Chapter 4
Complex Disease Genes and Their Discovery

Jeffrey C. Barrett and Mark J. Daly

Abstract The study of the genetic underpinning of heritable human diseases stretches back nearly a century. While thousands of mutations in single genes have been found that cause severe "Mendelian" disorders, attempts to find such single genes for complex diseases have been relatively unsuccessful. Instead it has become clear that complex diseases, like IBD, are affected by many (likely hundreds or even thousands) different genes as well as environmental factors. Here we describe the process by which that discovery was made, as well as the technological advances from small-scale candidate gene to genome-wide association studies. These approaches, especially when undertaken in large-scale collaborations, have unlocked thousands of complex disease genes, including 163 associated with IBD. Despite these exciting developments, the discovery of genes represents the first stage in translating that knowledge into biological understanding of disease and possible future treatments.

Background

It has long been appreciated that genetics plays an important role in susceptibility to a wide variety of complex human diseases. Indeed it has been almost 100 years since R. A. Fisher and others reconciled the discrete Mendelian inheritance of

J.C. Barrett, B.S., D.Phil. (✉)
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton,
Cambridge, CB10 1HH, UK
e-mail: barrett@sanger.ac.uk

M.J. Daly, Ph.D.
Simches Research Center, Massachusetts General Hospital,
185 Cambridge Street, CPZN 6818, Boston, MA 02114, USA

Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts
General Hospital, 185 Cambridge Street, CPZN 6818, Boston, MA 02114, USA
e-mail: mjdaly@atgu.mgh.harvard.edu

M. D'Amato and J.D. Rioux (eds.), *Molecular Genetics of Inflammatory Bowel Disease*,
DOI 10.1007/978-1-4614-8256-7_4, © Springer Science+Business Media New York 2013

individual genes with the continuous distribution of complex heritable traits, such as height [1]. The earliest geneticists at that time were realizing that while most traits were correlated among relatives and thereby appeared "heritable," only a few of them strictly followed Mendel's laws of inheritance. Instead, the majority of heritable traits and diseases quite evidently involved the action of many genes (as well as nongenetic or environmental factors). The suggestion that individual genetic variants might have relatively modest effects on these traits, and that if they were sufficiently numerous, would give rise to normal distributions in general populations is one which continues to reverberate through the most cutting-edge genetic studies of today.

Indeed, the history of complex disease genetics has been a story of reconciling the obvious family clustering of these diseases with an evolving understanding of the types of genetic variation that exist in human populations and how they affect disease risk. In this chapter we will describe how that process moved (over many decades) from relatively fruitless searches to find single genes explaining disease in individual families to international collaborations studying tens of thousands of patients with complex diseases at once. At each stage, a combination of dedicated clinical researchers, statistical analysts, and new technologies enabled new discoveries to be made. We will try to illustrate this process with examples from IBD and conclude by considering what biological lessons have been learned and what challenges remain.

Chasing a Successful Paradigm: Linkage Studies in Complex Disease

Even in the simplest, so-called "Mendelian" diseases, the earliest studies of inherited phenotypes often showed that two genetic traits cosegregated—that is, were correlated in their transmission from parents to children. This is of course the consequence of the genes being "linked" or closely located on the same chromosome, such that from generation to generation, the pair of genes is passed on intact from the same parental chromosome without intervening meiotic recombination. Sturtevant and Morgan 100 years ago conceived that this cosegregation of phenotypes could be used to create a linear map of the underlying order of the genes responsible for those phenotypes—thus creating the first linkage map [2] (in this instance in fruit flies). Mathematically, the principles of linkage analysis were worked out independent of even the definition of the structure of DNA [3], but it was not until the 1970s and 1980s that the molecular techniques to clone and sequence DNA permitted the experimental connection of genetic linkage maps of phenotypes to underlying DNA variation and thus to identify genes responsible for phenotype via linkage analysis, followed by detailed sequencing and functional studies to define the specific underlying causal gene and mutation—so-called positional cloning.

Use of this strategy to approach human disease was outlined at this time [4, 5] and first successfully demonstrated in the localization of the Huntington's disease

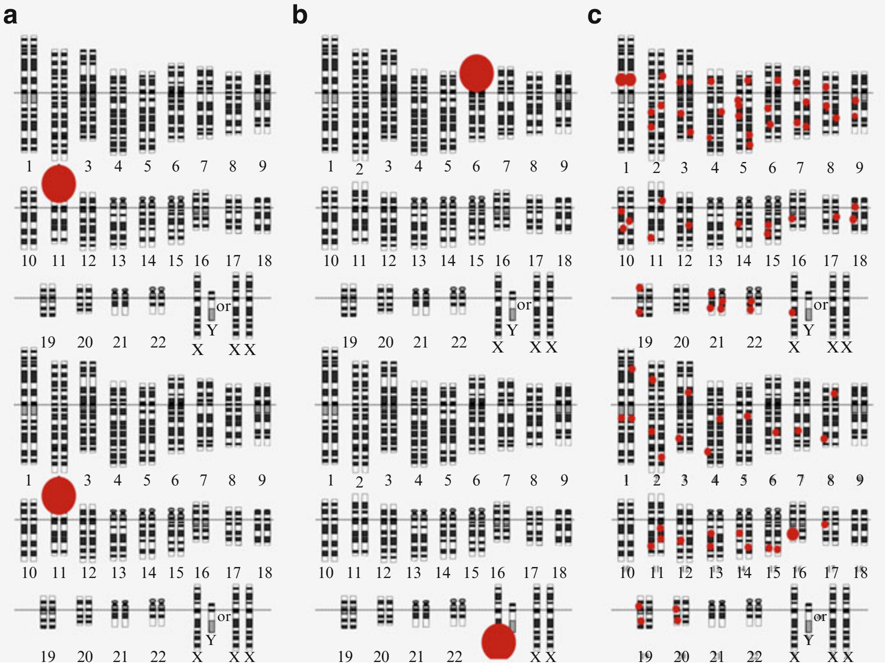

Fig. 4.1 Three illustrations of the spectrum of disease genes. Six images of imaginary "patient genomes" with *red circles* corresponding to risk alleles for a particular disease. Three columns (**a**–**c**) correspond to different disorders, with two patients of each for comparison. (**a**) Some single gene, or Mendelian, disorders, such as sickle cell disease, are caused by mutations in the same gene in all patients. (**b**) Other disorders, such as intellectual disability, are often caused by a single mutation in each patient, but can be in a variety of genes. (**c**) Complex diseases, like IBD, are instead affected by a large number of individually weak variants across the genome, none individually as strong as the Mendelian variants

gene in 1983 [6]—later coming to fruition in the identification of genes for cystic fibrosis and Duchenne muscular dystrophy as the 1980s drew to a close. Family-based linkage studies in humans are the most direct approach to analyzing the simple consequences of Mendelian inheritance from one generation to the next and the resulting sharing of relatively long segments of DNA *identical by descent*. Consider families with multiple individuals affected with a rare disease (see Fig. 4.1) such as Huntington's disease (which affects a handful of individuals per 100,000 populations [7]): it is very unlikely that such co-occurrence would happen by chance, so a genetic explanation is likely. If the family is large, with a sufficient number of affected individuals, a classic pattern of Mendelian inheritance might be clear, such as *recessive* (requiring two copies of a damaging mutation to be affected, one from each parent) or *dominant* (individuals with one copy of the mutation, from either parent, are affected—the pattern seen in Huntington's). Because the segments of shared chromosomes between nearby relatives are large (tens of megabases among first-degree relatives), it is possible to identify which parts of which chromosomes are shared between affected individuals with relatively few DNA markers. If one chromosomal segment is shared in a consistent way among all affected family members, but not those who are healthy, it likely carries a mutation that causes disease. The same region (containing the relevant gene) will be "linked" in this way in many different families, even if the individual mutations responsible vary across families.

For this reason, the earliest studies of complex disease genetics built upon the successes of family-based linkage studies in mapping Mendelian disease genes. The linkage approach was rapidly applied across a range of rare diseases suspected of Mendelian inheritance, leading to the identification of hundreds of disease genes throughout the 1990s. These successes led to the application of the same approach to more common, complex diseases, such as type 2 diabetes and inflammatory bowel disease. Much like Mendelian disease, these diseases were known to run in families, and studies of disease concordance in monozygous and dizygous twins rigorously established that genetics plays an important role in their etiology. With a few exceptions (see next chapters), however, linkage studies in these complex diseases did not lead to the discovery of major genetic risk factors.

The principal insight from this failure was the lesson learned from the earliest genetic studies in fruit flies, specifically that, in contrast to Mendelian disease, there was not a single gene (nor even a very small number) for most complex diseases. This was partially unsurprising, as multiply affected families with these diseases did not display the classic Mendelian patterns of inheritance, but instead appeared to be driven by combinations of many genetic factors each exerting a relatively weak effect—just as predicted by the biometrical models of Fisher. These realizations lead to the invention of more sophisticated statistical techniques for linkage analysis [8] aimed at discovering loci which only partially explain the disease state of family members.

A key observation was made [9] that the power of linkage studies falls rapidly with decreasing effect size of the associated genetic variant. If the genetic basis of complex disease was completely unlike Mendelian disease, and instead consisted of dozens or hundreds of small effects, then linkage would never, for practical purposes, be able to discover them. *Association studies*, where one simply compared the allele frequency of a particular variant between unrelated cases and controls, had the potential to discover these tiny effects. At the time of this publication, however, technology did not exist to make such studies possible, nor was there even a compelling estimate of how much genetic variation existed in the human population. By contrast to genome-wide linkage mapping, where the recombination map of humans had been described for a decade and could be conveniently assayed by fewer than 1,000 polymorphic markers, genome-wide association would need to wait.

Motivated by Biology: Candidate Gene Studies

The complex disease genetics community was thus faced with the twin realities that linkage mapping would be unlikely to discover risk loci for these diseases and that genome-wide application of the association study paradigm was still technologically impossible. One possible solution to this problem would be to prioritize genes for genetic study that seemed biologically plausible candidates for particular diseases. For instance, one of the few success stories in complex disease linkage mapping was the identification of a tandem repeat polymorphism in *INS* (the gene encoding the insulin protein) associated with type 1 diabetes [10]. This discovery fit neatly into the developing biological understanding of the disease and suggested that perhaps genetic discoveries could be made by first guessing the relevant candidates.

Unfortunately, three problems undermined this candidate gene approach. First, and perhaps most importantly, the ability of researchers to predict which genes would be associated with which diseases was poor: obvious connections like insulin and type 1 diabetes were not common. Second, the available patient collections typically numbered in dozens or low hundreds: too small to detect the very weak effects that would come to typify the contribution of common alleles to complex diseases. Finally even with the good fortune of picking the right gene, it was not possible to select SNPs that represented the diversity of variation within that gene in a systematic fashion. Just as it had been impossible to query the reference sequence of a gene before the draft human genome was finished, it was now impossible to look up how a particular gene commonly varied within a population of interest.

Maps of Common Human Genetic Variation

As noted above, a common feature of candidate gene studies was the selection of only a handful of SNPs in each gene being considered to test for association. This limitation was largely a result of a lack of comprehensive databases of genetic variation throughout the human genome. The first project aimed at producing such a database was the SNP Consortium [11], which undertook large-scale genome resequencing and identified over one million SNPs. This effort provided the substrate for a wide array of subsequent investigations into the number, distribution, and frequency of SNPs throughout the genome.

One research area transformed by this new abundance of variation data was the study of population genetics: the quantity and frequency of, and patterns of correlation among, genetic variation in different populations around the world provided empirical data with which to fit models of human demography and selection. Two forces increase variation in the genome: mutation, which introduces new variants, and recombination, which reshuffles the existing patterns. Random drift, evolutionary selection (either positive or negative), and human population history then shape this pool of variation into the patterns seen in modern humans. Previous population genetics work could be used to make very specific predictions about the extent of correlation between nearby SNPs (known as *linkage disequilibrium*, or LD) given certain assumptions about the history of humans and, crucially, that recombination occurred uniformly throughout the genome. Two simultaneous observations suggested, however, that recombination was instead clustered in punctate "hotspots"—the vast majority of historical human recombinations had happened in a relatively small fraction of the genome sequence. Molecular typing of multiple sperm from a single individual showed clustering of directly observable recombinations [12], and an analysis of the precise positions where LD decayed in a survey of general population variation suggested this process was consistently concentrated over many generations [13].

Prompted by these insights, the International HapMap [14] project was launched to create a genome variation reference for medical genetics in multiple human populations across the entire genome. The project was undertaken in two main phases which yielded a map of the frequencies and LD patterns of over 2.5 million SNPs in individuals of European, West African, and East Asian ancestry. The HapMap provided both a generic variation reference and revealed new specific insights into human population history, such as strong support for the out-of-Africa hypothesis of human migration and a realization that a huge fraction of variation is shared across the world. In addition, this large-scale collaboration contributed heavily to the development of high-throughput genotyping technologies. In the course of the project, it became possible to move from genotyping dozens of SNPs to thousands and then hundreds of thousands—technological advances which would prove to be just as transformative as the scientific discoveries of the project.

Nowhere were the implications of these data and technologies greater than for the study of the role of genetic variation in disease risk. It became clear that it was possible to select a small number of SNPs from a particular region of the genome that were highly correlated with all nearby SNPs. These "tag" SNPs could then be genotyped as an efficient means of capturing all the information contained in the full complement of SNPs in the region [15]. This tagging approach was quickly shown to be scalable genome wide, so that fewer than 500,000 carefully chosen SNPs could capture nearly all the common variation in populations of European descent [16]. The stage was set for a revolution in the discovery of genetic risk loci for common diseases.

Genome-Wide Association Studies

Several developments from the HapMap project presented new opportunities for disease gene mapping: an understanding of genome-wide LD patterns, algorithms and tools for selecting efficient tag SNP sets, and affordable technologies for genotyping hundreds of thousands of SNPs. Taken together, these offered the ability to genotype large groups of healthy individuals and cases of particular diseases in a way which captured nearly all the common variation in individuals of European ancestry on an affordable scale. It was also recognized that robust genome-wide statistics would likely only emerge from much larger sample collections than customarily used in candidate gene studies, and indeed even at this time both the lack of consistent marker maps and small samples with inconclusive statistical support were creating a cacophony of inconsistent candidate gene studies for many gene, disease pairings. These early *genome-wide association studie*s (GWAS) showed some early successes [17]; they also confirmed the increasing suspicion that individual common risk alleles generally exercised very weak effects on disease risk: few odds ratios were >1.2. Crohn's disease (as will be described in future chapters) benefited from a number of early GWAS discoveries, increasing the number of confirmed loci to a dozen [18–20].

It quickly became clear that data quality control of GWAS data was essential to producing interpretable and reproducible results [21]. While the genotyping platforms produced data that were extremely high quality on average, the sheer size of the datasets compared to earlier studies meant that even very low error rates could produce spurious associations. A suite of quality control metrics, including missing data rates, Hardy–Weinberg equilibrium, and overall heterozygosity quickly became standards in GWAS analysis, and geneticists became familiar with QQ-plots and other statistical tools as a rapid transition from genetic studies where each genotyping assay was manually inspected and scored to automated genome-wide typing technologies took place. It was also recognized that, even if genotyping data were perfect, false inference of association could arise if the ancestries of cases and controls were not well matched and the frequency differences characteristic of different populations were confounded with case–control status. Here a parallel set of methods emerged [22, 23] to measure and control for population structure in association studies that, like the QC standards, are still in wide use. Furthermore, the genetics community insisted on stringent statistical significance thresholds ($p<5\times10^{-8}$ being a common criterion for genome-wide significance [24]) and replication of any putative findings in independent samples [25] to generate ultimate assurance that novel genetic findings constituted truly durable insights into disease pathogenesis.

These rigorous guidelines for GWAS produced a substantial shift away from the contentious and generally irreproducible findings from linkage and candidate gene studies and rapidly produced a swath of bona fide associations to a wide variety of common diseases. Despite these early GWAS successes, however, it became apparent that associations from the first generation of studies explained only a very small fraction of the total genetic contribution to disease [26]. A variety of hypotheses were proposed to explain this so-called missing heritability, including a preponderance of rare variants, copy number variation, and complex interactions among risk loci [27]. None of these explanations lent themselves to straightforward post-GWAS experimentation nor had direct evidence that they explained the majority of what was not yet found. What was clear was that the confirmed findings from GWAS were both numerous and generally barely strong enough to have been detected, suggesting that many more results might lie just beneath the surface. Thus, the natural next step was for geneticists to set aside historical competitions in favor of combining GWAS datasets studying the same disease to investigate what additional associations might be discovered via collaboration.

Meta-analysis and the Importance of Sample Size

Early examples of GWAS *meta-analysis*, where individual scans were combined, often using summary association statistics from individual projects, began shortly after the initial GWAS publications [28, 29]. These studies, which typically consisted of a few thousand individuals, rapidly confirmed the suspicion that a large number of additional common alleles of small effect were waiting to be identified. Reassuringly, and in contrast to the experience of both complex disease linkage and candidate gene studies, GWAS meta-analyses also confirmed nearly all the previously published associations in the smaller scans.

The possibilities of this approach were most clearly realized by researchers studying quantitative traits, such as height or cholesterol levels, which had been measured in hundreds of thousands of individuals subjected to GWAS analysis. Unlike specific disease studies, which were limited by the incidence of diseases and the difficulty of recruiting large numbers of cases, these quantitative trait studies could draw samples collected for any number of different study designs, so long as the measurement of interest had been recorded in a consistent way. In the most recent meta-analysis of height GWAS [30], for instance, nearly 200 independent genomic loci showed significant association. A similar trend was observed across a wide variety of traits and diseases: as sample sizes increased, so did the number of associated loci. What often differed, however, was the number of samples required to make the earliest discoveries (i.e., find the biggest effects in that disease) and the rate at which loci subsequently accumulated. It is still unclear whether fundamental differences in genetic architecture, heterogeneity of diagnoses, or other factors might explain this locus discovery "coefficient." IBD once again reaped the benefits of these approaches via a series of successively larger meta-analyses culminating in the discovery of 163 independent loci [28, 31–33].

In addition to unleashing a torrent of individual associations across hundreds of diseases and traits, large meta-analysis sample sizes encouraged the application of newly developed statistical methods that analyzed the entire genome at once. Rather than focus on the most statistically significant associations, these methods [34] aimed to evaluate the total amount of phenotypic variance explained by common variation across the entire genome. These models suggested that a much larger fraction of total variance in many traits and risk of diseases could be explained by common variation than was explained by the genome-wide significant loci. It may be, therefore, that seeking to fully uncover the "missing heritability" is a fruitless effort since these hundreds or thousands of tiny effects will be impossible to pinpoint individually.

Biological Insights from Disease Gene Mapping

In parallel to the goal of trying to identify the specific regions of the genome associated with disease risk, or particular DNA variants which cause those associations or the overall contribution of common variation in general, disease gene mapping also provides an opportunity to better understand the biological mechanisms of health and disease. Indeed, given the difficulty in accurately predicting disease susceptibility from GWAS-type analyses [35], it is likely that new biological insights will be the most important long-term benefits of studying the genetics of complex diseases.

GWAS studies of fetal hemoglobin (HbF) levels in sickle cell disease (SCD) patients offer an informative example of this process. In some ways, SCD is the fundamental example of a Mendelian genetic disease, as the recessive mutation (in the hemoglobin beta gene) which causes the disease has been known for over 60 years [36]. It has also been long known that increased levels of HbF (encoded by a different gene and typically not expressed after birth) substantially reduce the severity of SCD. This observation led to a GWAS for HbF level [37], which identified a strong effect of variants near the *BCL11A* gene on persistence of HbF after birth. This discovery was followed by the remarkable discovery that reducing the activity of *BCL11A* could substantially alleviate SCD in mice [38]. A single new GWAS discovery opened a potential therapeutic avenue that had remained undetected despite decades of biologically motivated research into the relationship between HbF and SCD.

The discovery of *BCL11A* as a key regulator of HbF also serves as an illustration of the caution needed when predicting how quickly GWAS results will enable new diagnostics or treatments. They serve as a critical starting point, a biological truth that some functional unit in a particular part of the genome is related to disease risk. That piece of information alone offers little clue how to then move towards more complete knowledge of disease processes but certainly offers better prospects than aimlessly trying to make sense of those same processes in the context of the entirety of human biology.

Future Directions

The most substantial change to mapping complex disease genes at the present is the transition from GWAS-style data (where only a subset of common variation is studied) to complete genome sequencing, enabled by the plunging cost of sequencing compared to genotyping [39]. These technological shifts have the potential to open up the study of a wide spectrum of variation beyond the common alleles targeted by GWAS. It will be important, however, not to forget the lessons of that era principally that large sample sizes are critical to success. In addition to broadening the types of genetic variation which can be detected, sequencing-based studies (either directly or indirectly through imputation in projects like 1,000 Genomes) have the potential to more rapidly proceed from region of association to specific causal variants. While these discoveries only slightly affect the amount of variation explained in a disease or trait of interest, they have great potential to aid the biological inferences described above.

In a very real sense the progression of gene mapping described in this chapter is approaching its final stages: it will soon be possible to analyze the complete genome sequence of nearly all the patients seen with a particular disease. It is likely that even at that point, it will be impossible to perfectly predict an individual's risk of disease. Instead, we must ask whether we can use this limited information in a clinically useful way (in a similar sense to currently used risk measures, like cholesterol levels, which are strongly but imperfectly predictive of outcomes like heart disease) and simultaneously promote the utility of genetic association results in more fundamental biological studies of human health.

References

1. Fisher RA (1919) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52:399–433
2. Sturtevant AH (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J Exp Zool 14:43–59
3. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7:277
4. Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci 75:5631–5635
5. Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet 32:314
6. Gusella JF et al (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306:234–238
7. Wise J (2010) Prevalence of Huntington's disease is underestimated in UK. BMJ 340:c3516
8. Markianos K, Daly MJ, Kruglyak L (2001) Efficient multipoint linkage analysis through reduction of inheritance space. Am J Hum Genet 68(4):963–977
9. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517
10. Undlien DE et al (1995) Insulin gene region-encoded susceptibility to IDDM maps upstream of the insulin gene. Diabetes 44:620–625
11. Sachidanandam R et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933
12. Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nat Genet 29:217–222
13. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232
14. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
15. Pe'er I et al (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38:663–667
16. Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662
17. Frayling TM et al (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894
18. Duerr RH et al (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314:1461–1463
19. Rioux JD et al (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39:596–604
20. Parkes M et al (2007) Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. Nat Genet 39:830–832
21. Anderson CA et al (2010) Data quality control in genetic case–control association studies. Nat Protoc 5:1564–1573

22. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004
23. Price AL et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909
24. Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32:381–385
25. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
26. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88(3):294–305. doi:10.1016/j.ajhg.2011.02.002
27. Eichler EE et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450
28. Barrett JC et al (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 40:955–962
29. Voight BF et al (2010) Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet 42:579–589
30. Paper A. F. L. O. A. A. T. A. A. A. T. E. O. T (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467:832–838
31. Franke A et al (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. Nat Genet 42(12):1118–1125. doi:10.1038/ng.717
32. Anderson CA et al (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet 43:246–252
33. Jostins L, Ripke S, Barrett JC, Cho JH (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 490:119–124
34. International Schizophrenia Consortium et al (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752
35. Jostins L, Barrett JC (2011) Genetic risk prediction in complex disease. Hum Mol Genet 20:R182–R188
36. Ingram VM (1957) Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. Nature 180:326–328
37. Menzel S et al (2007) A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. Nat Genet 39:1197–1199
38. Xu J et al (2011) Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing. Science 334:993–996
39. Kilpinen H, Barrett JC (2013) How next-generation sequencing is transforming complex disease genetics. Trends Genet 29(1):23–30. doi:10.1016/j.tig.2012.10.001