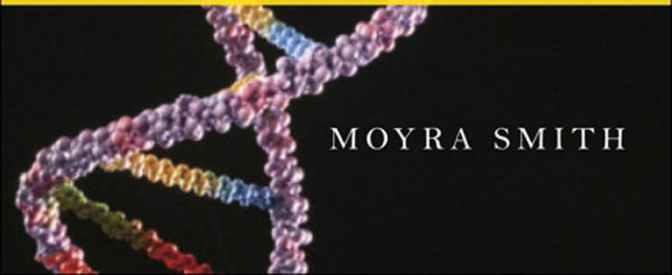
INVESTIGATING THE HUMAN GENOME

INSIGHTS INTO HUMAN VARIATION AND DISEASE SUSCEPTIBILITY



MOYRA SMITH

Investigating the Human Genome

Insights into Human Variation and Disease Susceptibility

Moyra Smith

Vice President, Publisher: Tim Moore
Associate Publisher and Director of Marketing: Amy Neidlinger
Acquisitions Editor: Kirk Jensen
Editorial Assistant: Pamela Boland
Operations Manager: Gina Kanouse
Senior Marketing Manager: Julie Phifer
Publicity Manager: Laura Czaja
Assistant Marketing Manager: Megan Colvin
Cover Designer: Chuti Prasertsith
Managing Editor: Kristy Hart
Senior Project Editor: Lori Lyons
Copy Editor: Hansing Editorial Services
Proofreader: Kathy Ruiz
Senior Indexer: Cheryl Lenser
Compositor: Nonie Ratcliff
Manufacturing Buyer: Dan Uhrig

Publishing as FT Press
Upper Saddle River, New Jersey 07458

FT Press offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales. For more information, please contact U.S. Corporate and Government Sales, 1-800-382-3419, corpsales@pearsontechgroup.com. For sales outside the U.S., please contact International Sales at international@pearson.com.

Printed in the United States of America

First Printing June 2011

ISBN-10: 0-13216814-6
ISBN-13: 978-0-13216814-4

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.

Library of Congress Cataloging-in-Publication Data
Smith, Moyra.
Investigating the human genome : insights into human variation and disease susceptibility / Moyra Smith.
p. ; cm.

Includes bibliographical references.
ISBN-13: 978-0-13-216814-4 (hardback : alk. paper)
ISBN-10: 0-13-216814-6 (hardback : alk. paper)
1. Human genome. 2. Disease susceptibility. I. Title.
[DNLM: 1. Genome, Human. 2. Genetic Predisposition to Disease. QU 470]
QH447.S625 2011
611'.01816—dc22

2011006494

This work is dedicated to the memory of three remarkable people who inspired and encouraged me long ago:

My mother Florence Van Eyk Smith,
my grandfather Manard James Van Eyk,
and our beloved family physician,
Dr. Colin Roy Cockcroft.

This page intentionally left blank

Contents

Acknowledgments

I express my gratitude to researchers and clinicians who have worked hard and long to further our knowledge of genetics and genetic diseases. Preparation of this book would not have been possible without the amazing resources of the University of California library services and the online resources of the National Library of Medicine—I want to acknowledge all who contribute to making these resources available.

I want to acknowledge Kirk Jensen, who encouraged me to write this book. I thank Lori Lyons of Pearson for moving this manuscript along in the publishing process and copy editor Krista Hansing for her thorough input.

Preface

In 2009, Georgina Ferry wrote, "The optimism of science is twofold: that its methods might reveal, one tiny pixel at a time, more of the wonder of the natural world and that this knowledge might be able to solve practical human problems."

Progress in the fields of genetics and genomics since a draft sequence of the human genome was published in 2001 is indeed a cause for optimism. However, this discovery has left some people disappointed because development of new therapies to treat disease has been slower than anticipated. Availability of this sequence information has fueled groundbreaking studies in genetics, genomics, and epigenetics that provide insight into human variation and the pathogenesis of both common and rare diseases. The goal of this book is to briefly review several of those groundbreaking studies and new insights.

My own experiences during a 40-year career as a clinical geneticist and researcher in genetics and genomics influenced the choice of topics discussed in this book. I explore new insights into human origins, migrations, and human population diversity gained though genomic analyses. I consider insights into the etiology of common diseases such as diabetes and coronary heart disease. I also consider studies on synapses and synaptic plasticity, representing pathways to understanding mind and cognition.

I discuss complexities of late-onset neurological diseases and efforts to utilize genetic and genomic methodologies to unravel the pathogenesis of these disorders. I also consider new insights into aspects of protein misfolding and clearance or deposition as aggregates that sequester other proteins. In considering gene environment interactions, I focus on aspects of DNA damage and repair and DNA

instability. An appropriate movement is underway toward translational research and greater emphasize on treatment. I review examples of treating primary defects and downstream effects of genetic disorders. I review new information on regulating gene expression at the levels of transcription, translation, and post-translational modifications. Growing evidence indicates that modifications of DNA, histones, and of nonhistone proteins greatly impact gene expression and the function of gene products, and I review aspects of research in these areas, sometimes referred to as epigenetics. In a chapter related to cancer, I review new discoveries in genetics and genomics that have direct relevance to therapy.

Growing evidence points to the importance of protein interactions and webs of molecular interactions that determine regulation and growth and the operation of systems, and I consider these topics. In a closing chapter I consider the relevance of genomics and systems biology to personalized medicine.

1

Genome architecture and sequence variation in health and disease

Availability of information on DNA sequence in human genomes and advances in technologies to amplify and sequence DNA have led to significant progress in delineating sequence differences that lead to disease. These techniques have also led to the discovery of sequence variants that occur in healthy individuals.

Studies of variation in the human genome are greatly facilitated through the availability of microarrays designed to detect single nucleotide polymorphisms (SNPs) that occur with frequencies greater than 1% to 5% in the population. Gene loci that are close to each other are often coinherited. SNP analyses can determine a series of alleles of loci in a specific region (a haplotype). Microarray technologies enable analysis of as many as one million SNPs on each array. These microarrays can also determine structural variation and copy number changes, defined as deletion or duplications greater than 1 kilobase (kb). Specific probes for regions known to frequently harbor copy number changes are also present on SNP microarrays such as the Affymetrix 6.0 array. Advances in technologies in DNA sequencing include massively parallel sequencing, often referred to as next-generation sequencing.

This chapter explores aspects of structural genomic variation and sequence variation in different populations and the role of sequence differences in the etiology of common disorders such as diabetes

mellitus, obesity, and coronary heart disease. It also covers next-generation sequencing and examples of its application to the discovery of gene defects that lead to disease.

Through the use of polymerase chain reaction techniques, samples with low concentrations of DNA can be used to derive material for DNA sequencing. This chapter discusses applications of these techniques to discover how the sequence in modern humans differs from that of Neanderthals and early modern humans. Also presented are reports of studies of DNA extracted from two teeth from a man who died in 1783. DNA analysis enabled researchers to diagnose the disease that afflicted him and analyze the specific mutation and surrounding polymorphisms that connected him to present-day patients with the same disease.

Structural variation

In the human genome, segmental duplications with highly identical sequence are usually interspersed and separated by more than 1 megabase. She, et al. (2006), identified more than 400 duplication blocks within the human genome. Segmental duplications are frequently clustered in pericentric and subtelomeric regions (Marques-Bonet, et al., 2009). Evidence indicates that pericentric and subtelomeric duplications evolved independently from intrachromosomal duplications. Core duplicons of 5–30kb occur in intrachromosomal duplications. One example of a core duplicon is LCR16a, which is rich in Alu repeats.

Unequal crossover between directly oriented duplicated segments may lead to dosage changes or altered structure and function of a gene. Marques-Bonet, et al., noted that most copy number polymorphisms result from this mechanism.

Regions between segmental duplications may be deleted, duplicated, or inverted as a result of unequal crossover. The existence of

highly similar duplicated segments on two different chromosomes may lead to translocation events. Polymorphisms also exist within the segmental repeats, and in different individuals, these regions may be larger or smaller. Segmental duplications are particularly abundant in certain chromosome regions, such as 15q11-q13, and these regions are frequent sites of deletions and duplications.

A key question is whether a specific structural variant, such as a deletion or a duplication (copy number variant) that includes unique sequence DNA, is a direct cause of phenotypic abnormality. Genomic syndromes often occur as a result of deletion or duplication of genomic regions that are flanked by segmental duplication blocks. In these syndromes, specific phenotypes result from the deletion of specific regions; for example, Williams syndrome results from the deletion of chromosome 7q11.2. Characteristic phenotypic features of this syndrome include cognitive and behavioral impairments, distinct facial features, and cardiac malformations.

Girirajan and Eichler (2010), reviewed findings in a subset of genomic structural changes in which a particular genomic change results in a series of phenotypes in which specific clinical features differ in different individuals. Differences occur in the degree to which individuals with the same defect are affected—that is, there are varying degrees of penetrance. The clinical consequences of a particular dosage change in a specific region may be influenced by dosage changes or mutations elsewhere in the genome.

Examples of specific regions where deletions are associated with a variety of phenotypes include 16p11.2. In some cases with deletion in this region, severe obesity occurs; other cases with the same deletion are diagnosed with autism, while in others, congenital malformations and developmental delay occur. Diverse phenotypes have been described in cases with deletion of 17q12; some cases present with hereditary neuropathy, with a tendency to pressure palsy (HNPP);

and in other cases, schizophrenia occurs. Other diagnoses encountered in patients with 17q12 deletion include renal cystic disease or maturity-onset diabetes of the young. Deletion in 1q21.1 may be associated with a learning disability in some cases and with congenital heart disease or schizophrenia in others.

The copy number variants associated with diverse phenotypes are sometimes found with low frequency in control populations. One question that arises is whether the different phenotypic consequences result from slight differences in the position of deletion breakpoints and whether sequence differences occur in the same region on the homologous chromosome.

Another genetic factor that may play a role in some cases is that the deletion of a specific locus on one chromosome unmasks a recessive mutant allele at that locus on the homologous chromosome. Other important possible explanations for the phenotypic variation are that additional genetic modifiers elsewhere in the genome modify the phenotype.

Girirajan and Eichler (2010), proposed that a two-hit genomic model most likely explains the variable phenotypes in individuals with copy number variants in 16p12.1 or 22q11.2.

Copy number variations and deletions, in particular, are most often considered to be of clinical relevance if they arise *de novo*—that is, if they are present in a child but absent in the parents. However, growing evidence indicates that parents who carry specific copy number variants may have subclinical manifestations attributable to the genomic change. CNVs and microarray are illustrated in Figure 1.1.

Human genetic sequence variation

During the past decade there have been significant advances in technologies for DNA sequencing that have facilitated studies of variation in ancient and modern humans.

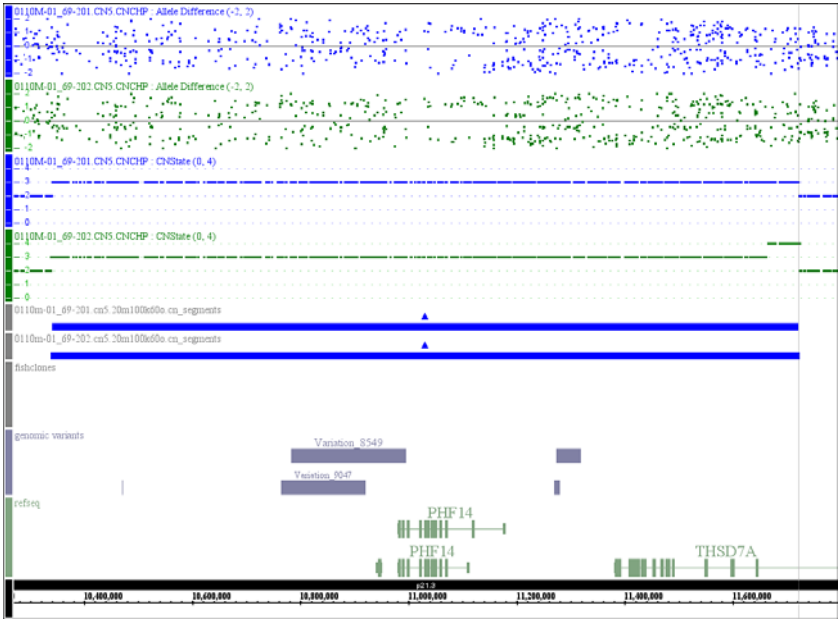


Figure 1.1 Results of analyzing SNP alleles and copy number variants using an Affymetrix 6.0 array and genotyping console on twin females with autism. Rows 1 and 2 at the top of the figure show the distribution of A and B alleles of specific SNPs. Note the identical patterns of alleles in the twins. Rows 3, 4, 5, and 6 show a chromosomal region with a copy number variant. Each twin has three copies of the CNV that encompasses three genes, shown at the bottom of the figure. A known population variant region is indicated, but the variant region is shorter and does not encompass a gene.

Studies in ancient fossil remains

In 2010, Green, et al., published data on four billion nucleotides of DNA sequence from three different Neanderthal individuals. They noted that DNA extracted from these late Pleistocene remains had degraded to segments less than 200 nucleotides in length and that it had been chemically modified. In addition, they found substantial contamination from microorganisms. To enrich the Neanderthal DNA, samples were digested with restriction endonucleases that selectively cleave microbial DNA.

Green, et al., examined the DNA sequence in loci with specific alleles that are known to differ in different modern human populations. They determined that Neanderthals shared 1% to 4% of genotypes at the sequences with Europeans and Asians. At these loci, Neanderthals did not share alleles with Sub-Saharan African populations.

Sequence analyses also led to the identification of genes that apparently underwent positive selection in modern humans. Specific sequences in these genes impact protein function. Green, et al., identified specific functional sequence changes that occurred in modern humans but were absent from Neanderthals, and the Neanderthal sequence matched the sequence present in chimpanzees.

Studies on an ancient Saqqaq individual

In the past decade, we have seen the confluence of paleontological analyses of bone fossils and cultural artifacts with DNA analyses. Rasmussen, et al. (2010), examined DNA recovered from the hair roots of an individual from Greenland, estimated to have lived 4,000 years ago, who was of member of the Saqqaq culture. Analysis of DNA polymorphisms from the hair roots indicated that the closest match was with individuals from eastern Siberia.

One advantage of analyses from hair roots is that they are less contaminated with fungi and bacteria than samples isolated from bone fossils. The high quality of the DNA isolated from hair roots of the Saqqaq individual enabled the analysis of 350,000 SNPs. Earlier studies by Rasmussen's group generated information on the complete mitochondrial DNA gene from permafrost-preserved Saqqaq individuals.

Given the length of the tracts of homozygosity they found, Rasmussen, et al., concluded that the inbreeding coefficient was high. Rasmussen studied DNA sequence at functional polymorphic sites. The combination of SNPs at the HERC2 and OCA2 (oculocutaneous

albinism gene2) indicated that the individual most likely had brown eyes and dark hair. Analyses also revealed that the Saqqaq individual was most closely related to three Northern Old World Arctic populations and was more distantly related to New World Amerinds. Researchers were not able to detect evidence of West Eurasian population admixture. Nuclear DNA SNP analyses and studies on the mitochondrial and Y chromosome haplotypes of the Saqqaq individuals matched most closely with those of North East Asian populations.

The Saqqaq culture is a component of the Arctic small tool transition and is estimated to have existed between 4,750 and 2,500 years ago.

Sequence variation in different populations and regions

In an analysis of 650,000 common SNPs, Li, et al. (2008), collected samples from populations in 51 geographic regions. Populations studied were drawn from Sub-Saharan Africa, North Africa, the Middle East, Europe, East, South, and Central Asia, Oceania, and the Americas. They carried out haplotype analysis to identify linked alleles at specific loci. They detected finer haplotype substructure in different regions. They noted, for example, that Palestinians, Druze, and Bedouins have haplotype contributions from the Middle East, Europe, and South and Central Asia.

Li, et al., concluded that nonrandom differences between populations have accumulated at a number of different loci. However, they also concluded that within population differences accounted for most of the genetic diversity. Results of their analyses revealed that heterozygosity was greatest in Africa and was reduced as geographic distance from Addis Ababa increased.

Tishkoff, et al. (2009), studied genotypes in 121 African populations, in 60 non-African populations, and in the African-American population. They studied microsatellite repeat polymorphisms and insertion deletion polymorphisms. They obtained evidence for

regional differences in the allele frequency of markers; however, their analyses also revealed evidence for substantial population admixture.

Homozygosity mapping

Because recombination occurs between homologous chromosomes during meiosis, the presence of identical alleles over long stretches of genomic DNA on homologous chromosomes (homozygosity) was thought to occur only in consanguineous pedigrees or inbred populations. Gibson, et al. (2006), studied 262 individuals in four different populations and identified 20 different genomic regions where homozygosity extended over 1 megabase or more. They noted that the lowest number of homozygous tracts occurs in the Yoruba population and that this reflects the more ancient roots of the population; over longer time periods, segments of chromosomes have broken up. Gibson, et al., noted that, in modern populations, tracts of homozygosity often occurred in similar genomic regions, indicating regions with a lower frequency of recombination. Examples of blocks of homozygosity are illustrated in Figure 1.2.

Studies on hereditary disorders and population history

Currently, 36 disorders are considered to comprise the Finnish disease heritage. Norio (2003) reviewed the history and studies of these disorders and related them to the historical origins, migrations, and settlements of the Finnish population. Thirty-two of these disorders have autosomal recessive inheritance patterns, two are autosomal dominant, and one is an X-linked disorder.

Norio noted that the Finnish population has been relatively stable for many years, without evidence of continuous migration into Finland. Internal migration of families from the southeastern parts of the country around Sevo into middle and northern regions of the country occurred around 1600. The migrant families settled in small clusters. Each cluster was often located at some distance from other

clusters, and with little admixture between clusters, mating occurred within clusters. In later generations, couples who married often shared founders six or seven generations ago.

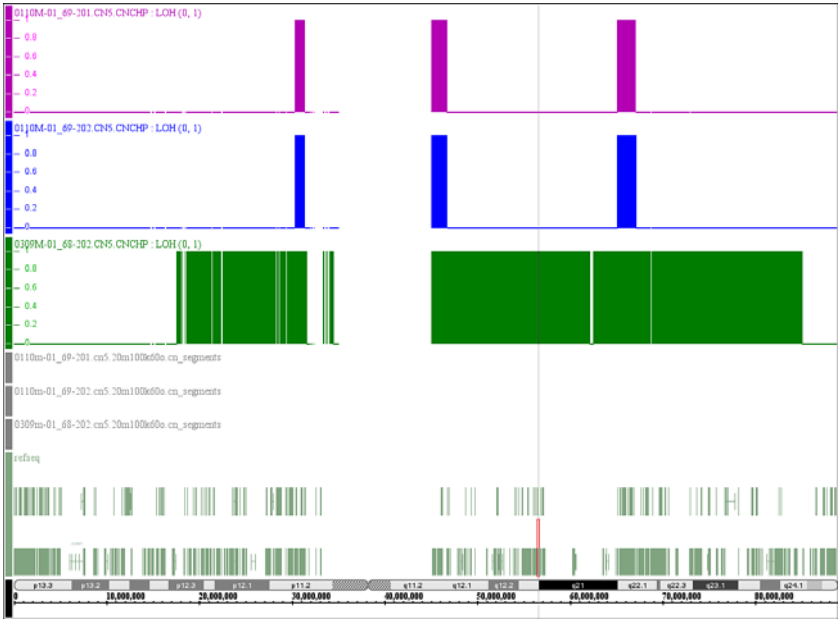


Figure 1.2 Blocks of homozygosity that are identical in twins (rows 1 and 2). Strikingly large blocks of homozygosity are present in the individual illustrated in row 3, likely due to consanguinity of parents. Rows 4 and 5 indicate positions of genes on chromosome 16.

In 1999, Peltonen, et al., reported that gene loci for 32 of the Finnish hereditary diseases were mapped to specific chromosomes and causative genes for 17 of the disorders were isolated. As expected, marked linkage disequilibrium occurred for markers in the vicinity of the disease alleles, and homogeneity of the disease-causing alleles was observed.

Peltonen and colleagues studied molecular mechanisms in these genetic disorders. Molecular analyses of products encoded by genes in regions where the diseases mapped resulted in the discovery of new proteins and enzymes. Peltonen emphasized that analysis of the

disease genes facilitated disease diagnosis, and, importantly, healthcare was available following diagnosis.

Peltonen and coworkers also reported that analyzing linkage disequilibrium is useful in identifying gene loci that contribute to the risk of complex common diseases. Kilpinen, et al. (2009), studied regions of linkage disequilibrium in a unique pedigree with multiple cases of autism. Individuals in this pedigree shared ancestors in the 17th century. Analyses revealed areas of linkage disequilibrium in three chromosomal regions at 15q11-q13, 19p13, and 1q23.

Genetic variations, single nucleotide polymorphisms (SNPs) and genome wide association studies (GWAS)

The design of genome wide association studies (GWAS) is predicated on the hypothesis that common DNA sequence variants contribute to the etiology of common disease. Results indicate that even when statistically significant associations between disease and a specific SNP are determined, the overall contribution of specific SNPs to disease risk is often low.

Genome wide association studies and insight into etiology of type 2 diabetes and obesity

In type 2 diabetes, the pancreatic beta cell–secreting capacity becomes inadequate to overcome the progressive peripheral resistance to insulin uptake. Factors that play roles in the development of peripheral insulin resistance include age, inactivity, and weight gain. McCarthy (2010) reviewed the discovery of genes that impact susceptibility to diabetes and obesity. He considered three waves of discovery. The first included family-based linkage studies. These studies led to the identification of genes involved in a number of Mendelian forms of early-onset diabetes, including neonatal diabetes and maturity-onset diabetes of the young (MODY). Genes that were found to

play a role in MODY included NEUROD1 (neurogenic differentiation 1); GCK (glucokinase); hepatic nuclear factor genes HNF1A, HNF1B, and HNF4A; and IPF (insulin promoter factor). Family studies also led to the discovery of a mitochondrial DNA mutation that predisposes carriers to diabetes and deafness.

McCarthy noted that family studies of childhood obesity led to the discovery of rare forms of this condition due to mutations in any one of three genes: leptin, leptin receptor, and pro-opiomelanocortin.

The second phase of investigation into diabetes and obesity involved searching for variants in candidate genes. These studies led to the identification of common variants of modest effect in PPARG (peroxisome proliferation activated receptor gamma) and KCNJ11 (potassium inwardly-rectifying channel, subfamily J, member 11). Resequencing of the melanocortin 4-receptor gene led to the identification of associated variants in 2% to 3% of cases of obesity.

A third wave of studies involved large-scale analysis of common DNA sequence variants (SNPs). McCarthy considered this to be the most successful wave of studies. Important diabetes-associated loci identified in these studies include the transcription factor that modulates pancreatic function TCF7L2; cyclin-dependent kinases CDKAL1, CDKN2A, and CDKN2B, which regulate cyclin; and HHEX, a gene involved in beta cell development. Each copy of a susceptibility allele at one of these loci leads to a 15% to 20% increase in the risk for diabetes.

McCarthy reported at least 40 known loci with alleles associated with increased risk of diabetes. Of interest is the fact that five of the loci with common variant alleles associated with increased risk of diabetes also harbor rare variants involved in familial or syndromic diabetes. These four loci are wolframin (NFS1); hepatocyte nuclear factors HNF1A and HNF1B; the melatonin receptor MTNR1B; and IRS1 insulin receptor substrate 1, which impacts insulin action.

Pathways involved in diabetes

McCarthy reported that the loci with the strongest evidence of association with type 2 diabetes impact insulin secretion. Examples include cyclin-dependent kinases CDKAL1, CDKNL2A, and CDKN2B. At these loci, risk variants predispose to reduced pancreatic beta cell mass. The diabetes risk alleles in TCF7L2, MTNR1B, and KCNJ11 predispose to beta cell dysfunction.

Risk alleles in the FTO locus (gene related to fat mass and obesity) contribute to obesity and to peripheral resistance to insulin. The PPARG and IRS1 (insulin receptor substrate 1) loci impact insulin resistance and obesity.

In the monogenic forms of diabetes and in autoimmune diabetes in adults, information about the underlying causative gene can influence therapeutic decisions, such as whether insulin is required, whether dietary management may be sufficient, or whether sulfonylureas are required. In type 1 diabetes commonly associated with HLA variants (latent autoimmune diabetes) or with defects in the insulin genes INS or PTPN22 (protein tyrosine phosphatase non receptor type 22), insulin is likely necessary. Maturity-onset diabetes of the young (MODY) due to GCK (glucokinase) deficiency may respond adequately to dietary management. MODY due to HNF1A deficiency may require treatment with sulfonylureas.

McCarthy reviewed the results of genome wide association studies designed to identify common variants associated with increased body mass index (BMI) and noted that at least 30 such loci have been associated. The strongest signal is associated with the FTO locus (gene related to fat mass obesity). He noted that signals of risk alleles were also detected in genes with neuronal function, such as BDNF (brain derived neurotrophic factor), SH2B1 (signaling protein), and NEGR1 (neuronal growth regulator). He indicated that obesity may be partly a disease of disordered hypothalamic function. In studies that involved analyzing fat mass distribution, risk alleles in 15 loci

were identified. Evidence indicates that risk alleles at these loci impact adipocyte development and function.

McCarthy noted that clinical translation of these findings is impacted at least partly by the modest effect of the risk alleles. Homozygotes for the FTO risk allele are an average of 2 to 3 kilograms heavier than in individuals without the risk allele. However, he noted that identifying risk-altering genes contributes to our understanding of the biology of disease. Another important consideration is that most of the risk alleles lie outside the coding regions of genes, and it's not clear how they impact the regulation of gene expression.

McCarthy predicted that large-scale genome-wide resequencing efforts now underway would clarify relationships between sequence variants and clinical phenotypes.

Tracking genes involved in coronary heart disease after GWAS

In 2007, Samani, et al., carried out genome wide association studies in coronary heart disease subjects. They identified several genetic loci that affect the risk of coronary artery disease (CAD), including loci at 9p21.3 and 1p13.5. In 2008, Kathiresan, et al., identified two loci associated with abnormal levels of low-density lipoprotein cholesterol (LDL cholesterol), one locus mapped to chromosome 1p13 and the other mapped to 19p13. They noted that the 1p13 locus maps near the gene SORT1 (sortilin 1).

Kjolby, et al. (2010), demonstrated that sortilin protein encoded by SORT1 is an intracellular sorting receptor for apolipoprotein ApoB100. They noted that SORT1 regulates plasma low-density lipoprotein levels through hepatic export of ApoB100 containing lipoproteins. In studies on mice, they determined that sortilin 1 overexpression stimulates the hepatic release of lipoproteins and increases plasma LDL levels.

Musunuru, et al. (2010), carried out studies in cohorts of human subjects and in human-derived hepatocytes. They determined that a noncoding polymorphism SNP rs12740374 in 1p13 impacts a transcription factor binding site that alters hepatic expression of SORT1. The risk allele G in rs12740374 disrupts the C/EBP transcription factor binding site and is significantly associated with LDL cholesterol levels, $p=1X10^{-170}$. In studies on mouse livers, Musunuru, et al., determined that sortilin 1 impacts plasma levels of low-density lipoprotein (LDL) and very low-density lipoprotein (VLDL) cholesterol. They demonstrated that knockdown of sortilin 1 expression in mice led to a 46% increase in total cholesterol compared with controls.

The studies of Musunuru, et al., demonstrated the clinical relevance of non-protein-coding DNA variants identified through GWAS. They concluded that the sortilin pathway is a promising new target for therapeutic intervention in hyperlipidemia and myocardial infarction. These investigators noted that, in some individuals, aggressive treatment with statins fails to lower the levels of LDL cholesterol. Statins inhibit cholesterol synthesis through inhibiting hydroxy-3methyl-glutaryl coenzyme A reductase and reduce levels of both LDL cholesterol and total cholesterol.

In GWAS analysis of blood lipids on 100,000 individuals, Linsel-Nitschke, et al. (2010), reported evidence for the involvement of 18 genes that were previously shown to play roles in Mendelian lipid disease. The significance values for association were much higher with these variants than those obtained for other variants. Highly associated loci included LPL (lipoprotein lipase) $2X10^{-115}$, APOA1 (Apolipoprotein A1) $7X10^{-240}$, CETP (cholesterol ester transfer protein) $7X10^{-380}$, LDLR (LDL receptor) $4X10^{-117}$, APOE (apolipoproteins A) $9X10^{-147}$, APOB (Apolipoprotein B) $4X10^{-114}$, SORTL1 (sortilin1) $1X10^{-170}$, and GCKR (glucokinase regulator) $6X10^{-133}$.

Therefore, evidence indicates that genes that play roles in the etiology of rare Mendelian forms of diseases such as diabetes and hyperlipidemia also play roles in the common polygenic forms of these diseases.

Disease-specific mutation in a Hunterian museum skeleton and his living relatives

In 2011, Chahal, et al., reported that they had identified a specific mutation in the arylhydrocarbon receptor interacting protein (AIP) in four families from Northern Ireland in whom familial isolated pituitary adenoma occurred. The specific mutation in these families was a nucleotide substitution c.(910 CtoT); p.(R304X). A termination codon replaces amino acid 304, leading to a loss of 26 amino acids from the AIP protein.

Chahal, et al., obtained permission from the directors of the Hunterian museum in London to extract DNA from two teeth of the skeleton of an Irish giant who died in 1783. Harvey Cushing examined this skeleton in 1909. He concluded on the basis of the degree of enlargement of the pituitary fossa that the man had a pituitary adenoma. Chahal, et al., discovered that the same AIP mutation was present in the Hunterian giant with pituitary adenoma and in the four families from Northern Ireland they studied. Analysis of DNA polymorphisms (microsatellite repeat polymorphisms) revealed that the giant skeleton DNA and adenoma patients in the four Irish families shared a haplotype that extended for 2,068 megabases on chromosome 11q13.2 and included the AIP gene. Taking into account polymorphisms, mutation rates, and generation length, Chahal, et al., concluded that the skeleton and the four families shared a common ancestor 57 to 66 generations ago.

Discovery of familial-inherited adenomas in different populations and the role of AIP

In 2006, Vierimaa, et al., reported two clusters of families from Northern Finland who had familial pituitary adenoma that led to increased secretion of growth hormone and prolactin.

Analysis of SNP polymorphisms in these families revealed a link between adenoma development and chromosome 11q12-q13.

Sequencing of genes in this region revealed a defect in the aryl hydrocarbon interacting protein AIP. Subsequent analyses in the Finnish population led to the identification of a Q14X mutation in 6 out of 45 patients with acromegaly.

Karhu and Aaltonen (2007) noted that the function of AIP was not known.

The amino-terminal region of AIP contains FKBP domains. These domains usually are involved in protein folding and trafficking. In the carboxyterminal region of AIP are three tetratricopeptide repeats. These repeats usually form scaffolds for the formation of multiprotein complexes. The carboxy-terminal region of AIP interacts with arylhydrocarbon receptor and with the HSP 90 heat shock protein. Low expression of AIP in pituitary adenomas is a marker for invasive growth hormone producing tumors.

In 2009, Jennings, et al., reported studies on Polynesian kindred with three members who presented with pituitary macro-adenoma in childhood or adolescence. These patients had AIP mutation R271W. They presented with headaches, visual disturbances, and excessive height. Features of acromegaly were absent. Acromegaly features include frontal bossing and overgrowth of hands and feet.

In 2010, Daly, et al., reported results of an international study on 96 patients with germline AIP mutations and pituitary adenomas. They noted that the patients were usually young and that the first symptoms occurred in children or adolescents. Males constituted 63.6% of the patients. The majority of the tumors were macro-adenomas. Excessive secretion of growth hormone occurred in 78% of tumors. In 13 of the 96 cases, prolactin secretion was excessive; 7 tumors were nonsecreting.

In 2010, Chahal, et al., reported that they had identified 49 different AIP mutations in patients with familial-inherited pituitary adenomas. These included deletions, insertion, segmental duplications, nonsense, missense, and splice site mutations. In addition, whole

exon deletion or deletion of the entire AIP gene occurred in some patients. They noted that in the cohort of families they studied, approximately 30% of the individuals who carried a germline AIP mutation presented with pituitary tumors.

Chahal, et al. (2010), concluded that the physiological role of the arylhydrocarbon receptor (ARH) likely includes cell proliferation and differentiation. ARH occurs in the cytoplasm as a multiprotein complex with AIP, HSP90, and co-chaperone p23. This complex binds to xenobiotics. It is transferred to the nucleus, where its binds with hypoxia inducible factor HIF1b, also known as ARNT. They noted that several proteins involved in the regulation of hypoxia-induced proteins play a role in tumor susceptibility. These include succinate dehydrogenase fumarate hydratase and Von Hippel Lindau proteins. (These proteins are discussed further in Chapter 5, "Pathways, Phenotypes, and Phenocopies.")

Evidence also indicates that AHR binds to ubiquitin ligase and plays a role in the degradation of estrogen and androgen receptors.

Next-generation sequencing

Key elements in next-generation sequencing are the miniaturization of sequencing reactions, the sequencing of short fragments of DNA bound to solid matrices, and real-time photo-capture of the sequencing reactions. Different companies have developed a number of different platforms for next-generation sequencing; precise methods for capturing fragments and sequencing vary depending on the sequence platform used. In some cases, fragments are ligated with specific oligonucleotides at each end, and these are hybridized to matching oligonucleotides on the solid matrix sequencing platform. In other cases, DNA fragments are biotin labeled and then captured with streptavidin beads; the beads are subsequently captured on the sequencing platform. Detecting the sequencing reaction is enabled through use of nucleotides A G C T, each labeled with different colors

of fluorescent dyes, and fluorescent images are captured. The flow cells used as sequencing platforms are partitioned into several channels so that a number of samples can be simultaneously analyzed.

Next-generation sequencing is referred to as massively parallel sequencing because thousands of short sequences are read at the same time and each sequence is read optimally about 30 times. Sequence data generated on the platform is submitted to a computer and is subsequently aligned to reference sequence.

Whole-genome sequencing in humans generates a very large amount of data for analysis. In determining disease-causing mutations in humans, capturing specific genomic regions and capturing the human exome represent methods that reduce the complexity of the data analysis.

Data analysis may be further simplified by prioritizing genomic regions or genes to be studied through filtering at the levels of bioinformatic analysis. Roach, et al. (2010), carried out studies on two siblings affected with an autosomal recessive disorder called Miller syndrome and their parents. They applied previous information on polymorphic markers and haplotype analysis to select areas of the genome for computational analysis following whole-genome sequencing. They focused their analysis on 22% of the genome where both affected offspring inherited the same genomic segments from both parents.

In reviewing the application of next-generation sequencing to the discovery of rare gene defects that cause Mendelian diseases, Ng, et al. (2010b), emphasized that linkage information may narrow the genomic region that needs to be sequenced or computationally analyzed.

Additional studies are usually required to definitively establish the significance of sequence alterations that are likely candidates for disease causation. Significant changes include chain termination substitution, deletions, and nonsynonymous nucleotide substitution that cause amino acid substitutions that likely alter the structure or

localization of a gene product. Follow-up studies include applying PCR (polymerase chain reaction amplification) and Sanger sequencing. Downstream follow-up includes biochemical and physiological studies.

Whole-genome sequencing and the discovery of mutation leading to Charcot-Marie-Tooth Neuropathy

Charcot-Marie-Tooth neuropathies (CMT) are a group of disorders characterized by peripheral motor and sensory neuropathies with different modes of inheritance, including autosomal dominant, autosomal recessive, and X-linked inheritance. They are characterized clinically by symmetric distal polyneuropathy. Progressive muscle weakness and atrophy occur particularly in the peroneal muscles, leading to foot-drop and abnormal gait.

CMT results from mutations in at least 39 different genes. Lupski, et al. (2010), reported that mutation testing is available in the United States for 15 of the 39 genes and costs $15,000.

Lupski, et al., reported results of whole-genome sequencing and follow-up analysis on a family with CMT. Sequencing yielded 89 gigabytes of sequence data; the depth of coverage was 30, indicating that each base was sequenced 30 times. The sequence derived from the affected proband was compared to the human reference genome sequence, and differences between the two were documented. These differences included single base substitutions, small deletions, and insertions and copy number changes.

Copy number variants were examined by array comparative hybridization and by sequence analysis. No copy number variants were identified that impacted genes known to play roles in CMT.

Lupski, et al., focused attention on the 9,069 single nucleotide substitutions that led to nonsynonymous codon changes. Of these 121 were nonsense mutations. Data was examined to search for single nucleotide substitutions in the proband that impacted genes known

to cause neuropathic conditions. Two nucleotide substitutions were found in the SH3TC2 gene, one missense mutation that led to Y169H and one nonsense mutation R954X.

Lupski, et al., noted that mutations in the SH3TC2 gene were previously described in Eastern European, Turkish, and Spanish gypsy patients and that the R954X mutation was present in some of these patients.

In the family reported by Lupski, et al., the R954X mutation occurred in one parent of the proband and the Y169H mutations occurred in the other parent. The proband had three siblings affected with CMT, and all three carried both the R954X and Y169H mutations. Subclinical phenotypes revealed by neurophysiological studies occurred in heterozygotes for each of the mutations.

Earlier studies on SH3TC2 and Charcot-Marie-Tooth neuropathy

Demyelinating autosomal recessive CMT was mapped to chromosome 5q23-q33 through homozygosity mapping in consanguineous families. Subsequently, sequence analysis of genes in this region revealed mutations in the SH3TC2 gene (Azzedine, et al., 2006). In ten consanguineous families, eight different mutations were found. Six of the mutations occurred in exon 11. Two cases had R954X mutations. Azzedine, et al., noted that the patients had foot deformities and that spinal abnormalities (kyphoscoliosis) also occurred.

In an analysis of 23 English patients with autosomal recessive CMT, Houlden, et al. (2009), identified 5 patients with SH3TC2 mutations. Affected members in four families were homozygous for the R954X mutation, and in one family, the affected members were compound heterozygotes for the R954X mutation and E657K mutation. Houlden, et al., noted clinical heterogeneity in the families with

respect to the severity of neuropathy. Neuropathology on sural nerve biopsies revealed demyelinating fibers and an abnormal Schwann cell that formed onion bulb–like structures.

The SH3TC2 protein localizes to the cellular plasma membrane and to the membrane of vesicles in the endocytic membrane trafficking pathway (Lupo, et al., 2009).

Disruption in this pathway apparently leads to impaired interactions between Schwann cells and axons.

Roberts, et al. (2010), demonstrated interaction between SH3TC2 protein and the membrane small GTPase Rab 11. Rab11 is known to regulate the recycling of internalized membranes in the endosomal pathway.

Exome sequencing

Analysis is simplified when exome sequencing rather than whole-genome sequencing is carried out, because the exome constitutes approximately 1% of the genome, approximately 30 megabases (Mb). Ng, et al. (2010a), carried out exome sequencing on four unrelated individuals affected with Miller syndrome. Clinical features in Miller syndrome include micrognathia, cleft lip and palate, and eye and limb anomalies. To derive the sequence, Ng, et al., used array-based capture of exomes. Their study was initiated using DNA from affected siblings, which facilitated a search for changes in regions where siblings had identical polymorphisms and nucleotide substitutions. They identified a mutation in the dihydro-orotate dehydrogenase gene DHODH. They subsequently carried out studies on individuals affected by Miller syndrome in three unrelated families. Sequence analysis established that these affected individuals were compound heterozygotes for DHODH mutations. The DHODH gene product plays a role in pyrimidine metabolism.

Next-generation sequencing continues to shed light on DNA sequence changes and their potential roles in diseases. Bioinformatic analysis of sequence data is challenging, and continued development of resources for analysis is important. Equally important will be downstream analysis of the biochemical and physiological effects of sequence changes.

INDEX

Numbers

A

B

C

D

E

F

G

H

I

J–K

L

M

N

O–P

Q–R

S

T

U–V

W–Z