Published in final edited form as:

Psychiatr Clin North Am. 2010 March ; 33(1): 1–13. doi:10.1016/j.psc.2009.12.006.

Genetics of psychiatric disorders: Methods: Molecular approaches

Dimitrios Avramopoulos, M.D., Ph.D.

McKusick Nathans Institute of Genetic Medicine and Department of Psychiatry, Johns Hopkins University, School of Medicine

Summary

The launch of the Human Genome Project in 1990 triggered unprecedented technological advances in DNA analysis technologies, followed by tremendous advances in our understanding of the human genome since the completion of the first draft in 2001. During the same time the interest shifted from the genetic causes of the Mendelian disorders, most of which were uncovered through linkage analyses and positional cloning, to the genetic causes of complex (including psychiatric) disorders that proved more of a challenge for linkage methods. The new technologies, together with our new knowledge of the properties of the genome, and significant efforts towards generating large patient and control sample collections, allowed for the success of genome-wide association studies. The result has been that reports currently appear in the literature every week identifying new genes for complex disorders. We are still far from completely explaining the heritable component of complex disorders, but we are certainly closer to being able to use the new information towards prevention and treatment of illness. Next-generation sequencing methods, combined with the results of association and perhaps linkage studies, will help us uncover the missing heritability and achieve a better understanding of the genetic aspects of psychiatric disease, as well as the best strategies for incorporating genetics in the service of patients.

Today, the genes responsible for the majority of Mendelian disorders are known. The Online Mendelian Inheritance in Man (OMIM; <http://www.ncbi.nlm.nih.gov/omim/>) database lists 2517 Mendelian phenotypes with known and 1741 with unknown molecular basis. This success is mainly due to the power of linkage analysis which allowed the genetic mapping of diseases to narrow genomic intervals that, even before the availability of the genome sequence, could be readily investigated for the presence of genes and mutations responsible for each disease. This valuable tool, however, proved much less effective for more common and genetically more complex disorders including psychiatric disorders. The need for more powerful tools for molecular genetic analysis became obvious and the importance of developing such tools was clear given the public health impact of these disorders. The human genome project was the first big step towards the development of technologies and tools that today allow for the successful genetic investigation of complex phenotypes. In the post-genomic era the importance of understanding such phenotypes provides the major motivation driving the emergence of new technologies. There have been many scientific breakthroughs in genetics in the last century, however, when it comes to technological breakthroughs, this is undoubtedly one of the most significant times in the history of genetics.

BRB-509, 733 N. Broadway, Baltimore, MD 21205, tel 410 955-8323, fax 410 955-7397, adimitr1@jhmi.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The genome and genetic variation

The human genome project was launched by the National Institutes of Health, the Department of Energy and international partners in 1990 and reached its first major landmark with the publication of a first working draft of the human genome in 2001 (1). The simultaneous publication of a genome draft from a parallel genome project outside the public sector (2) highlighted the tremendous technological advances that made possible ahead of schedule what originally seemed to many an overambitious undertaking. The availability of the human and other genomes subsequently led to renewed interest and further advances in the field of population genetics and provided new tools and information for the study of polymorphism, recombination, linkage disequilibrium and genetic association leading to knowledge that has been instrumental for the study of complex disorders.

One of the first things that became apparent in the study of complex disorders was that the practice of testing one or a few polymorphisms within a gene for association with a disease was at best insufficient. The number of known genetic variants increased exponentially making it clear that a gene can contain dozens or hundreds of single nucleotide polymorphisms (SNPs) that could influence its function. Although information on coding sequence and phylogenetic conservation information from the newly emerged field of comparative genomics could provide a means to assess the likelihood of function for any given SNP and reduce the number of tested SNPs, this is clearly an approach that could miss important variants.

Fortunately, the first studies that performed high throughput genotyping showed that determining the genotypes of all common SNPs is not necessary to survey all common variation (3–5). This is a result of the phenomenon of *linkage disequilibrium* (LD), an old genetic concept that came to enjoy renewed popularity. As shown in Figure 1, when a mutation arises in the population it generates a new variable location. The newly generated allele resides on a preexisting *haplotype* – a DNA strand that carries a specific sequence of alleles on the other variable positions. The new allele will then be transmitted to the next generations always on this haplotype except where the haplotype continuity is broken by recombination. As a result, the genotype of every variant in the genome is correlated with the genotypes of neighboring variants, and this correlation is reduced with increasing distance and with increasing phylogenetic age of the variant. Further, due to the non-uniform recombination rates in the genome and the existence of recombination “hot-spots” (6) the correlation between genotypes, or LD, has a patchy distribution with regions of higher LD separated, often abruptly, by regions of lower LD, a phenomenon that gave rise to the term “islands” of LD (5) (see Figure 2). Regions of high LD are often referred to as haplotype blocks, referring to short haplotype fragments that contain only some of the possible combinations of alleles across their length. These blocks should not be confused with the traditional meaning of the word haplotype, which does not assume any correlation between alleles at the population level, but only their coexistence in the same strand in an individual.

Analysis of LD in the genome has made it clear that the correlations of genotypes are often so strong that one variant can fully predict (or “tag”) the genotype of another (perfect LD-see Figure 1), making it possible to examine all common variation by genotyping only a fraction of common SNPs. The HapMap project launched in 2002 (7) has genotyped millions of SNPs in multiple populations, achieving its goal to characterize LD across the genome. These results have already been utilized to enhance the efficiency of current genotyping technologies and they are a valuable everyday tool for complex disease genetics researchers.

The sequencing of the human genome and the high throughput technologies that emerged in part as a consequence of the human genome project resulted in the identification of an additional source of genetic variation whose significance had previously remained unrecognized. Using

high throughput methods originally developed for the study of cancer Iafrate et al (8) found that the human genome of normal individuals (not affected by a specific disorder) contains multiple sites that are often deleted or vary in copy number, sites that sometimes include one or more genes. Today we know that more than 10% of the human genome is subject to copy number variation (CNV) and that these regions often include genes. Recent literature (9,10) has shown that these CNVs often influence the transcription of genes, not only those included in the CNV, but also at some distance, enforcing their perceived importance and the necessity to examine them when exploring the genome for complex disease liability loci.

In 2003 the ENCODE project was launched (<http://www.genome.gov/10005107>) with the ambitious goal to identify all functional elements in the human genome. Recently the investigators completed and published (11) results of the pilot phase of the project (for a series of resulting articles see *Genome Research* vol. 17 issue 6 at <http://www.genome.org/content/vol17/issue6/>) and the NHGRI funded new awards to scale the project to the genome level and to perform additional pilot studies. The encode data is incorporated in the UCSC genome browser together with multiple other tracks of information including the HapMap data track, providing researchers with a valuable and continuously improving tool for genome-wide or targeted genetic analyses.

Today, less than 10 years after from the completion of the Human Genome Project, technological advances have allowed the launch of the 1000 Genomes Project (12) (<http://www.1000genomes.org>), an international collaboration for sequencing the genomes of approximately 1,200 people from around the world and acquiring an unprecedented wealth of information on human polymorphism and diversity. Like the Human Genome and the HapMap project before it, the 1000 Genomes Project promises to bring this information to the fingertips of researchers, providing a tremendous push forward to human genetics research.

Linkage analyses

We mentioned above that linkage analysis led to the identification of most genes causing Mendelian genetic disorders. Such disorders are caused by rare mutations that cause disease under a specific model (for example autosomal dominant) with a high penetrance (close to 100% of mutation carriers are affected) and with very rare phenocopies (almost everyone affected has the same genetic condition). Unfortunately these criteria are not met by the more common human disorders that, although usually milder than Mendelian disorders, represent overall a much bigger burden to our society. Psychiatric diseases, which are both common and complex, are often characterized by a relatively early age of onset and require lifelong treatment. As treatments are often only partially effective these diseases have devastating consequences for the quality of life of the patients and their caregivers.

When linkage was first applied to complex disorders it was performed using the established methods of the time, which required determining, *a priori*, an inheritance model. Many early efforts failed to find linkage and often declared exclusion of parts of the genome based on such analyses. It was slowly becoming clear however that the common disorders were much more likely to involve more than a single gene and that the “mutations” involved were likely to cause disease in some but not all carriers, consistent with prior segregation analyses results for schizophrenia (13). Traditional linkage methods were not meant for the investigation of such diseases. Numerous efforts for the development of new approaches to linkage analysis were undertaken in the 1990’s, most of which involved what eventually became the standard for linkage analysis of complex disorders, the analysis of shared alleles in each genomic location between affected relatives compared to the expected allele sharing for each type of pair of relatives. This makes a sacrifice in power, but avoids assumptions about inheritance, bypassing a significant problem of parametric linkage. At the same time, technologies were developed

allowing genotyping of short tandem repeat (STR) markers without polyacrylamide gels or radioactivity and with extensive marker multiplexing. It became possible to survey the genome at a cost and time investment that was realistic. Whole genome linkage studies were soon being published monthly or weekly. The results, however, did not deliver on the expectations. Linkage signals were most often weak, not providing the strong evidence for disease loci everyone hoped for. In 1995 Lander and Kruglyak (14) formalized the criteria for declaring genome-wide significant linkage. Very few studies met these criteria and the few loci that did almost invariably did not replicate in other samples. Complex disease genetics and psychiatric genetics were going through a confidence crisis.

Just one year later, in a paper that had a very strong influence in the field of complex disease genetics, Risch and Merikangas (15) showed that, if complex disorders were due to common genetic variants with small effects, linkage studies would need to examine many thousands of families to identify such loci. They also showed that performing genetic association studies rather than linkage was far more powerful for discovering these loci. At that time the genotyping technologies did not allow a genome-wide scan for association and such studies were only performed in small scale for individual candidate genes and were often reserved for following up linkage signals. Today, through the use of microarrays and other similar array-based technologies, genotyping a million SNPs across the genome has become both fast and affordable and genome-wide association studies (GWAS) have mostly replaced linkage in the exploration of complex disorders. The few linkage studies that are now performed use SNP markers, genotyped in large numbers to match and exceed the high information content of microsatellites. The genetic effect sizes of common alleles uncovered by most GWAS are so small that one would need millions of families to detect them by linkage, which seems to be a justification for favoring GWAS. One, however, needs to be careful when comparing linkage and association. As opposed to association studies, linkage would detect a disease locus with multiple rare disease risk alleles, provided their combined effect is substantial. In view of the most recent GWAS results where the identified variants do not explain much of the heritability of the respective disorders (16), linkage is starting to look attractive again as it could provide target regions for next generation sequencing and allow the inference of mutation segregation across pedigrees, leading to the discovery of disease variants not visible by GWAS.

Genome-Wide Association Studies

The concept of GWAS was formed in the 1990's and supported by theoretical work such as that of Risch and Merikangas (15), however at that time genotyping the genome at an adequate density was not possible. Association studies were performed for one, ten or sometimes 100 SNPs, covering targeted genes or regions. Claims of "lack of association" were obviously exaggerated and to everyone's frustration positive findings rarely replicated and almost never consistently. Genome-wide scans became possible around the turn of the millennium, initially at low marker densities and often with sample pooling to reduce cost. Today, before the end of the decade, studies testing 1 million SNPs across the genome on many thousands of subjects are published weekly. There have been many success stories that have confirmed the value of the GWAS approach. In just the last three years GWAS have led to hundreds of associations of common DNA variants with over 80 diseases and traits (17) including schizophrenia, bipolar disorder, major depression, autism, attention deficit disorder, neuroticism, alcohol dependence, smoking behavior and other psychiatric or related disorders and traits (see <http://www.genome.gov/gwastudies/>). In addition to connecting genetic variation to disease we have learned many lessons from these studies. The newly identified variants are common and have small effect sizes, typically allelic odds ratios not exceeding 1.5 (18). The implicated genes are most often not among those considered candidates, and they are often involved in multiple disorders that were previously thought to be unrelated (18). Although the identified associations involve non-synonymous coding variants and 5 kb gene promoter regions more

often than expected by chance, 88% are in the intronic or intergenic space, and only slightly more often than expected in conserved regions (17). Although it is thought, and in many cases has been shown, that many of those trait-associated variants are involved in gene regulation (19), in the majority of cases the functional link between the DNA variant and disease biology remains unclear.

Despite the tremendous success of GWAS in the last few years there remain skeptics of the value of the approach. Some question the value of observing associations of DNA variants to disease in the absence of knowledge of the underlying biology. The answer to such reservations is simple. GWAS should only be considered a first step toward the identification of causal relationships between gene biology and disease. There is much work that needs to be done after a GWAS in order to move the knowledge through basic research to translational research and into the clinic. Others argue that, since the identified genetic variants only explain a small fraction of the disease heritability, they are of limited value. It is true that the heritability explained by variants identified through GWAS is most often less than 2–3% of the disease heritability (20), a phenomenon that has triggered much discussion about the underlying reasons, which might include a great number of small effect variants, gene-gene interactions, multiple individually rare variables that are not detectable by GWAS, copy number variations (a type of variation that remains incompletely examined although it represents a significant fraction of inter-individual variation) and perhaps epigenetic variation (16,21,22). Whichever the case, it is our view that the fact that they explain little of the phenotypic variance takes nothing away from the value of the identified associations. Knowing a gene is involved in a disease provides knowledge about the disease mechanisms and potential therapeutic targets, especially once the relationship between the DNA variants and the gene's function in the relevant pathways is determined. Although the normally occurring variation within a gene might influence the risk only by little, pharmacological intervention targeting that gene's function could have a significant impact. For example, although genetic associations between PPAR gamma and diabetes mellitus are relatively weak, with odds ratios below 1.2 (23), PPAR gamma is the target of thiazolidinediones, a class of oral antidiabetic agents (24). Another striking example is the LDL receptor where, while mutations only account for a small fraction of hyperlipidemias their discovery has been instrumental to the development, and our understanding of the mechanism of action, of statins that revolutionized hyperlipidemia treatment (25).

The insights from GWAS specific to psychiatric disorders are similar, but they include some additional interesting observations, suggesting new ways to look at GWAS data. One is the possible involvement of rare copy number variations in schizophrenia and autism, a possibility that has received substantial support (26–37). Another is the possible involvement of thousands of genes with very small effects in schizophrenia, genes that overlap significantly with those involved in other psychiatric disorders like bipolar disorder (38). As shown in the paper from the International Schizophrenia Consortium (38), a risk score can be generated from multiple variants with weak associations with schizophrenia (certainly including many false positives), and such a score will differ when comparing controls to schizophrenic or bipolar patients, but not to patients with non-psychiatric disorders. Although, once again, the explained variance from such a score was small the approach provides a glimpse of the power that genetics will have when such groups of variants become free of false positives and of how dissecting such groups of variants/genes will help us understand how different disorders both overlap and differ.

Genome wide gene expression (GWGE) studies

Microarray technologies provided a strong push forward for high-throughput genotyping, however they were originally introduced for the high-throughput analysis of gene expression.

They were a natural evolution of classic DNA and RNA hybridization methods like Southern and Northern blotting through technological advances that made it possible to print thousands of features on a chip and hybridize them to the RNA (or cDNA) under investigation, allowing the quantification of gene transcripts across the genome and comparisons between healthy and diseased tissues. What could be considered the first expression array experiment was performed as early as 1987 by Kulesh et al (39) searching for genes responding to interferon. The first microarray experiment was reported by Schena et al (40) in 1995, followed by thousands of other reports.

GWGE studies can be useful for the identification of genes involved in a disease as they can point to the genes whose expression differs between patients and controls, presumably as a cause or consequence of the disease process, or genes that respond to disease-relevant medication. The value of such studies to our understanding of the disease biology is significant and the knowledge of disease expression profiles can have many applications. It is important to remember that a gene that is found to have altered transcript levels in a disease is not necessarily a culprit in its pathogenesis. Although one can clearly claim a relationship between the gene and the disease process, this might be distant and mediated by indirect associations, so that the gene might bear no meaningful connection to the disease risk. When GWGE studies are seen in the context of their potential limitations, they are invaluable for our understanding of complex disorders.

More recently, GWGE studies have also been used to identify functional DNA variation in the genome, variation that influences the expression or splicing of genes. The data generated from such studies is expected to be very useful for the study of complex disorders as it will allow the assessment of the functionality of variants identified through GWAS. In a recent review Cookson et al (19) calculated that 10–15% of GWAS signals involved a known regulatory variant, often called an expression quantitative trait locus (eQTL). This number will likely increase as we identify more eQTLs through new gene expression mapping studies. The NIH has recognized the necessity of linking gene regulation to genetic variation and has recently launched the Genotype-Tissue Expression (GTEx) project (<http://nihroadmap.nih.gov/GTEx/>), a pilot project that will ultimately lead to a gene regulation database involving multiple tissues from 1,000 donors with genome-wide variation information, a much anticipated resource.

Many limitations, potential pitfalls and confounding factors must be considered when interpreting the information provided by GWGE studies. Unlike analyzing constitutional DNA which is expected to be almost identical across cell types and tissues (with the notable exception of tumors), the source of RNA is a defining parameter in GWGE studies. Different tissues and different cell types have distinctly different expression profiles and likely involve different regulatory mechanisms. Solid tissues contain multiple cell types that are hard to separate and, although techniques like laser capture microdissection (41) have been developed for this purpose, their use in large-scale gene expression studies is not always practical. As a result, one must be aware that the observed expression profile is often the summation from more than one cell type and in certain cases (for example in the brain in neurodegenerative disorders) the relative cell type abundance might differ between cases and controls. Further, the acquisition of some types of tissues (for example, brain) is almost entirely limited to postmortem specimens, introducing additional confounding parameters like the cause of death and the delay in the dissection of the tissue, which could lead to RNA degradation. Many studies attempt to bypass such limitations by examining lymphocytes or immortalized cell lines that are often readily available or easy to acquire. This reduces significant sources of variation, however it might be of limited value for the study of a psychiatric disease. Although it can be argued that some of the gene expression differences will be similar in brain and lymphocytes, only a fraction of genes will be expressed in both tissues and those are likely to be subject to different

regulatory mechanisms in tissues that have different functions. Many other limitations including alternative splicing, environmental factors like diet and drugs, as well as platform-specific characteristics introduce additional levels of complexity and highlight the importance of careful experimental design in GWGE studies.

In the more than 20 year history of expression array analysis we have learned a lot from GWGE studies and have developed tools to help us tackle their complexities and interpret their results. Those results combined with the data from GWAS will provide one more powerful tool for untangling the genetics of complex diseases.

Epigenetics

The term epigenetics refers to changes that involve the genetic material and lead to phenotypic changes but do not alter the DNA sequence. Epigenetic changes mainly include the methylation of DNA and modifications of chromatin like methylation and acetylation of the histones, the DNA's packaging material. Epigenetic changes are acquired during the life of an organism and they are important for gene regulation, with big differences observed in epigenetic marks across different tissues. Environmental factors can also influence epigenetic marks through life, before they are reprogrammed in gametogenesis and early embryogenesis (42). Occasionally epigenetic changes can escape reprogramming and be vertically transmitted across generations and as a result an acquired epigenetic state can persist in the next generation, in other words it can be inherited. The involvement of epigenetic modifications in cancer is well known (43) and their potential importance in complex diseases has been argued by many (44–46), and it has been suggested that such epigenetic changes could be a source of missing heritability (22). Although this view has been challenged (47) it is clear that epigenetic variation can be causally linked to complex diseases including psychiatric disease, and recognizing the interplay between epigenetics and genetics will help us in the discovery of complex disease genes.

There are now multiple tools available that allow assessing epigenetic variation across the genome, based on modification of methylated DNA or chromatin immunoprecipitation and microarray hybridization methods (ChIP-chip), the latter now being replaced by modern high throughput sequencing methods (ChIP-seq, see Figure 3) (48) (49,50). These tools, when appropriately applied to complex disorders, will likely lead to significant new discoveries of the mechanisms of disease, like they have for cancer. Many of the same problems encountered in GWGE studies also need to be considered here, including the differences between tissues and cell types that could mask differences or confound results. Additional complexity is added by the many different types of histone modification one needs to examine to obtain a thorough epigenetic assessment. Yet, however many the complexities, the promise is great, especially as pharmacological agents that can affect epigenetic modifications are already available (51).

Next-generation sequencing

For more than three decades the leading method for DNA sequencing was that described by Sanger in 1977 (52) based on oligonucleotide primed DNA synthesis and dideoxy termination and it was the method used for sequencing the human genome. Sanger sequencing remains the tool of choice in the modern laboratory for small scale DNA sequencing, however new technologies that evolved over the last few years, collectively termed next generation sequencing, have taken over in the field of high throughput sequencing (53). These new technologies, that include platforms like Roche 454 (Roche, CT, USA) Genome Analyzer (Illumina/Solexa, CA, USA), SOLiD™ (Applied Biosystems, CA, USA), Heliscope™ (Helicos Biosciences, MA, USA) and SMRT (Pacific Biosciences, CA, USA), each have specific advantages and one common feature that clearly set them apart from Sanger sequencing; they are highly parallel generating 500 Mb to 40 Gb of sequence in every run. These new technologies have allowed the price of sequencing a genome to drop by many orders of

magnitude (an estimate for the human genome today is under \$100,000) and have made possible research projects that involve sequencing extensive genomic regions across multiple individuals. Beyond this revolution in sequencing capacity, they have allowed new approaches to genomics and epigenomics, including RNA-seq analysis for gene expression (54) (see Figure 3), ChIP-seq analysis for DNA-protein interactions (50) and metagenomics applications for the study of biodiversity by sampling specific environmental niches (55). All of the above applications take advantage of the new technologies' high throughput as well as the ability to generate sequence from DNA libraries without prior knowledge of the content.

RNA-seq is the application of next generation sequencing to survey the content of the transcriptome (54). It is a step up from previously applied sequencing methods for gene expression like SAGE (56). It can provide a quantitative assessment of transcription including sequence information across the transcriptome. That allows the identification and assessment of normal or disease-causing alternative splicing, the discovery of coding variation including pathogenic mutations, and the assessment of allele-specific expression differences revealed by transcribed polymorphisms. It is clear that in the study of disease this approach has significant advantages over hybridization-based microarray analysis.

ChIP-seq (57,58) is the sequencing of DNA that has been reversibly bound to protein and pulled down by immunoprecipitation, an evolution of ChIP-chip analysis which used hybridization on tiled arrays instead of sequencing (see Figure 3). Although the name implies that the proteins of interest are in the context of chromatin, any DNA associating protein can be analyzed. The method allows the identification of the DNA sequences at the location of interaction, and it has been used extensively in the study of epigenetic modifications where the modified histones are used for immunoprecipitation, but also in the study of nucleosome positioning and transcription factor binding. The additional information provided by ChIP-seq compared to ChIP-chip will enhance and accelerate our functional annotation of the genome sequence and bring us one step closer to linking DNA variation to function and to disease.

Next generation sequencing also allows the complete sequencing of large genomic regions around association signals or linkage peaks for multiple individuals. The large number of DNA variants such studies identify, especially variants in intergenic space, make it difficult to determine their likely role in disease. However many of the genomic approaches we mentioned here, including gene regulation mapping which can utilize RNA-seq, epigenomic and other DNA-protein interaction analyses utilizing ChIP-seq, together with a multitude of other laboratory and *in silico* approaches to genetic function that we do not have the space to discuss, will provide guides that will help us sort through the identified sequence variants and identify those that underlie disease.

Conclusion

Despite significant challenges in the genetics of psychiatric and other complex disorders the sequencing of the human genome, together with a series of advances in biotechnology, are leading to new gene discoveries and the recognition of new disease mechanisms. As the pace of discovery accelerates, the geneticist's tool box is also becoming richer and the future looks more promising than ever. The practice of psychiatry has long suffered from the limited information available on the biological basis of mental disorders, as compared to other conditions. This limitation is now coming to an end, and exciting new possibilities are on the horizon for psychiatry in the 21st century.

Acknowledgments

The author thanks Megan Szymanski for critical review of the manuscript. DA is supported by funding provided from the National Institute of Aging (grant R01AG022099 to DA) and an award from the Neurosciences Education and Research Foundation.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921. [PubMed: 11237011]
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. The sequence of the human genome. *Science* 2001;291:1304–51. [PubMed: 11181995]
3. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. High-resolution haplotype structure in the human genome. *Nat Genet* 2001;29:229–32. [PubMed: 11586305]
4. Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, et al. Linkage disequilibrium in the human genome. *Nature* 2001;411:199–204. [PubMed: 11346797]
5. Goldstein DB. Islands of linkage disequilibrium. *Nat Genet* 2001;29:109–11. [PubMed: 11586289]
6. Arnheim N, Calabrese P, Tiemann-Boege I. Mammalian meiotic recombination hot spots. *Annu Rev Genet* 2007;41:369–99. [PubMed: 18076329]
7. International HapMap Consortium. The International HapMap Project. *Nature* 2003;426:789–96. [PubMed: 14685227]
8. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. Detection of large-scale variation in the human genome. *Nat Genet* 2004;36:949–51. [PubMed: 15286789]
9. Cahan P, Li Y, Izumi M, Graubert TA. The impact of copy number variation on local gene expression in mouse hematopoietic stem and progenitor cells. *Nat Genet* 2009;41:430–7. [PubMed: 19270704]
10. Henrichsen CN, Vinckenbosch N, Zollner S, Chaignat E, Pradervand S, Schutz F, Ruedi M, Kaessmann H, Reymond A. Segmental copy number variation shapes tissue transcriptomes. *Nat Genet* 2009;41:424–9. [PubMed: 19270705]
11. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799–816. [PubMed: 17571346]
12. Siva N. 1000 Genomes project. *Nat Biotechnol* 2008;26:256. [PubMed: 18327223]
13. Risch N, Baron M. Segregation analysis of schizophrenia and related disorders. *Am J Hum Genet* 1984;36:1039–59. [PubMed: 6496472]
14. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995;11:241–7. [PubMed: 7581446]
15. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996;273:1516–7. [PubMed: 8801636]
16. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53. [PubMed: 19812666]
17. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 2009;106:9362–7. [PubMed: 19474294]
18. Manolio TA, Brooks LD, Collins FS. A HapMap harvest of insights into the genetics of common disease. *J Clin Invest* 2008;118:1590–605. [PubMed: 18451988]
19. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 2009;10:184–94. [PubMed: 19223927]
20. Goldstein DB. Common genetic variation and human traits. *N Engl J Med* 2009;360:1696–8. [PubMed: 19369660]
21. Maher B. Personal genomes: The case of the missing heritability. *Nature* 2008;456:18–21. [PubMed: 18987709]

22. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* 2008;17:R156–65. [PubMed: 18852205]
23. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;316:1341–5. [PubMed: 17463248]
24. Day C. Thiazolidinediones: a new class of antidiabetic drugs. *Diabet Med* 1999;16:179–92. [PubMed: 10227562]
25. Goldstein JL, Brown MS. The LDL receptor. *Arterioscler Thromb Vasc Biol* 2009;29:431–8. [PubMed: 19299327]
26. International Schizophrenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 2008;455:237–41. [PubMed: 18668038]
27. Stefansson H, Rujescu D, Cichon S, Pietilainen OP, Ingason A, Steinberg S, Fossdal R, Sigurdsson E, Sigmundsson T, Buizer-Voskamp JE, et al. Large recurrent microdeletions associated with schizophrenia. *Nature* 2008;455:232–6. [PubMed: 18668039]
28. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* 2008;320:539–43. [PubMed: 18369103]
29. Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M. Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 2008;40:880–5. [PubMed: 18511947]
30. Kirov G, Gumus D, Chen W, Norton N, Georgieva L, Sari M, O'Donovan MC, Erdogan F, Owen MJ, Ropers HH, et al. Comparative genome hybridization suggests a role for NRXN1 and APBA2 in schizophrenia. *Hum Mol Genet* 2008;17:458–65. [PubMed: 17989066]
31. Vrijenhoek T, Buizer-Voskamp JE, van der Stelt I, Strengman E, Sabatti C, Geurts van Kessel A, Brunner HG, Ophoff RA, Veltman JA. Recurrent CNVs disrupt three candidate genes in schizophrenia patients. *Am J Hum Genet* 2008;83:504–10. [PubMed: 18940311]
32. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, et al. Strong association of de novo copy number mutations with autism. *Science* 2007;316:445–9. [PubMed: 17363630]
33. Christian SL, Brune CW, Sudi J, Kumar RA, Liu S, Karamohamed S, Badner JA, Matsui S, Conroy J, McQuaid D, et al. Novel submicroscopic chromosomal abnormalities detected in autism spectrum disorder. *Biol Psychiatry* 2008;63:1111–7. [PubMed: 18374305]
34. Kumar RA, Karamohamed S, Sudi J, Conrad DF, Brune C, Badner JA, Gilliam TC, Nowak NJ, Cook EH Jr, Dobyns WB, et al. Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet* 2008;17:628–38. [PubMed: 18156158]
35. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, et al. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med* 2008;358:667–75. [PubMed: 18184952]
36. Cho SC, Yim SH, Yoo HK, Kim MY, Jung GY, Shin GW, Kim BN, Hwang JW, Kang JJ, Kim TM, et al. Copy number variations associated with idiopathic autism identified by whole-genome microarray-based comparative genomic hybridization. *Psychiatr Genet* 2009;19:177–85. [PubMed: 19407672]
37. Glessner JT, Wang K, Cai G, Korvatska O, Kim CE, Wood S, Zhang H, Estes A, Brune CW, Bradfield JP, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature* 2009;459:569–73. [PubMed: 19404257]
38. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 2009;460:748–52. [PubMed: 19571811]
39. Kulesh DA, Clive DR, Zarlenga DS, Greene JJ. Identification of interferon-modulated proliferation-related cDNA sequences. *Proc Natl Acad Sci U S A* 1987;84:8453–7. [PubMed: 2446323]
40. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995;270:467–70. [PubMed: 7569999]
41. Best CJ, Emmert-Buck MR. Molecular profiling of tissue samples using laser capture microdissection. *Expert Rev Mol Diagn* 2001;1:53–60. [PubMed: 11901800]

42. Rakyan V, Whitelaw E. Transgenerational epigenetic inheritance. *Curr Biol* 2003;13:R6. [PubMed: 12526754]
43. Feinberg AP, Tycko B. The history of cancer epigenetics. *Nat Rev Cancer* 2004;4:143–53. [PubMed: 14732866]
44. Feinberg AP. Epigenetics at the epicenter of modern medicine. *Jama* 2008;299:1345–50. [PubMed: 18349095]
45. Ptak C, Petronis A. Epigenetics and complex disease: from etiology to new therapeutics. *Annu Rev Pharmacol Toxicol* 2008;48:257–76. [PubMed: 17883328]
46. Jiang Y, Langley B, Lubin FD, Renthall W, Wood MA, Yasui DH, Kumar A, Nestler EJ, Akbarian S, Beckel-Mitchener AC. Epigenetics in the nervous system. *J Neurosci* 2008;28:11753–9. [PubMed: 19005036]
47. Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics* 2009;182:845–50. [PubMed: 19416939]
48. Shendure J. The beginning of the end for microarrays? *Nat Methods* 2008;5:585–7. [PubMed: 18587314]
49. Park PJ. Epigenetics meets next-generation sequencing. *Epigenetics* 2008;3:318–21. [PubMed: 19098449]
50. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10:669–80. [PubMed: 19736561]
51. Papait R, Monti E, Bonapace IM. Novel approaches on epigenetics. *Curr Opin Drug Discov Devel* 2009;12:264–75.
52. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 1977;74:5463–7. [PubMed: 271968]
53. Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;9:387–402. [PubMed: 18576944]
54. Morozova O, Hirst M, Marra MA. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet* 2009;10:135–51. [PubMed: 19715439]
55. Blow N. Metagenomics: exploring unseen communities. *Nature* 2008;453:687–90. [PubMed: 18509446]
56. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science* 1995;270:484–7. [PubMed: 7570003]
57. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 2007;4:651–7. [PubMed: 17558387]
58. Mardis ER. ChIP-seq: welcome to the new frontier. *Nat Methods* 2007;4:613–4. [PubMed: 17664943]

Figure 1.

Linkage disequilibrium. Three SNPs and their observed haplotypes in the population are shown. A new mutation generates an additional variant on one of the haplotypes. In the simplistic example shown here the mutation is in **perfect** LD with SNP2 ($r^2=1$) and any disease association would also be observed with that SNP. The other SNPs are correlated with the mutation, but not perfectly ($r^2 < 1$). As long as recombination does not break this haplotype block, the variants remain in **complete** LD ($D'=1$) and certain combinations of alleles are never observed.

Figure 2.

Example of LD islands. A 100 Kb region around the *APOE* gene is shown using data from the HapMap project and the Haploview software (www.broadinstitute.org/mpg/haploview). Colored diamonds show the strength of LD between the SNPs corresponding to the extension of their sides up to the SNP line-up at the top. Red is for strong LD (high D') while purple signifies insufficient information to infer LD. Islands of LD are revealed as large red triangles. An island of strong LD is labeled.

Figure 3.

ChIP-chip vs. ChIP-seq. The first three steps are identical. DNA is cross-linked to chromatin and fragmented and then antibodies specific to modifications of interest are used for immunoprecipitation (IP). Cross links are reversed and the selected DNA is analyzed for content. This analysis was traditionally done through hybridization on a chip (hence ChIP-chip) and is now often done by next-generation sequencing (ChIP-seq).