The Impact of Genomics on Pediatric Research and Medicine

abstract

In this review, we discuss some of the most recent developments in genomics research and their relevance to the field of pediatrics. In particular, we examine 3 major approaches that are being used to identify genetic correlates of disease: genome-wide association studies, copy number variation studies, and next-generation sequencing. In the past few years, these approaches have yielded major insights into the causes and pathophysiology of a wide range of diseases but are also constrained by certain limitations. This review provides an overview of the genomic landscape in complex pediatric disorders and sets the stage for translating new discoveries into clinical practice, the future of genomic medicine. *Pediatrics* 2012;129:1150–1160

AUTHORS: John J. Connolly, PhD,^a and Hakon Hakonarson, MD, PhD^{a,b}

^aCenter for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania; and ^bDepartment of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania

KEY WORDS

CNV, genome, genomics, GWAS, NGS

ABBREVIATIONS

ADHD—attention-deficit/hyperactivity disorder
AMD—age-related macular degeneration
ASD—autism spectrum disorder
CD—Crohn disease
CNVs—copy number variants
GWAS—genome-wide association studies
HDL-C—high-density lipoprotein cholesterol
IL—interleukin
NGS—next-generation sequencing
SNP—single nucleotide polymorphism
WES—whole exome sequencing

Both authors have made substantive intellectual contributions to this article. Respective responsibilities include the following: Dr Connolly: (1) substantial contributions to conception and design of the manuscript, (2) drafting the manuscript and revising in consultation with Dr Hakonarson, and (3) final approval of the submitted manuscript. Dr Hakonarson: (1) substantial contributions to conception and design of the article; (2) revising the manuscript, offering critical review for important intellectual content, and drafting key content sections; and (3) final approval of the submitted manuscript.

www.pediatrics.org/cgi/doi/10.1542/peds.2011-3636

doi:10.1542/peds.2011-3636

Accepted for publication Feb 27, 2012

Address correspondence to Hakon Hakonarson, MD, PhD, Director, Center for Applied Genomics, The Children's Hospital of Philadelphia, 3615 Civic Center Blvd, Abramson Research Center, Suite 1216E, Philadelphia, PA 19104. E-mail: hakonarson@email.chop.edu or John Connolly, PhD, Center for Applied Genomics, The Children's Hospital of Philadelphia, 3615 Civic Center Blvd, Abramson Research Center, Suite 1016H, Philadelphia, PA 19104. E-mail: connollyj1@chop.edu

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275).

Copyright © 2012 by the American Academy of Pediatrics

FINANCIAL DISCLOSURE: The authors have indicated they have no financial relationships relevant to this article to disclose.

FUNDING: Research was supported by the Daniel Kubert Estate and by a grant from the National Institutes of Health, Neurodevelopmental Genomics: Trajectories of Complex phenotypes (RC3 MH089924, Co-PIs Hakon Hakonarson/Raquel Gur). Funded by the National Institutes of Health.

In this review, we provide a narrative of recent developments in genomics, with the aim of generating a broad overview of the field as a whole. We focus on the 3 main approaches (genome-wide association studies (GWAS), copy number variation analysis, and next-generation sequencing [NGS]) that have been instrumental in propelling the field forward.

GENOME-WIDE ASSOCIATION STUDIES

GWAS use single nucleotide polymorphisms (SNPs) to identify and compare allele frequencies in target populations, typically by using either case-control or family-based designs. The former is the most common because it is conducive to large-scale recruitment and is not constrained by limits on the numbers of controls. The latter has advantages in terms of controlling for population stratification and the types of analysis that are possible. GWAS use microarrays to tag up to several million SNPs at once, which gives broad coverage of genic and nongenic regions. The approach is hypothesis-independent, and all SNPs carry equal weight during statistical analyses. When a significant difference in SNP frequency is observed between cases and controls, we infer a difference in the underlying genomic locus, which may affect gene expression or regulation. Because of the large number of comparisons being made, most GWAS require large numbers of patients and controls to achieve requisite statistical power, and sample sizes in excess of several thousand are the norm. A major advantage of the approach is its applicability to complex disease, where numerous loci may be implicated as causal factors.

GWAS are constrained by the fact that they can only examine SNPs found in relative abundance in the population of interest; generally SNPs with a population frequency of ~5% or greater. A major assumption, therefore, is that the

variants under investigation are common. By extension, a second assumption is that the phenotype of interest is caused by the cumulative result of many small-effect variants. These assumptions constitute the so-called common disease, common variant model,¹ which has successfully identified variants in many complex disorders, some of which are reviewed here. For diseases caused by rare variants, including many Mendelian syndromes, GWAS are underpowered, and sequencing approaches are required, either of linked/associated candidate regions or through whole exome/whole genome approaches.

Notable Studies

The first association study commonly considered to be truly genome-wide was published in March 2005; it was a screen of ~100 000 SNPs in 96 patients with age-related macular degeneration (AMD) and 50 control subjects.² By November 2011, 1061 GWAS with sample sizes of up to 250 000 had been cataloged by the National Human Genome Research Institute (<http://www.genome.gov/GWASstudies>), demonstrating rapid perpetuation in academia and industry. Arguably the phenotype with which GWAS have had most success is AMD. To date, 6 GWAS have examined common variants as a cause of this disorder, where heritability has been estimated at 71%.³ Prominent candidate loci include *ARMS2/HTRA1* and the complement pathway genes *CFH*, *CFB*, *C3*, and *C2*. *P* values as low as 4×10^{-322} for allele frequency differences between cases and controls have been reported.⁴ By using a standard liability threshold model,⁵ Yu et al (2011)³ calculated the explained variance of 12 replicated loci at ~39%, or 55% of heritability. Although a large proportion of the variance remains unexplained, this total is high compared with other phenotypes (this is further discussed in the Supplemental Information). A possible reason for the

comparative success of GWAS of AMD likely relates to evolutionary pressure. For a variant to remain common in a population, it must be resistant to negative selection. Because AMD is late-onset, it is more resilient to this phenomenon. GWAS of Alzheimer disease, which explain >23% of heritability,⁶ are consistent with this conclusion.

The success of the approach with late-onset phenotypes should not be taken to imply that GWAS are inappropriate for pediatric research, however. Where candidate loci do not account for a large proportion of explained variance, they nevertheless remain highly informative. For example, our laboratory previously identified 6 genetic markers on chromosome 5p14 that confirm susceptibility to autism spectrum disorders (ASDs).⁷ The region is flanked by two genes, *CDH9* and *CDH10*, both of which encode type II classic cadherins (transmembrane proteins that promote cell adhesion). The association of cadherins with neurodevelopment is consistent with the cortical-disconnectivity model of ASD,⁸ which proposes that the phenotype may result from increases or decreases in functional connectivity/synchronization of relevant pathways. This hypothesis is supported by recent functional studies that have correlated altered activity in neuronal networks with social, communication, cognitive, and sensorimotor impairment.⁹

There are numerous other examples in the GWAS literature in which candidate genes of relatively small effect have been highly informative in terms of defining the broader phenotype. A large meta-analysis ($n = \sim 100\,000$) of blood lipid levels by Teslovich et al (2010)¹⁰ identified 95 lipid-associated loci. Mouse models of 3 of these loci demonstrate the global importance of individual common variants. Thus, decreased expression of the novel candidate ortholog *Galnt2* significantly decreased levels of high-density lipoprotein cholesterol

(HDL-C), overexpression of the ortholog gene *Ppp1r3b* in the liver significantly lowered plasma HDL-C, whereas a knock-down of a *Ttc39b* transcript correlated with increased HDL-C. A related study examined blood lipid levels in myocardial infarction and found that an associated noncoding locus at chromosome 1p13 alters the expression of the *SORT1* gene.¹¹ Overexpression of *Sort1* in the liver was found to alter plasma low-density lipoprotein cholesterol and very low-density lipoprotein particle levels, which constitutes a novel regulatory pathogenic pathway. Collectively, these studies highlight the potential of individual candidate genes to defining pathogenesis, even in noncoding regions. These points are potentially important to bear in mind when evaluating GWAS of childhood-onset diseases (see below and Supplemental Information).

Attention-Deficit/Hyperactivity Disorder

Attention-deficit/hyperactivity disorder (ADHD) affects ~6% to 10% of Americans^{12,13} and is a major economic and health care burden.¹⁴ Although at least 11 GWAS have been conducted on children with ADHD, no standout candidate gene has emerged. A recent meta-analysis¹⁴ of 2064 trios, 896 cases, and 2455 controls found that no single SNP achieved a genome-wide significance below the 5×10^{-8} threshold, although a number of loci warrant additional investigation. These include a region on chromosome 7q21, where 8 SNPs featured among the top 50 associations. The closest gene (albeit 200 kb away) to this locus is *SHFM1*, which may be involved in proteolysis and the regulation of the cell cycle.¹⁵ Other genes of interest include *CHMP7* on chromosome 8p21, which may play a role in endosomal sorting and vesicular transport,¹⁶ and *CHMP2B*, associated with frontotemporal dementia, disinhibition and executive dysfunction.¹⁷ Recently, we performed the largest whole-genome study of ADHD in a cohort

of 3000+ cases and 12 000+ healthy children of European ancestry, genotyped with 550 000 SNP markers. Although no GWAS signal meeting standard criteria was uncovered, copy number variants (CNVs) affecting metabotropic glutamate receptor genes were significantly enriched in ADHD cases. These include deletions in *GRM5*, *GRM7*, *GRM8*, and duplications in *GRM1*, which are involved in glutamatergic neurotransmission, an important mediator for of neurodevelopment.¹⁸

Asthma

Asthma is a similarly complex disease that affects >6% of children in the developed world.¹⁹ The Cookson Laboratory conducted the first GWA study of asthma,²⁰ identifying multiple markers on chromosome 17q21 that have since been independently replicated.^{21–23} Two genes within this locus have been linked to susceptibility to asthma: *ORMDL3* and *GSDML*.²⁰ *ORMDL3* belongs to a family of genes that encode transmembrane proteins in the endoplasmic reticulum.²⁴ *GSDML* encodes a member of the gasdermin protein family, which are expressed in epithelial cells and regulate apoptosis. Another gene on chromosome 5q12, *PDE4D*, has subsequently been associated with asthma,²⁵ albeit outside the threshold of genome-wide significance. However, functional data linking (lung) expression of this phosphodiesterase to airway contractility marks it as a gene of interest. Similarly, *PDE11A*, which encodes a related phosphodiesterase, has also been linked with asthma.²⁶

We recently conducted a case-control study of pediatric asthma,²⁷ identifying a novel asthma locus on 1q31 containing *DENND1B*, which is expressed by natural killer (NK) cells and dendritic cells. Homolog proteins of *DENND1B* are thought to interact with the TNF α receptor.²⁸ The study also found that the same locus is associated with asthma

susceptibility in African American children and with inflammatory bowel disease and primary biliary cirrhosis.²⁹

Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) has been extremely amenable to GWAS, and multiple studies^{30–35} have identified and replicated numerous associations. The tally of currently identified IBD loci is 99, of which 71 are associated with Crohn disease (CD),³⁶ and 47 with ulcerative colitis.³⁷ A recent meta-analysis of CD³⁶ found that the 71 risk loci explained >23% of heritability, comparatively high in the realm of complex disease. Furthermore, we recently showed that GWAS may underestimate heritability at the *NOD2* locus³⁸ (see below). If a similar scenario exists for other IBD-related genes, explained variance may increase substantially.

Pediatric IBD

Pediatric IBD is associated with a higher incidence of family history, increased colonic involvement, and a more chronic clinical course.³³ Our group analyzed cohorts of children with early-onset IBD, identifying multiple novel loci, including 20q13 (*TNFRSF6B*)³² and 16p11 (*IL27*).³⁹ We also corroborated the majority of known adult-onset IBD loci, confirming a close genetic relationship between early- and adult-onset phenotypes. In individuals of European descent, the most prominent candidates include *NOD2*, *IL23R*, and *ATG16L1*. *NOD2* is located on chromosome 16q12 and is primarily expressed in peripheral blood leukocytes. It is a cytosolic innate receptor able to sense intracellular bacteria and to trigger RIP2- and NF- κ B-mediated pro-inflammatory responses.⁴⁰ *IL23R* is a type I cytokine receptor and, as with *NOD2*, has a long history of association with IBD and autoimmunity. Primary functions include maintaining T helper 17 cells, restraining *Foxp3*(+) regulatory T-cell activity, and initiating expression of T helper 17–like cytokines in non-T cells.⁴¹

ATG16L1 is located on chromosome 2q27 and is an important component of the autophagy pathway.⁴²

As with most of the genes discussed here, it is important to stress that these genes are not universally associated with the phenotype. Indeed, Lee and Parkes (2010)⁴³ point out that *NOD2*, *IL23R*, and *ATG16L1* are not associated with IBD in East Asians.

Obesity

Obesity is a major and growing health problem in Western societies, particularly in children.⁴⁴ It is a causal risk factor for type 2 diabetes, cardiovascular disease, hypertension, and numerous other chronic diseases.^{45,46} Approximately 70% of obese adolescents become obese adults,⁴⁷ and adolescent obesity is associated with increased adult mortality.⁴⁸ The first candidate gene identified by GWAS was insulin-induced gene 2 (*INSIG2*).⁴⁹ The *INSIG2* protein is an endoplasmic reticulum protein that inhibits processing of sterol regulatory element binding proteins. However, this candidate has been inconsistently replicated. An association with the fat mass and obesity-associated gene (*FTO*) has been more robustly reproduced,^{50–52} and we recently confirmed the association in European American and African American pediatric cohorts.⁵³ We have additionally uncovered several novel CNVs associated with obesity in both European American and African American children.⁵⁴

GWAS PATHWAY ANALYSIS

Pathway analysis, which leverages existing biological knowledge about gene function to examine how causal factors interact, helps assimilate the glut of GWAS data, and can also be used to identify risk variants. The approach typically examines whether test statistics for a predefined set of candidate loci simultaneously deviate from chance. It is based on pathway association as applied

to gene expression microarray analysis, in which examination of groups of related genes has yielded major insights into functional capacity.^{55,56} We recently used pathway analysis to determine risk factors for CD.³⁸ The study examined enrichment of association signals in genes previously identified in pathway networks as defined by Gene Ontology, Biocarta, and KEGG (with careful adjustment for gene size, number of SNPs per gene, and pathway size). The pathway most significantly enriched by association signals was the interleukin (IL)-12 gene pathway, which harbors IL-12 and IL-23 (Fig 1). These cytokines share a cellular receptor subunit, numerous intracellular signaling components, and have previously been shown to associate with CD.⁵⁷ Initially only 3 genes (*IL12B*, *IL23R*, and *IL12RB2*) at 2 loci (5q23, 1p31) showed genome-wide signals. Now, 3 additional genes in the IL-12–IL-23 pathway (*JAK2*, *CCR6*, and *STAT3*) were confirmed as candidates in replication studies, and 6 more were supported by pathway association, all of which have been previously reported as CD susceptibility genes in earlier association and functional studies.^{58–61} From a starting point where only 3 genes surpass the threshold for genome-wide significance, we develop a much richer picture of the pathophysiology of the disease through the pathway-based approach. This includes related variants that have remained above significance criteria but collectively contribute significantly to the risk variance. Moreover, given that the strongest candidate gene is often not a viable drug target, the pathway approach highlights alternatives for targeted intervention.

COPY NUMBER VARIANTS

As discussed earlier, GWAS have been most successful in identifying risk variants that are common. Recently, we have witnessed a renewed emphasis on rare variants, as a number of techniques have

been developed that complement the GWA approach. In particular, CNVs, which leverage the GWAS platform, have become a primary research area. CNVs are insertions, deletions, or inversions in the genome that vary in length from many megabases to a kilobase or less. Although CNVs are often not associated with any observable phenotype, their presence in genic regions has been linked to a number of major diseases, including ADHD,¹⁸ autism,⁶² schizophrenia,^{63,64} bipolar disorder⁶⁵ and many others. The origin of most CNVs is unknown, but causal mechanisms can include replication errors, meiotic recombination particularly in areas of segmental duplications, and homologous/nonhomologous repair of double-strand breaks.⁶⁶ As shown in Fig 2, CNVs can be detected by the same SNP arrays used for GWA by examining changes in the intensity of SNP signals.

Studies by the Wellcome Trust Consortium⁶⁷ and Conrad et al (2010)⁶⁸ report that common CNVs are well covered by SNPs in existing arrays, and many have been indirectly examined in a range of GWAS. A resequencing study by Pang et al (2010),⁶⁹ however, suggest that the impact of rare CNVs may be substantial. The authors reexamined data from the Venter genome and identified >12 000 structural variants spanning >40 mb of sequence that was initially unreported. These variants were found in 4867 genes, which are often large and under negative selection. In total, >24% of CNVs would not be imputed by SNP-association. Because rarer alleles are more likely to confer greater effect sizes and have a higher penetrance, these results strongly support the role of CNVs as causal factors in genetic diseases. Furthermore, this requires us to stress the limitations of using conventional SNP arrays to identify rare CNVs, which can be more accurately identified using high-density custom arrays or NGS methods.

Since the first CNV studies in 2004,^{70,71} the Database of Genomic Variants

FIGURE 1

Pathways linked to CD.¹⁰⁰ Wang et al (2010) manually compiled a pathway centered on IL-12 and IL-23 that is important to CD. Only the main proteins in this pathway are shown. For each gene, the most significant P value among SNPs closest to the gene was annotated. Reproduced with the permission of Nature Publishing Group.

(<http://projects.tcag.ca/variation>) has cataloged >66 000 known variants. The majority of CNVs are inherited and where they occur de novo are more likely to be pathogenic⁷² (although CNVs inherited from an unaffected parent have been associated with increased susceptibility⁷³). CNVs are the primary mode through which an individual acquires a mutation, and occur at a rate of $\sim 1.7 \times 10^{-6}$ per locus (as opposed to 1.8×10^{-8} for sequence variation⁷⁴), the cumulative result of which may constitute >10% of the human genome.

CNV Studies

CNV studies of ASD are perhaps the best known of pediatric disorders. An early family-based study⁷⁵ examined CNVs in 118 simplex families (ie, families with only 1 child with an ASD), 47 multiplex families, and 99 control families. De novo CNVs were identified in 10% of

patients with sporadic autism, compared with 3% in multiplex families, and 1% in controls. Disease-associated CNVs were identified at 17 loci on 11 chromosomes, supporting the widely held assumption that many different loci can contribute to ASDs, and that pathogenesises do not necessarily overlap. The sheer volume of loci identified by this approach affirms the extraordinary complexity of ASDs.

A number of subsequent studies have greatly expanded the number of ASD candidate loci. We reported >150 CNVs in 912 ASD families that were not found in 1488 controls.⁷⁶ Of these, 27 were replicated in an independent cohort of 859 ASD cases. Some of the rare variants had previously been associated with ASDs, including the cell adhesion molecule, *NRXN1*,⁷⁷ and the ubiquitin gene, *UBE3A*.⁷⁸ Another CNV study from our laboratory uncovered CNVs affecting 3 other prominent members of the

ubiquitin gene family, *PARK2*, *RFWD2*, and *FBXO40*, as well as CNVs in several other neuronal cell adhesion molecules, such as *CNTN4*, *NLGN1*, and *ASTN2*, all of which were observed to be significantly enriched in ASD cases. These findings indicate that dysregulated cell adhesion and ubiquitination networks may be integral to the etiology of ASDs. Together with neuronal cell adhesion molecules, the ubiquitin-proteasome system operates at pre- and postsynapses, and is involved in a number of critical synapse functions, including regulation of neurotransmitter release, recycling synaptic vesicles in presynaptic terminals, and modulating changes in dendritic spines and postsynaptic density.⁷⁹ Cell adhesion molecules additionally contribute to neurodevelopment by facilitating axon guidance, synapse formation, and plasticity, and neuron-glia interactions. Similar to individual GWAS candidates, these findings highlight the role of rare

FIGURE 2

Signal intensity patterns in 4 members of a simplex autism family.¹⁰¹ As shown in the Genome Browser (bottom), this CNV region encompasses *DGCR6* and *PRODH*. A horizontal bar represents the region for each individual (light/dark = 3/4 copies). We infer the first child inherits duplications from both parents. Reproduced with the permission of *Neuroscientist*.

variants in informing our understanding of the broader phenotype.

Schizophrenia

In a study of 150 European schizophrenia patients and 268 ancestry-matched controls, Walsh et al (2008)⁸⁰ identified novel CNVs of ≥ 100 kb in 15% of cases versus 5% of controls. Two subsequent studies in Chinese⁸¹ and European⁸² participants, however, failed to replicate this difference. A study from our group⁶³ compared CNVs in 977 European cases and 2000 healthy controls, and also used an independent replication cohort of 758 cases and 1458 controls. Genes

involved in synaptic transmission were found to be enriched in cases, with the calcium-signaling genes *CACNA1B* and *DOC2A* the most prominent, the latter was nominated as an ASD candidate.⁶² From the 1735 total patients analyzed, we identified an average of 45.4 CNVs per individual. For the combined 3485 controls, we identified an average of 45.1 CNV calls. Unlike the Walsh et al study, cases were not enriched for CNVs > 100 kb, nor for overall CNV frequency. The average size of CNVs for case-control groups was very similar at 88.4 kb and 87.9 kb respectively. This suggests that the location of CNVs, rather than

frequency, may be preeminent in determining association with schizophrenia.

The majority of CNV studies now converge on the conclusion that rare variants across multiple loci confirm susceptibility to disease and that de novo CNVs are globally overrepresented in cases. This conclusion has potentially important consequences for pediatric diseases as a whole because it highlights the prominent role of rare variants in complex phenotypes, which may be punctuated by a trivial subset of monogenic subtypes/syndromes or have less complex inheritance patterns. There is also evidence for a 2-hit CNV model in severe

neurodevelopmental cases, as evidenced by a study of microdeletions of the childhood developmental delay locus, 16p12.1. Girirajan⁸³ et al (2010) found that in 95% of cases, a 520-kb deletion was inherited from a parent. This is a much higher proportion than in most genomic disorders, of which de novo mutations are typical. Probands' families were subsequently found to be enriched for less severe neuropsychiatric disorders, including depression, bipolar disorder, mild learning disability, and seizures. The group hypothesized that differences in phenotype severity between probands and parents may be attributable to a second hit, and indeed another large CNV (>500 kb) was identified in 23.8% of probands. The authors subsequently reported second hits for a range of syndromes of which recurrent microdeletions/microduplications are a feature. These include Williams syndrome, DiGeorge syndrome, 15q13.3 deletion, 16p11.2 deletion, 1q21.1 deletion, and 22q11.2 duplication, and account for 5% to 15% of cases.

NEXT-GENERATION SEQUENCING

NGS differs from its predecessors in that it relies on the massive parallelization of biochemical and measurement steps, which produce thousands or millions of sequences simultaneously. This is accomplished by splitting the entire genome into small segments, which are then ligated to adapters during DNA synthesis. The first complete genome sequenced using NGS technology was published in 2008⁸⁴ and proved to be considerably more efficient in terms of cost and timescale than the Sanger approach used in the Human Genome Project.⁸⁵

Because NGS looks at the raw sequence, it essentially resolves criticisms of tagging coverage that have been aimed at SNP-based approaches. However, this increased resolution comes at the cost of a heavy burden on data-management and

analysis, and the infrastructures required to process the huge data output is immense. The current cost of sequencing an entire human genome is ~\$5000, the time required is several days (for a reasonable coverage of 30–50×), and the outlay required in terms of data-storage and analysis is considerable. For these reasons, it is not yet practical to sequence entire genomes on the large scale required for complex diseases. However, the resolution of monogenic disease can be accomplished by sequencing a family or even a single individual⁸⁶ and this represents the most prominent use of genome resequencing in the current climate.

To date, the NGS approach that has been most widely used to tackle complex disease is targeted resequencing, which leverages existing knowledge (eg, linkage or GWAS) to guide exploration of a candidate locus. Unlike whole-genome approaches, targeted resequencing is not hypothesis-neutral (agnostic) and typically concentrates on a finite number of loci that are decided in advance. Depending on gene size and coverage, the average cost to resequence a single gene in 1 individual is ~\$20. The efficacy of this approach is contingent on careful phenotyping and selection of participants, who are assumed a priori to be collectively enriched for variants at the target locus/loci. Specially developed software (eg, SampleSeq⁸⁷) can help enrich the yield of rare disease alleles in unrelated samples. Targeted approaches have successfully detected rare variants in a range of phenotypes, including nonsyndromic deafness,⁸⁸ cancer tumors,⁸⁹ and Rothmund-Thomson syndrome.⁹⁰ These conform to the “common disease rare variant” model,⁹¹ which is complementary to the GWA approach.

Whole Exome Sequencing Studies

A 2010 study by Ng et al⁹² was the first to successfully apply whole exome sequencing (WES) to a clinical disorder,

with the resolution of Miller syndrome, a Mendelian disorder characterized by a number of dysmorphologies including absent digits, and facial abnormalities. The group performed WES on 4 individuals (including 2 siblings) and 8 controls, identifying 6 rare variants in the *DHODH* gene that were subsequently confirmed by Sanger sequencing in 3 independent families. Since this landmark publication, the flow of NGS articles has accelerated rapidly, with new syndromes being resolved on an almost weekly basis (review at Ku et al 2011⁹³).

The Miller syndrome study is also notable in that it drew attention to a secondary phenotype in the sibling pair who had a medical history of recurrent infection. As is the case with many rare disorders, it can be difficult to determine whether these infections constituted an uncommon symptom of Miller syndrome or an unrelated disorder. In this case, the authors found that both siblings were compound heterozygote for *DNAH5*, which is known to cause primary ciliary dyskinesia and is characterized by recurrent infections of the respiratory tract⁹⁴ unrelated to Miller syndrome. Although essentially a coincidental finding, this nevertheless represents an important result in demonstrating the power of NGS as a diagnostic application.

Recent results from our laboratory yielded a similar conclusion. A WES⁹⁵ examined a severe form of ADHD in a father and 2 sons. Although the underlying cause of this primary phenotype remains to be resolved, the team identified 2 rare nonsynonymous mutations in the *PKLR* gene, which resolved a (coincidental) rare form of hemolytic anemia (Fig 3), raising the question of if and how to share genomic data with participants. In this instance, after consultation with the institutional review board, results were shared with the patient's hematologist, with an offer of genetic counseling. As sequencing becomes increasingly ubiquitous, coincidental findings are

FIGURE 3

Identification of rare variants by progressive filtering.⁹⁵ Combining genomic software and manual review identifies 2 variants that cause idiopathic hemolytic anemia. indels, insertions and deletions. Reproduced with the permission of Discovery Medicine.

increasingly likely to emerge and highlight the importance of appropriate institutional policies.

Whole Genome Sequencing Studies

To date, the majority of rare variants in monogenic disorders have been identified in exons. As such, whole genome sequencing studies, which are 4 to 10 times more expensive and produce ~50 times more sequence data, have not been commonplace. However, a recent study sequenced the whole genome of a patient with Charcot-Marie-Tooth disease and identified 3.4 million SNPs that were different to the reference genome.⁹⁶ By focusing on nonsynonymous SNPs in ~40 genes, the group was able to zero in on *SH3TC2* as harboring the causal variant. A similar approach identified *ABCG5* as the source of 2 nonsense

mutations as a cause of severe hypercholesterolemia.⁹⁷

The bias toward WES is largely based on expediency because it maximizes return on the investment of time and resources. It is applicable to sporadic cases (unlike linkage approaches) and, for Mendelian disorders, can succeed where the sample size is as low as 1.⁹⁸ One major drawback is that full capture of all exons is currently not possible because of incomplete knowledge of the genome. Similarly, as we know from GWAS, numerous intergenic regions have long been known to predispose to a wide range of genetic diseases. Although WES is currently the favored approach, it is clear that whole genome sequencing will eventually become the primary mode of discovery, and will have a large role to play in resolving complex disorders.

SUMMARY AND FUTURE DIRECTIONS

In the course of this review, we have attempted to summarize how advances in statistical models, technology, and software have transformed the field of genomics. Developing in parallel are a range of new approaches in the areas of epigenomics, proteomics, transcriptomics, and replication profiling, all of which offer unprecedented insight into how the genome functions and is regulated. Integrating these areas represents a major conceptual and logistical challenge but is necessary to fully understand the biology of complex disease. Essentially, GWA, CNV, and NGS studies provide a static view of the human genome, and an important next step will be to incorporate dynamic, non-nucleotide data. This is particularly true of pediatric disease, in which the effects of developmental stages are enormous.

Nevertheless, an increasing number of rare syndromes are being resolved and clinical tests (including chromosomal microarray tests for genomic deletions and duplications) are available to diagnose idiopathic forms of intellectual disability, neuropsychiatric subtypes, and congenital malformations.^{98,99} The next several years will inevitably yield additional insight into the elementary relationship between genotype and phenotype, resulting in a deeper understanding of pediatric development and diseases as the genome era emerges from its infancy. This will open up new avenues for targeted therapies in which dominant gene networks and pathways are targeted in subsets of patients who harbor variations that are most likely to be responsive to therapeutic approaches in functional pathways impacted by genetic variations.

REFERENCES

1. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273(5281):1516–1517
2. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*. 2005;308(5720):385–389
3. Yu Y, Bhangale TR, Fagerness J, et al. Common variants near *FRK/COL10A1* and *VEGFA* are associated with advanced age-

- related macular degeneration. *Hum Mol Genet.* 2011;20(18):3699–3709
4. Seddon JM, Cote J, Page WF, Aggen SH, Neale MC. The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences. *Arch Ophthalmol.* 2005;123(3):321–327
5. Barrett JC, Hansoul S, Nicolae DL, et al; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet.* 2008;40(8):955–962
6. So HC, Sham PC. Effect size measures in genetic association studies and age-conditional risk prediction. *Hum Hered.* 2010;70(3):205–218
7. Wang K, Zhang H, Ma D, et al. Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature.* 2009;459(7246):528–533
8. Gepner B, Féron F. Autism: a world changing too fast for a mis-wired brain? *Neurosci Biobehav Rev.* 2009;33(8):1227–1242
9. Müller RA, Shih P, Keehn B, Deyoe JR, Leyden KM, Shukla DK. Underconnected, but how? A survey of functional connectivity MRI studies in autism spectrum disorders. *Cereb Cortex.* 2011;21(10):2233–2243
10. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature.* 2010;466(7307):707–713
11. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature.* 2010;466(7307):714–719
12. Boyle CA, Boulet S, Schieve LA, et al. Trends in the prevalence of developmental disabilities in US children, 1997–2008. *Pediatrics.* 2011;127(6):1034–1042
13. Bloom B, Cohen RA, Freeman G. Summary health statistics for U.S. children: National Health Interview Survey, 2009. *Vital Health Stat 10.* 2010; (247):1–82
14. Minkoff NB. ADHD in managed care: an assessment of the burden of illness and proposed initiatives to improve outcomes. *Am J Manag Care.* 2009;15(5 Suppl):S151–S159
15. Neale BM, Medland SE, Ripke S, et al; Psychiatric GWAS Consortium: ADHD Subgroup. Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry.* 2010;49(9):884–897
16. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet.* 2007;81(5):1084–1097
17. Urwin H, Ghazi-Noori S, Collinge J, Isaacs A. The role of CHMP2B in frontotemporal dementia. *Biochem Soc Trans.* 2009;37(pt 1):208–212
18. Elia J, Gai X, Hakonarson H, White PS. Structural variations in attention-deficit hyperactivity disorder. *Lancet.* 2011;377(9763):377–378, author reply 378
19. Fanta CH. Asthma. *N Engl J Med.* 2009;360(10):1002–1014
20. Moffatt MF, Gut IG, Demenais F, et al; GABRIEL Consortium. A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med.* 2010;363(13):1211–1221
21. Bisgaard H, Bønnelykke K, Sleiman PM, et al. Chromosome 17q21 gene variants are associated with asthma and exacerbations but not atopy in early childhood. *Am J Respir Crit Care Med.* 2009;179(3):179–185
22. Galanter J, Choudhry S, Eng C, et al. ORM DL3 gene is associated with asthma in three ethnically diverse populations. *Am J Respir Crit Care Med.* 2008;177(11):1194–1200
23. Leung TF, Sy HY, Ng MC, et al. Asthma and atopy are associated with chromosome 17q21 markers in Chinese children. *Al-lergy.* 2009;64(4):621–628
24. Cantero-Recasens G, Fandos C, Rubio-Moscardo F, Valverde MA, Vicente R. The asthma-associated ORM DL3 gene product regulates endoplasmic reticulum-mediated calcium signaling and cellular stress. *Hum Mol Genet.* 2010;19(1):111–121
25. Himes BE, Hunninghake GM, Baurley JW, et al. Genome-wide association analysis identifies PDE4D as an asthma-susceptibility gene. *Am J Hum Genet.* 2009;84(5):581–593
26. DeWan AT, Triche EW, Xu X, et al. PDE11A associations with asthma: results of a genome-wide association scan. *J Allergy Clin Immunol.* 2010;126(4):871–873.e9
27. Sleiman PM, Flory J, Imielinski M, et al. Variants of DENND1B associated with asthma in children [published correction appears in *N Engl J Med.* 2012;366(7):672]. *N Engl J Med.* 2010;362(1):36–44
28. Al-Zoubi AM, Efimova EV, Kaithamana S, et al. Contrasting effects of IG20 and its splice isoforms, MADD and DENN-SV, on tumor necrosis factor alpha-induced apoptosis and activation of caspase-8 and -3. *J Biol Chem.* 2001;276(50):47202–47211
29. Mells GF, Floyd JA, Morley KI, et al; UK PBC Consortium; Wellcome Trust Case Control Consortium 3. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis [published correction appears in *Nat Genet.* 2011;43(11):1164]. *Nat Genet.* 2011;43(4):329–332
30. Baldassano RN, Bradfield JP, Monos DS, et al. Association of the T300A non-synonymous variant of the ATG16L1 gene with susceptibility to paediatric Crohn's disease. *Gut.* 2007;56(8):1171–1173
31. Baldassano RN, Bradfield JP, Monos DS, et al. Association of variants of the interleukin-23 receptor gene with susceptibility to pediatric Crohn's disease. *Clin Gastroenterol Hepatol.* 2007;5(8):972–976
32. Kugathasan S, Baldassano RN, Bradfield JP, et al. Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat Genet.* 2008;40(10):1211–1215
33. Imielinski M, Hakonarson H. Breaking new ground in inflammatory bowel disease genetics: genome-wide association studies and beyond. *Pharmacogenomics.* 2010;11(5):663–665
34. Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science.* 2006;314(5804):1461–1463
35. Libioulle C, Louis E, Hansoul S, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 2007;3(4):e58
36. Franke A, McGovern DP, Barrett JC, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet.* 2010;42(12):1118–1125
37. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet.* 2011;43(3):246–252
38. Wang K, Dickson SP, Stolle CA, Krantz ID, Goldstein DB, Hakonarson H. Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am J Hum Genet.* 2010;86(5):730–742
39. Imielinski M, Baldassano RN, Griffiths A, et al; Western Regional Alliance for Pediatric IBD; International IBD Genetics Consortium; NIDDK IBD Genetics Consortium; Belgian-French IBD Consortium; Wellcome Trust Case Control Consortium. Common variants at five new loci associated with

- early-onset inflammatory bowel disease. *Nat Genet.* 2009;41(12):1335–1340
40. Lecat A, Piette J, Legrand-Poels S. The protein Nod2: an innate receptor more complex than previously assumed. *Biochem Pharmacol.* 2010;80(12):2021–2031
 41. Morrison PJ, Ballantyne SJ, Kullberg MC. Interleukin-23 and T helper 17-type responses in intestinal inflammation: from cytokines to T-cell plasticity. *Immunology.* 2011;133(4):397–408
 42. Cadwell K, Stappenbeck TS, Virgin HW. Role of autophagy and autophagy genes in inflammatory bowel disease. *Curr Top Microbiol Immunol.* 2009;335:141–167
 43. Lee JC, Parkes M. Genome-wide association studies and Crohn's disease. *Brief Funct Genomics.* 2011;10(2):71–76
 44. Troiano RP, Flegal KM. Overweight children and adolescents: description, epidemiology, and demographics. *Pediatrics.* 1998;101(3 pt 2):497–504
 45. Nguyen NT, Nguyen XM, Lane J, Wang P. Relationship between obesity and diabetes in a US adult population: findings from the National Health and Nutrition Examination Survey, 1999–2006. *Obes Surg.* 2011;21(3):351–355
 46. Zalesin KC, Franklin BA, Miller WM, Peterson ED, McCullough PA. Impact of obesity on cardiovascular disease. *Med Clin North Am.* 2011;95(5):919–937
 47. Parsons TJ, Power C, Logan S, Summerbell CD. Childhood predictors of adult obesity: a systematic review. *Int J Obes Relat Metab Disord.* 1999;23(Suppl 8):S1–S107
 48. Must A. Does overweight in childhood have an impact on adult health? *Nutr Rev.* 2003;61(4):139–142
 49. Herbert A, Gerry NP, McQueen MB, et al. A common genetic variant is associated with adult and childhood obesity. *Science.* 2006;312(5771):279–283
 50. Scuteri A, Sanna S, Chen WM, et al. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* 2007;3(7):e115
 51. Hinney A, Nguyen TT, Scherag A, et al. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PLoS ONE.* 2007;2(12):e1361
 52. Dina C, Meyre D, Gallina S, et al. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet.* 2007;39(6):724–726
 53. Grant SF, Li M, Bradfield JP, et al. Association analysis of the FTO gene with obesity in children of Caucasian and African ancestry reveals a common tagging SNP. *PLoS One.* 2008;3(3):e1746
 54. Glessner JT, Bradfield JP, Wang K, et al. A genome-wide study reveals copy number variants exclusive to childhood obesity cases. *Am J Hum Genet.* 2010;87(5):661–666
 55. Wei Z, Wang K, Qu HQ, et al. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.* 2009;5(10):e1000678
 56. Frayling TM, Colhoun H, Florez JC. A genetic link between type 2 diabetes and prostate cancer. *Diabetologia.* 2008;51(10):1757–1760
 57. Neurath MF, Fuss I, Kelsall BL, Stüber E, Strober W. Antibodies to interleukin 12 abrogate established experimental colitis in mice. *J Exp Med.* 1995;182(5):1281–1290
 58. Martínez A, Varadé J, Márquez A, et al. Association of the STAT4 gene with increased susceptibility for some immune-mediated diseases. *Arthritis Rheum.* 2008;58(9):2598–2602
 59. Zhernakova A, Festen EM, Franke L, et al. Genetic analysis of innate immunity in Crohn's disease and ulcerative colitis identifies two susceptibility loci harboring CARD9 and IL18RAP. *Am J Hum Genet.* 2008;82(5):1202–1210
 60. Leach ST, Messina I, Lemberg DA, Novick D, Rubenstein M, Day AS. Local and systemic interleukin-18 and interleukin-18-binding protein in children with inflammatory bowel disease. *Inflamm Bowel Dis.* 2008;14(1):68–74
 61. Sato K, Shiota M, Fukuda S, et al. Strong evidence of a combination polymorphism of the tyrosine kinase 2 gene and the signal transducer and activator of transcription 3 gene as a DNA-based biomarker for susceptibility to Crohn's disease in the Japanese population. *J Clin Immunol.* 2009;29(6):815–825
 62. Glessner JT, Wang K, Cai G, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature.* 2009;459(7246):569–573
 63. Glessner JT, Reilly MP, Kim CE, et al. Strong synaptic transmission impact by copy number variations in schizophrenia. *Proc Natl Acad Sci USA.* 2010;107(23):10584–10589
 64. Levinson DF, Duan J, Oh S, et al. Copy number variants in schizophrenia: confirmation of five previous findings and new evidence for 3q29 microdeletions and VIPR2 duplications. *Am J Psychiatry.* 2011;168(3):302–316
 65. Chen X, Li X, Wang P, et al. Novel association strategy with copy number variation for identifying new risk loci of human diseases. *PLoS ONE.* 2010;5(8):e12185
 66. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics.* 2008;1(1):4
 67. Craddock N, Hurles ME, Cardin N, et al; Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature.* 2010;464(7289):713–720
 68. Conrad DF, Pinto D, Redon R, et al; Wellcome Trust Case Control Consortium. Origins and functional impact of copy number variation in the human genome. *Nature.* 2010;464(7289):704–712
 69. Pang AW, MacDonald JR, Pinto D, et al. Towards a comprehensive structural variation map of an individual human genome. *Genome Biol.* 2010;11(5):R52
 70. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet.* 2004;36(9):949–951
 71. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science.* 2004;305(5683):525–528
 72. Morrow EM. Genomic copy number variation in disorders of cognitive development. *J Am Acad Child Adolesc Psychiatry.* 2010;49(11):1091–1104
 73. Wang K, Chen Z, Tadesse MG, et al. Modeling genetic inheritance of copy number variations. *Nucleic Acids Res.* 2008;36(21):e138
 74. Lupski JR. Genomic rearrangements and sporadic disease. *Nat Genet.* 2007;39(suppl 7):S43–S47
 75. Sebat J, Lakshmi B, Malhotra D, et al. Strong association of de novo copy number mutations with autism. *Science.* 2007;316(5823):445–449
 76. Bucan M, Abrahams BS, Wang K, et al. Genome-wide analyses of exonic copy number variants in a family-based study point to novel autism susceptibility genes. *PLoS Genet.* 2009;5(6):e1000536
 77. Kim HG, Kishikawa S, Higgins AW, et al. Disruption of neurexin 1 associated with autism spectrum disorder. *Am J Hum Genet.* 2008;82(1):199–207
 78. Guilmatre A, Dubourg C, Mosca AL, et al. Recurrent rearrangements in synaptic and neurodevelopmental genes and shared biologic pathways in schizophrenia, autism, and mental retardation. *Arch Gen Psychiatry.* 2009;66(9):947–956
 79. Yi JJ, Ehlers MD. Ubiquitin and protein turnover in synapse function. *Neuron.* 2005;47(5):629–632

80. Walsh T, McClellan JM, McCarthy SE, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*. 2008;320(5875):539–543
81. Shi YY, He G, Zhang Z, et al. A study of rare structural variants in schizophrenia patients and normal controls from Chinese Han population. *Mol Psychiatry*. 2008;13(10):911–913
82. Need AC, Ge D, Weale ME, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet*. 2009;5(2):e1000373
83. Girirajan S, Rosenfeld JA, Cooper GM, et al. A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay. *Nat Genet*. 2010;42(3):203–209
84. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008;452(7189):872–876
85. Lander ES, Linton LM, Birren B, et al; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome [published correction appears in *Nature*. 2001;409(6822):860–921]. *Nature*. 2001;409(6822):860–921
86. Yandell M, Huff C, Hu H, et al. A probabilistic disease-gene finder for personal genomes. *Genome Res*. 2011;21(9):1529–1542
87. Edwards TL, Song Z, Li C. Enriching targeted sequencing experiments for rare disease alleles. *Bioinformatics*. 2011;27(15):2112–2118
88. Rehman AU, Morell RJ, Belyantseva IA, et al. Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. *Am J Hum Genet*. 2010;86(3):378–388
89. Levin JZ, Berger MF, Adiconis X, et al. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol*. 2009;10(10):R115
90. Volpi L, Roversi G, Colombo EA, et al. Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. *Am J Hum Genet*. 2010;86(1):72–76
91. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001;69(1):124–137
92. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*. 2010;42(1):30–35
93. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. *Hum Genet*. 2011;129(4):351–370
94. Olbrich H, Häffner K, Kispert A, et al. Mutations in DNAH5 cause primary ciliary dyskinesia and randomization of left-right asymmetry. *Nat Genet*. 2002;30(2):143–144
95. Lyon GJ, Jiang T, Van Wijk R, et al. Exome sequencing and unrelated findings in the context of complex disease research: ethical and clinical implications. *Discov Med*. 2011;12(62):41–55
96. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*. 2010;362(13):1181–1191
97. Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Hum Mol Genet*. 2010;19(22):4313–4318
98. Shen Y, Dies KA, Holm IA, et al; Autism Consortium Clinical Genetics/DNA Diagnostics Collaboration. Clinical genetic testing for patients with autism spectrum disorders. *Pediatrics*. 2010;125(4). Available at: www.pediatrics.org/cgi/content/full/125/4/e727
99. Schaaf CP, Wiszniewska J, Beaudet AL. Copy number and SNP arrays in clinical diagnostics. *Annu Rev Genomics Hum Genet*. 2011;12:25–51
100. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010;11(12):843–854
101. Glessner JT, Hakonarson H. Genome-wide association: from confounded to confident. *Neuroscientist*. 2011;17(2):174–184