Methods for Investigating the Genomic Basis of Complex Diseases

Increasingly, scientists are finding that many diseases—from cancer to rare disorders—contain a genetic component. Some disorders result from mutations in a single gene, while others, known as complex diseases, do not obey the single-gene dominant or single-gene recessive Mendelian patterns of inheritance. They include Alzheimer's disease, asthma, Parkinson's disease, multiple sclerosis, diabetes, heart conditions, autoimmune disorders, and others.[1] Complex diseases arise from a combination of genetic and environmental factors, many of which are not understood (Figure 1).

Recent advances in genomic tools have enabled researchers to look more closely at the genetic variants that contribute to complex diseases.[2] From the Human Genome Project[3] to the International HapMap Project,[4] researchers have collaborated to understand human genetic variation. Recently, the Encyclopedia of DNA Elements Consortium (ENCODE) project explored transcription, regulation, DNA–protein binding, and epigenetic patterns to provide functional information about the human genome.[5] These endeavors have been made possible by developments in technology that enable investigators to assess many genes at one time.

Due to the multifactorial nature of complex diseases, pinpointing the cause or causes of a disorder is challenging. The presence of multiple molecular targets—including genes, coding and noncoding RNA, and proteins—potentially contributing to the disease requires flexible, accurate tools for assessing all these factors at one time. High-throughput genomic technologies such as microarrays and next-generation sequencing (NGS) are introducing new approaches to understanding the etiology of complex diseases on a molecular level, providing detailed insight into the functional consequences of variation.

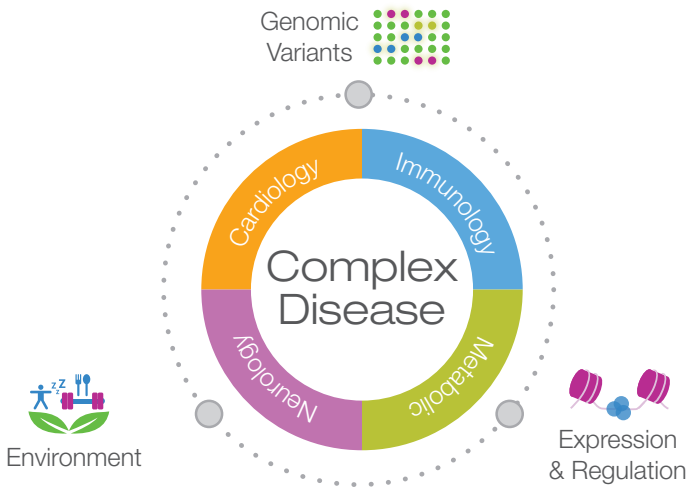

Figure 1: Genetic and Environmental Influences on Complex Disease. A combination of genetic and environmental factors contributes to complex disease phenotypes. Many have not yet been identified. Analyzing how these factors influence each other can provide a deeper understanding of disease biology, ultimately assisting researchers with developing targeted approaches to managing disease.

illumina®

Methods for Causal Variant Discovery

With the advent of high-throughput genomic technologies, an increasing amount of research is dedicated to investigating the connection between the human genome and disease. Studies continue to reveal associations between specific genetic variants and disease phenotypes.[6]

Identifying Single-Base Changes

Microarrays are traditionally used to measure gene expression levels or genotype multiple regions in a genome. Using array-based technology to detect single nucleotide polymorphisms (SNPs) is an efficient method for variant discovery. Large genome-wide association studies (GWAS) enable researchers to interrogate large sample sizes to detect variants that occur at a significant frequency in a diseased cohort relative to a healthy cohort. Array-based genomic surveys can generate data from tens of thousands of samples, providing the statistical confidence needed for disease association. GWAS can identify large numbers of common variants associated with a disease in a single assay, uncovering multiple genes for further study.

Recent findings show that rare variants have implications in complex diseases.[7,8] While SNP arrays are effective for common variant identification, the loci selection required for SNP arrays can limit detection of rare variants occurring at a low frequency within a population. Sequencing the genomes or exomes—the protein-coding portion of the genome—of individuals or families is becoming a more common approach for finding causative variants in rare or complex disease cases.[9–11] Researchers can use exome sequencing to investigate the coding regions of the genome when sequencing an entire genome is not practical or necessary. NGS is an unbiased approach to variant detection that evaluates all loci and alleles, regardless of prior expectations. Additionally, sequencing benefits trio and pedigree analyses by requiring small sample sizes and providing complete genomic information, including detection of potential *de novo* mutations.

While sequencing can be used on a case-by-case basis to determine the variant responsible for a particular phenotype in a single occurrence, large sample numbers are necessary to understand the role of genetic variation in a population. The HiSeq X™ Ten Sequencing System provides the high throughput needed for large-scale whole human genome sequencing, contributing to the understanding of genomic variation as studies progress from GWAS to population sequencing. Large-scale sequencing can potentially uncover more rare variants than traditional methods, providing a closer look at genomic complexities that were previously hidden.

Detecting Large Structural Variations

Copy number variations (CNVs) are genomic alterations that result in an abnormal number of copies of one or more genes. Structural genomic rearrangements such as duplications, deletions, translocations, and inversions can cause CNVs. Like SNPs, certain CNVs on particular chromosomes affecting specific genes have been associated with disease susceptibility.[12,13]

Array-based approaches for detecting CNVs offer reliable, efficient methods for large-scale analysis. Researchers can process multiple samples on a single microarray for broad surveys of genomic structural variation. Designed by an international community of experts, the Infinium® CytoSNP-850K BeadChip provides comprehensive coverage of cytogenetically relevant genes correlated with high disease risk, benefiting studies of constitutional disorders and cancer. With the CytoSNP-850K BeadChip, researchers can profile chromosomal aberrations such as amplifications, deletions, rearrangements, and copy-neutral loss of heterozygosity accurately. For broader studies of disease-associated variants, Illumina Omni microarrays offer an unbiased, non-targeted approach, providing high coverage across the human genome.

Due to marker spacing, genotyping arrays cannot detect CNVs that are smaller than 50 kilobases (kb). By providing a base-by-base view of the genome, NGS detects small CNVs that microarrays often miss. With sequencing, researchers can investigate CNVs as small as a few hundred base pairs, which can be useful for studies of missing heritability in complex disease. The detailed resolution offered by sequencing complements the high-throughput

capability of arrays, providing a comprehensive view of both large and small variations in the genome. By leveraging both technologies, researchers can obtain a complete view of genetic variation to drive discovery and enhance research (Figure 2).

Figure 2: A Complete View of the Genome with NGS and Microarrays. When used together, NGS and microarrays provide a comprehensive view of genomic variation, gene expression patterns, and downstream functional consequences. Genotyping results can spur sequencing studies, or vice versa, driving discovery and informing study design. By combining genotype information with epigenetic signatures and RNA transcription patterns, researchers can characterize the effects of environmental changes or genetic variation on disease biology.

Approaches to Validating Candidate Variants

After genetic variants are identified and associated with a particular disease phenotype using GWAS or sequencing, further study is necessary to refine the loci of interest and ultimately understand their biological effects. A genetic variant might change the structure of a protein directly, affect regulation of gene expression, or have no effect on RNA transcription and translation into protein. Array- and sequencing-based methods can help to determine the molecular consequence of a genetic variant.

Fine Mapping of Associated Variants

Fine mapping of genomic regions of interest can clarify the biological effects of genetic variation.[14] With current genotyping technologies, the ability to detect disease association depends on the linkage of the genotyped markers with an adjacent disease-causing SNP. Linkage disequilibrium—the non-random association of alleles at two or more loci—implies that the alleles are located near each other and are inherited together as a result.

Linkage disequilibrium can complicate disease studies. After identifying candidate SNPs through GWAS, researchers must make sure that the association between the SNP and the disease is a true association rather than a result of linkage disequilibrium. Further studies focusing on a variant of interest can clarify the connection between SNPs and phenotype. With nucleotide-level resolution, targeted resequencing of these candidate regions provides a closer look at the molecular consequences of each variant, such as predicted changes to protein function. Targeted approaches, such as TruSeq® Custom Amplicon assays, enable researchers to sequence many gene amplicons in a single reaction, providing an efficient method for discovering, validating, and screening variants of interest.[15]

In addition to targeted resequencing, custom genotyping arrays can be used to investigate candidate SNPs further. Illumina iSelect® custom genotyping BeadChips offer the ability to interrogate virtually any SNP. Custom arrays enable researchers to choose only relevant content for the disease under study and increase SNP density in regions of interest, conserving resources by avoiding irrelevant regions of the genome.

Consortium formation offers another opportunity to design arrays according to focused, custom content of interest to advance research for a particular condition. Consortia reduce the fiscal barrier of entry to genomics for individual researchers, as members of a consortium work together and pool resources to improve the collective understanding of human disease. Through collaboration, consortium members can develop a genotyping tool that meets the needs of multiple researchers, resulting in significant cost savings.

Understanding How Variants Affect Phenotype

Although many genetic variants have been associated with complex diseases, gene mutations do not always affect protein function. Most genomic variants lie outside coding regions and are putative regulators of gene expression.[16] By characterizing the downstream effects of identified variants, researchers can better understand the molecular mechanisms of disease, leading to improved development of potential targeted therapies.

Current methods for investigating RNA involve expression quantitative trait loci (eQTL), genomic loci that regulate mRNA expression. Standard methods for eQTL mapping test the linkage between variation in expression and genetic polymorphisms. Analysis of eQTL indicates how a variant affects expression. Microarrays offer high throughput for screening large numbers of samples to identify variants or expression patterns of interest, and subsequent focused sequencing studies can characterize their biological effects.

RNA sequencing (RNA-Seq) offers several advantages for functional genomics studies. Unlike microarrays, which measure probe intensities from entire genes, RNA-Seq quantifies RNA activity at single-base resolution, capturing subtle gene expression changes. RNA-Seq enables researchers to identify and quantify both rare and common transcripts, sequence splice junctions, and detect differential isoforms, novel transcripts, and gene fusions. Sequencing mRNA using NGS can identify strand information, indicating from which of the two complementary DNA strands a particular mRNA transcript was derived. Strand orientation allows identification of antisense expression, an important mediator of gene regulation.[17] The ability to capture the relative abundance of sense and antisense expression provides visibility into regulatory interactions that might otherwise be missed. Without requiring probe or primer design, RNA-Seq offers an unbiased method for investigating the transcriptome.

Gene Expression and Regulation

Transcriptome analysis provides a stronger indication of functional changes in a cell compared to genome analysis alone, but RNA transcription patterns do not always reflect protein translation. Many other factors are involved in the translation of RNA into protein. Gene expression regulation by epigenetic modifications, small RNAs, and other factors can alter protein translation, which in turn affects intracellular dynamics and can contribute to a disease phenotype.

Ribosome Profiling

Ribosome profiling, a novel technique for investigating translational control, relies on deep sequencing of ribosome-protected mRNA fragments to provide a complete view of the ribosomes that are active in a cell at a given time. This information can then be used to determine which proteins are actively translated in a cell. ARTseq™ Ribosome Profiling Kits (Epicentre, an Illumina company) produce sequencing-ready libraries from mRNA so that researchers can measure gene expression, identify translation start sites, predict protein abundance, and investigate translational and co-translational processes *in vivo*. This method offers a complete view of the functional consequences of genomic and transcript variants for a better understanding of the molecular mechanisms behind disease phenotypes.

Noncoding RNA Analysis

Small, noncoding RNA species, such as microRNAs (miRNAs), play a critical role in regulating gene expression through interaction with the enzyme Dicer and the RNA-induced silencing complex (RISC). Small RNA analysis reveals the differential expression of all small, noncoding RNAs in any sample without prior sequence or secondary structure information. Sequencing detects miRNAs and other small RNA molecules that influence gene silencing and expression, as well as novel miRNA targets.

Epigenetic Regulation

DNA–Protein Binding

In some disease cases, environmental factors can alter histone proteins in chromosomes and change trait behavior.[2] Chromatin immunoprecipitation sequencing (ChIP-Seq) can be used to survey interactions between protein, DNA, and RNA. ChIP-Seq using NGS enables researchers to identify the binding sites of multiple protein targets, including transcription factors and histones, across the entire genome. Analysis of DNA–protein interactions provides insight into regulation events that are critical for many biological processes and disease states. With ChIP-Seq, researchers can better understand how chromatin modifications and local structural changes affect gene expression.

Methylation

DNA methylation plays an important and dynamic role in regulating gene expression and chromatin remodeling. The presence or absence of cytosine methylation can trigger or silence RNA transcription. Aberrant DNA methylation and its impact on gene expression have been implicated in many disease processes, including Alzheimer's disease,[18] and research into methylation and other epigenetic patterns continues to grow (Figure 3).

Both microarrays and NGS offer the ability to investigate methylation patterns quantitatively. Arrays enable researchers to survey methylation sites across many genomes. Designed by experts, the Infinium HumanMethylation450 BeadChip is ideal for epigenome-wide association studies with large sample sizes. Sequencing-based methylation analysis applies the high coverage density and flexibility enabled by NGS for greater clarity in epigenetic studies. It relies on bisulfite sequencing, which uses a sodium bisulfite conversion reaction to distinguish between methylated and non-methylated cytosine bases. The EpiGnome™ Methyl-Seq Kit (Epicentre, an Illumina company) uses a unique post-bisulfite conversion method that generates high-quality sequencing libraries for characterizing CpG methylation. Whole-genome methylation sequencing provides a comprehensive view of methylation patterns at single-base resolution across an entire genome. By leveraging the benefits of sequencing- and array-based technologies, researchers can gain a holistic view of functional changes that affect disease biology (Table 1).

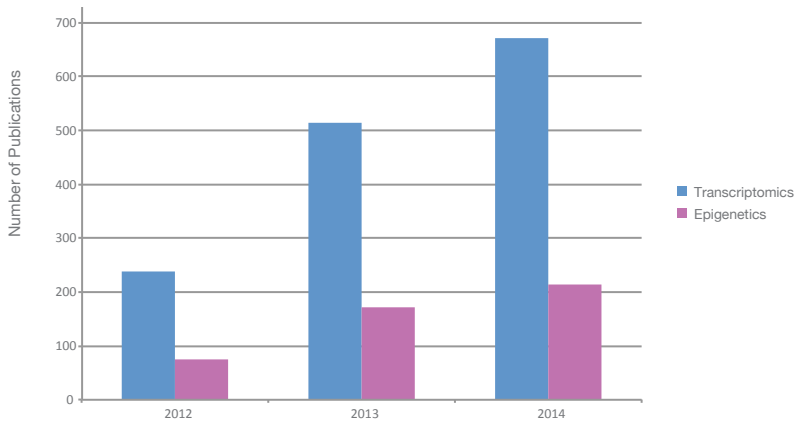

Figure 3: Complex Disease Research Growth. Existing research publications that use transcriptomics and epigenetics to investigate mechanisms in neurology, immunology, cardiology, and metabolic disease are represented cumulatively from 2012 to present day.

Table 1: Approaches to Characterizing the Functional Genomics of Complex Diseases[a]

Research Goal	Approach	Advantages	Illumina Solutions
Identify SNPs associated with disease	GWAS	• Cost-effective for large studies • Large amounts of data for statistical confidence	Omni microarrays
	• Whole-genome sequencing • Whole-exome sequencing	• Single-base resolution of the genome • Rare variant detection	• TruSeq DNA library prep kits • Nextera® DNA library prep kits • HiSeq® and HiSeq X Ten Systems
Identify CNVs associated with disease	Array comparative genomic hybridization	• Cost-effective for large studies • Large amounts of data for statistical confidence	• CytoSNP-850K BeadChip • Omni arrays
	Whole-genome sequencing	• Small CNV detection • Ideal for small studies, such as trios or pedigrees	TruSeq DNA library prep kits
Verify causative variants	Fine mapping using custom genotyping arrays	• Cost-effective for large studies • Increased SNP density in regions of interest	• iSelect custom genotyping BeadChips • Consortia
	eQTL analysis using RNA-Seq	• Does not require prior knowledge • Detects isoforms • Quantifies expression	• TruSeq Stranded mRNA Library Prep Kit • TruSeq Targeted RNA Expression Kit
	Fine mapping using targeted resequencing	• Nucleotide-level resolution • Ability to sequence many target regions in a single reaction	TruSeq Custom Amplicon Assay
Understand cellular activity	Ribosome profiling	Indicates which proteins are actively translated in a cell	ARTseq Ribosome Profiling Kit
	Total RNA sequencing	• Does not require prior knowledge • Can identify long, noncoding RNA transcripts	TruSeq Total RNA Library Prep Kit
	Small RNA sequencing	• Detects small, noncoding RNA species • Does not require prior sequence or secondary structure information	TruSeq Small RNA Library Prep Kit
Characterize epigenetic patterns	ChIP-Seq	Enables identification of protein binding sites in chromatin	TruSeq ChIP Library Prep Kit
	Methylation analysis using microarrays	Cost-effective for epigenome-wide association studies	HumanMethylation450 BeadChip
	Methylation sequencing	Single-base resolution of CpG methylation across the genome	EpiGnome Methyl-Seq Kit

a. All listed products are for research use only.

Summary

Recent technological advances provide new tools for investigating the many factors that influence complex diseases, including gene mutations, RNA regulation, and epigenetic patterns. The advent of high-throughput genomic technologies is changing the way researchers assess complex diseases, enabling them to process high sample volumes in less time, publish faster, and identify potential therapeutic targets earlier. The discoveries made today using arrays and sequencing provide opportunities for better understanding of the molecular factors that govern disease biology, ultimately leading to improved quality of life for people affected by complex conditions.

References

1. Craig J (2008) Complex diseases: research and applications. Nature Scitable (www.nature.com/scitable/topicpage/Complex-Diseases-Research-and-Applications-748)
2. National Institutes of Health (2010) Genetics of common, complex disease. NIH Fact Sheets (report.nih.gov/NIHfactsheets/Pdfs/GeneticsofCommonComplexDisease(NHGRI).pdf)
3. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001;409:860-921.
4. International HapMap Consortium. The international HapMap project. *Nature.* 2003;426:789-796.
5. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 2011;9:e1001046.
6. Catalog of Published Genome-Wide Association Studies (www.genome.gov/gwastudies) Accessed 23 September 2014.
7. Kiezun A, Garimella K, Do R, et al. Exome sequencing and the genetic basis of complex traits. *Nat Genet.* 2012;44:623-630.
8. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci USA.* 2014;111:E455-464.
9. Fromer M, Pocklington AJ, Kavanagh DH, et al. *De novo* mutations in schizophrenia implicate synaptic networks. *Nature.* 2014;506:179-184.
10. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature.* 2014;506:185-190.
11. Holmen OL, Zhang H, Fan Y, et al. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat Genet.* 2014;46:345-351.
12. Stefansson H, Meyer-Lindenberg A, Steinberg S, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature.* 2014;505:361-366.
13. Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature.* 2010;464:713-720.
14. Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet.* 2011;43:1193-1201.
15. Illumina (2013) Personalized cancer therapies through precision diagnostics. (www.illumina.com/documents/products/other/interview-daber-robert.pdf)
16. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA.* 2009;106:9362-9367.
17. Nagai K, Kohno K, Chiba M, et al. Differential expression profiles of sense and antisense transcripts between HCV-associated hepatocellular carcinoma and corresponding non-cancerous liver tissue. Int J Oncol. 2012;40:1813-1820.
18. Lunnon K, Smith R, Hannon E, et al. Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer's disease. *Nat Neurosci.* 2014;17:1164-1170.

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

 Pub. No. 1070-2014-007 Current as of 20 October 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTA
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACG
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA