizophrenia [29] and bipolar affective disorder [30, 31].

The lack of success in classical linkage analysis for the majority of cases of complex disease suggests that the genetic variants involved in such disorders typically have a small effect on disease risk. Thus multiple genetic variants are often involved in a single family, with the result that the co-segregation between disease and any single variant would be imperfect and different families will show linkage at different loci. It follows that for complex disorders, very large family samples are required to demonstrate conclusive linkage between disease and genetic markers.

A different version of linkage analysis, called non-parametric linkage, is based on excess local allele sharing between the affected relatives, above the level expected for the degree of relationship. For example, sibling pairs are expected to share on average one of the two alleles at any locus, and a locus which shows a significant excess of allele sharing above this level for affected sibling pairs would constitute evidence of linkage with the disease. This method has been considered to be more appropriate
for complex diseases as it does not assume an SML. The non-parametric linkage approach, usually based on affected sib-pairs, became popular in the 1990s. The approach was used successfully, for example, on late-onset Alzheimer’s disease, to identify a linkage region on chromosome 19 [32], which was subsequently found to contain a major susceptibility variant, the APOE ε4 allele [33]. Other studies using this approach (for example on multiple sclerosis [34, 35], schizophrenia [36–39] and autism [40]) have been less successful. The problem is that, while non-parametric linkage does not require the assumption of a single major locus model, the statistical power to detect linkage is nevertheless highly sensitive to the effect size of the susceptibility variant. For realistic sample sizes, only genetic effects which account for a substantial fraction (e.g., 20%) of the disease heritability are likely to be detected. The lack of success of non-parametric linkage analysis for many complex diseases would exclude the presence of genes of major effect. However, even variants which account for as much as 10% of the disease heritability are likely to go undetected because of inadequate statistical power.

DISEASE GENE MAPPING: ASSOCIATION STUDIES

Association analysis has a potentially far greater power than linkage analysis for detecting variants with modest effect on disease risk, provided that the genetic marker is close enough to exhibit strong linkage disequilibrium (LD) with the functional variant. This fact was recognized by Risch and Merikangas [41], who suggested that the future of complex disease genetics lies with systematic association rather than linkage studies. However, tight LD between polymorphisms requires them to be very close to each other, usually within 50 kb or less. The use of association analysis to identify disease susceptibility variants therefore depends either on prior biological knowledge that points to particular polymorphisms in candidate genes, or requires a very high density of genetic markers in candidate regions suggested by linkage studies or cytogenetic abnormalities, or indeed throughout the entire genome.

Association studies of candidate genes have revealed many susceptibility genes. A comprehensive review of candidate gene association studies, mostly performed when such studies were at their peak of popularity, found that over 600 “significant” associations for complex disease had been reported [42]. However, reviewing the subset of 166 loci which had been tested in three or more studies found that only six were consistently replicated. This suggests that our ability to pick candidate genes may be limited by our current lack of knowledge concerning the biology of complex diseases.
A systematic, genome-wide approach to association studies using dense marker sets is therefore much needed. Single-nucleotide polymorphisms (SNPs) are the most abundant type of sequence variants in the genome, occurring approximately once in every 100 to 300 base-pairs. The systematic cataloging of SNPs began in 1998 when the NIH Human Genome Project set the creation of an SNP map of at least 100,000 markers as one of its objectives. This resulted in the creation of the SNP Consortium, formed in 1999, with the goal of identifying 300,000 SNPs, but ultimately finding 1,400,000 SNPs by the end of 2001 [43]. The International HapMap Project [44], which aimed to create a map of 1 million common SNPs (defined as those where both alleles have frequency at least 5%), with not only their genomic locations but also genotype frequencies and LD relationships among each other, in three populations (Europeans, Africans and East Asians). The project involved the genotyping of enough SNPs to ensure at least one common SNP in every 5 KB bin of the genome, in 270 individuals (90 from each of the three populations). Subsequently, in Phase 2, the HapMap had been extended to include over 3 million SNPs on the same samples [45] and those samples plus additional ones were later genotyped using the latest SNP chip technology (discussed in the following paragraph) from both Affymetrix and Illumina, in Phase 3 of the project. The HapMap provides not only a very high density of common SNPs which can be used as markers in association studies, but also the genotype frequency data to determine the extent to which a set of selected SNPs can serve as surrogates (or tags) for all the other common SNPs in any targeted region of the genome.

The latest technologies for high-throughput SNP genotyping, the Affymetrix SNP Array 6.0 and the Illumina 1M BeadChip, can assay 1 million SNPs in a single reaction. These commercial genotyping products have been evaluated against HapMap data to provide adequate “coverage” for nearly 90% of all common variants in the genomes of European and East Asian populations [46], meaning that nearly 90% of all common variants are either included as one of the genotyped SNPs, or are in such strong LD with one or more of the genotyped SNPs that its genotype can be predicted (and therefore imputed) from those of the genotyped SNPs. Current genotyping technology therefore enables the systematic examination of nearly all common variants in the genome by association analysis, an approach called the genome-wide association study (GWAS). The first successful GWAS, published in 2005, detected a common risk allele for age-related macular degeneration in the gene coding for Complement Factor H [47]. Since then, the GWAS approach has identified multiple risk variants for many complex diseases. For example, over 10 risk variants for breast cancer have been identified, and over 20 for prostate cancer [48]. A consistent finding from all GWAS conducted to date is that nearly all the detected variants have very modest
effects on risk of disease (typically with odds ratios, OR, in the 1.2–1.5 range) and explain a very small proportion of the population variance in liability to disease (typically 0.2–0.5%). Even in aggregate, all the risk SNPs identified to date for any complex disease explain a very modest proportion of variance in liability, typically 5–10%. Since the overall heritability of these complex diseases are typically in the range of 40–80%, the GWAS approach has only just begun to characterize the genetic components of these disorders.

These emergent findings from GWAS have important implications for the genetic architecture of complex diseases. First, the number of sequence variants that influence disease risk must be very large, at least in the hundreds if not thousands. Second, since statistical power for detecting an associated variant is determined by both effect size and allele frequency, many risk variants may have escaped detection because of either a very small effect size (OR <1.2) or low allele frequency (<5%). Third, the genome-wide coverage of 90% calculated for current genotyping products applies only to common SNP variants, and it is known that the coverage is much lower for rare SNPs. Lastly, there may be other types of variants for which coverage is lower (e.g., copy number variants, CNV).

There are compelling theoretical reasons why rare variants may be important for complex diseases. Compared to common functional variants, which are likely to have arisen from ancient mutations and have adapted to the rest of the genome (those failing to adapt would have been eliminated by negative selection), rare functional variants have a more recent origin, are more numerous, and are less well-adapted to the rest of the genome. For these reasons, rare functional variants may tend to have a larger effect on disease risk, and collectively explain a substantial proportion of population variance in disease liability [49]. Much evidence has been mounting for the role of rare variants in complex disease. In 2009, a paper in *Science* [50] reported four rare (<3%) variants in the IFIH1 gene which decrease risk for developing type 1 diabetes (T1D). This gene plays a role in the immune response to enterovirus infection and the finding of protective rare variants suggests that T1D may be a result of an overaggressive response to these particular foreign invaders. All four variants discovered are predicted to result in severe functional disruption of the gene. One introduces a premature stop signal, two are found in conserved RNA splicing sites, and the fourth alters an evolutionarily conserved site. One of these variants has a much larger effect than those typically found from GWAS, halving the risk of developing T1D.

For some complex diseases, the combined sample size from all the GWAS is now in the thousands, and yet the associations detected still explain only a small proportion (typically 20% or less) of the total heritability [51]. A number of possible explanations have been proposed for
this problem of “missing heritability.” First, it has been suggested that heritability estimates for complex diseases may be inflated due to methodological problems. Another possible explanation is that the SNP sets used in current GWAS offer poor tagging, especially for rare variants and structural variations. This would both reduce the number of associations detected, and underestimate the true effect sizes of the detected loci. It is also possible that many susceptibility loci simply have very small effect sizes, so that many have not been detected due to the inadequate statistical power of current studies. Finally, it has been suggested that gene–gene and gene–environment interactions account for a substantial portion of the heritability estimates, but these interactions have been largely neglected in GWAS to date.

CONCLUSION

The rapid development of molecular genetic technologies has allowed highly detailed examination of genome sequence variation, and led to rapid progress in our understanding of the genetic architecture of complex diseases. The field is still rapidly moving, with increasingly higher-powered GWAS being conducted to detect loci with diminishing allele frequency and effect sizes. At the same time, next-generation whole-genome sequencing is becoming less expensive, and it will soon become feasible to examine both known and (previously) unknown sequence variations for association with disease liability. Our picture of genetic architecture of complex diseases will therefore likely change quite rapidly and become much more detailed in the next few years.

Nevertheless, it is useful to consider our current picture of the genetic architecture of complex diseases, taking into account the most recent findings from large-scale association studies. It now appears that most complex diseases are under the influence of a very large number, probably hundreds, of sequence variants. Both common and rare variants are involved, with a wide range of effect sizes. High-penetrance mutations are responsible for some particularly familial and early-onset forms of complex diseases, but these usually account for a very small proportion of cases. Examples include BRCA mutations for early-onset breast cancer, and APP for early-onset familial Alzheimer’s disease. A few common genetic polymorphisms have been identified which have moderately strong effects (allelic odds ratio >2); for example, APOE ε4 for late-onset Alzheimer’s disease, CFH variant in age-related macular degeneration. It may be that the late age-of-onset of these disorders have reduced selective pressure against these deleterious diseases, and allowed high-risk variants to become quite common in populations. The majority of the genetic
factors for complex diseases appear to have very small effect size, with an allelic odds ratio of less than 1.5. At present, the importance of gene—gene and gene—environment interactions is unclear. Although interactions are likely to be widespread, it appears from family data that they probably account for only a modest proportion of the variance in disease liability. However, this remains to be confirmed by studies with larger sample sizes and more comprehensive coverage of sequence variants in the genome. As future studies identify more and more sequence variants that account for individual differences in disease liability, the genes and pathways that determine the development of complex diseases will become clearer. This will enable further studies to focus on the function of these genes and pathways, to elucidate the mechanisms that lead to disease.

References


