Genome-Wide Association Studies in Pediatric Endocrinology

Andrew Dauber, Joel N. Hirschhorn

Abstract
Genome-wide association (GWA) studies are a powerful tool for understanding the genetic underpinnings of human disease. In this article, we briefly review the role and findings of GWA studies in type 1 diabetes, stature, pubertal timing, obesity, and vitamin D deficiency. We then discuss the present and future implications of these findings with regards to disease prediction, uncovering basic biology, and the development of novel therapeutic agents.

Introduction
Traditionally, endocrinologists have focused on hormone levels and their effects on human disease. Genetics has played a limited role in pediatric endocrinology, primarily utilized in the diagnosis of chromosomal disorders, such as Turner syndrome and Klinefelter syndrome, or rare mendelian conditions (e.g. congenital adrenal hyperplasia) that present with endocrine disorders. The majority of the disorders treated by most pediatric endocrinologists do not fall into these categories, but instead are complex genetic or polygenic disorders, meaning that multiple genes influence the pathogenesis of these conditions.

Over the past few years, tremendous progress has been made in understanding the genetic basis of complex diseases through the use of genome-wide association (GWA) studies. In this article, we will briefly review the role and findings of GWA studies in type 1 diabetes (T1D), stature, pubertal timing, obesity, and vitamin D deficiency. We will then discuss the present and future implications of these findings with regards to disease prediction, uncovering basic biology, and the development of novel therapeutic agents.

Prior to embarking on the individual topic reviews, we provide a brief overview of three major genetic approaches that have been used to study complex diseases [for a more detailed review, see 1]. We do not discuss in detail large-scale sequencing studies, which will likely be applied with greater regularity as sequencing costs continue to fall.

The first approach that has been used to study polygenic diseases is the classic linkage study. In a linkage study, large families with multiple individuals affected with a particular disease are studied with a moderately
sparse set of highly polymorphic genetic markers spread across the genome. Investigators then use the genotype data to search for a region of the genome that is consistently inherited by individuals with the disease but not by unaffected family members. The advantage of this genome-wide approach is that it is unbiased and allows for discovery of novel genes without prior knowledge of their function. The disadvantage is that it works best in large families with multiple affected individuals and with genetic variants that are highly penetrant, meaning that individuals who inherit the variant will almost certainly develop the disease. Ultimately, linkage studies can identify a genomic region of interest, but their success has been largely limited to single-gene disorders (with a few notable exceptions such as T1D, as described below). Even in those cases, the region of linkage can be quite large and contain many possible causal genes, necessitating follow-up sequencing studies.

The second approach that has been used is a candidate-gene association study. In this type of study, individual candidate genes are selected and then the frequencies of variants in those genes are compared in subjects with the disease versus matched controls (or, in the case of quantitative traits, correlations are sought between the genotype and the phenotype). This study design is excellent for pursuing an interesting candidate gene or short list of genes. Its major limitation is that it is a hypothesis-driven approach to genetic investigation, as the study design requires the investigator to choose the candidate genes a priori. This approach has had some limited success in determining the genetic etiology of complex traits.

The third approach is the GWA study [for more detailed reviews, see 1–3]. The design is based on the traditional association study where genetic variants are compared in cases versus controls. The major distinction is that GWA studies are unbiased approaches that examine common variants, usually single nucleotide polymorphisms (SNPs), across the entire genome. In recent years, genotyping platforms have been developed that allow for the simultaneous genotyping of hundreds of thousands of SNPs across the entire genome on a single chip. This approach allows investigators to test each SNP for an association with the disease of interest. It is important to realize that these SNPs are often not the actual causal genetic variant but rather that they mark a region (referred to as a locus) in which the causal variant lies [3]. The major disadvantage of GWA studies is that due to the multitude of variants tested, one must have strict statistical thresholds for calling a variant statistically significant to avoid false-positive results from multiple hypothesis testing. A traditionally used cutoff which accounts for testing the approximately 1 million independent common variants across the genome is a p value below $5 \times 10^{-8}$ [for detailed statistical consideration, see 2]. In order to reach this stringent level of significance, especially in complex traits where each variant often has a small effect size, very large numbers of subjects are needed (often tens of thousands in each arm of the study). This requirement for large samples has led to the formation of many large collaborative consortia that have performed meta-analyses, some of which will be discussed in the individual topic reviews that follow.

Type 1 Diabetes

The genetic underpinnings of T1D have been the source of much investigation. Twin studies suggest that genetic variation may explain up to 70% of the variability in disease risk. However, there is clearly an environmental component as well, supported by the recent increasing incidence and the variability in incidence between genetically similar populations [reviewed in 4]. Early association studies showed that approximately half of the increased risk to siblings of patients with T1D is explained by genetic variants in the HLA region [5, 6]. Variants in the HLA locus increase risk for many autoimmune diseases, consistent with the fact that many genes in this region encode for key components of the immune response [6]. Prior to the advent of GWA studies, linkage and candidate gene studies highlighted 4 additional genes that play a significant role in diabetes pathogenesis, the insulin gene, CTLA4, PTPN22, and IL2RA [reviewed in 7]. CTLA4 (cytotoxic T-lymphocyte-associated protein 4) is a co-stimulatory receptor that functions in regulatory T cells. PTPN22 encodes an inhibitor of T cell activation while IL2RA encodes CD25, which is important in suppressing T cell proliferation. These genes, along with the HLA locus, strongly implicate the T cell pathway in T1D pathogenesis. This theme has been further supported by genes emerging from GWA studies, but novel pathways have been implicated as well.

The most recent meta-analysis of GWA studies in T1D [8], including a total of 7,514 cases and 9,045 controls, identified over 40 loci that reached genome-wide significance. The effect sizes of the non-HLA loci range from an odds ratio of 1.05–2.38 for the associated insulin gene variant. These loci include multiple genes involved in the immune response. They also include genes expressed in β-cells such as IFIHI, GLIS3, and PTPN2 [7].
ingly, a number of the genes involved in immune response that influence T1D risk have also been highlighted by GWA studies of other autoimmune diseases. These shared genetic risk factors partially explain the shared epidemiology of autoimmune diseases. By examining the overlap and non-overlap between these various studies, scientists are beginning to unravel the common and distinct underlying pathogenesis of autoimmune diseases [6].

As mentioned above, a GWA study identifies a region of the genome that is associated with a phenotype, but it does not definitively identify the causal gene. The majority of the associated variants do not cause a change in a protein sequence, i.e., they are either non-coding or synonymous variants. Thus, identifying an associated variant is only the first step: in total the ~40 T1D loci include more than 300 genes. For some loci, it is quite obvious which gene is involved, as there is one clear biological candidate. However, in other instances, there are multiple plausible causal genes and further fine mapping of the association signals and functional studies are required to identify the true causal genes. Furthermore, some of the loci are located in areas of the genome without annotated genes, making the source of the association difficult to determine. Presumably, these regions have some type of long-range regulatory function affecting expression of a distant neighboring gene, but this hypothesis remains speculative.

A recent study provides an excellent example of the potential for follow-up investigation to identify causal genes within loci highlighted by GWA studies. Nejentsev et al. [9] performed follow-up sequencing to identify rare functional genetic variants in 10 T1D candidate genes, many of which were implicated by GWA studies. They identified 4 rare functional variants in the IFIHI gene that were strongly associated with T1D risk. Each of these variants significantly affects the expressed protein sequence, thereby impairing protein function. This finding strongly supports the notion that IFIHI is the causal gene within the associated locus. IFIHI (interferon induced with helicase C domain 1) encodes a protein that influences interferon response to viral DNA [10]. This gene provides evidence of an interaction between virus exposure and the host immune response in the development of T1D. Loss of function variants in IFIHI were protective for T1D risk, suggesting that chemical inhibitors of this gene product may have potential for diabetes prevention.

In addition to illuminating the pathogenesis of T1D, GWA studies have also identified variants that affect glycemic control and the development of medical complications. The Diabetes Control and Complications Trial group recently identified a variant near the SORC3 gene which was associated with hemoglobin A1c and mean glucose levels. The mechanism of this effect is currently unknown, but this gene has been implicated in the control of fasting insulin levels in rats and possibly humans [11]. Interestingly, this locus has not been associated with the risk of developing T1D or T2D. This suggests that its role in glycemic control may be independent of the underlying pathogenesis of diabetes. The Genetics of Kidneys in Diabetes (GoKinD) group performed a GWA study looking for susceptibility loci for diabetic nephropathy in T1D [12]. They found 4 loci with suggestive associations, but these loci require replication before considering them definitive. If confirmed, these findings may point to novel biology in the pathogenesis of diabetic nephropathy and could open up new fields of investigation.

### Stature

Until recently, little was known about the genetic determinants of height in the general population. Based on twin studies, it is estimated that up to 90% of the variation in height is heritable. Rare mutations have been described in genes underlying severe monogenic forms of short stature, such as the prototypical growth hormone receptor mutations in Laron dwarfism, but these genes do not explain the majority of the polygenic variation in height in the general population [reviewed in 13]. In 2008, three GWA studies [14–16] marked the number of loci associated with stature, together identifying over 40 loci associated with adult height. This effort was recently expanded through the Genetic Investigations of ANthropometric Traits (GIANT) consortium meta-analysis of 183,727 individuals, which has brought the number of independent loci that are associated with adult height to 180 [17]. Despite this large number of loci, the associated variants only explain ~10% of the variance in adult height. Nevertheless, the genes surrounding these loci are enriched for genes that are known causes of skeletal growth disorders, supporting the hypothesis that individual loci can harbor both rare and common variants that manifest a range of phenotype severity (i.e. significant skeletal dysplasia vs. mild short stature). Furthermore, these loci highlight all of the current growth-promoting therapies, as they include the genes encoding growth hormone, IGF-1 receptor, aromatase, and the estrogen receptor. This result suggests that other loci may include potential novel therapeutic targets. Finally, gene

---

Horm Res Paediatr 2011;75:322–328

Dauber/Hirschhorn
Pubertal Timing

Pubertal timing is influenced by both genetic and environmental factors. Recent studies confirm a trend for earlier onset of puberty in females [18] over a time frame that must be due to changing environmental influences rather than population-wide genetic changes. Nevertheless, 50–80% of the variation in pubertal timing is thought to be due to genetic factors [19]. Many studies have investigated the role of rare variants in candidate genes in cases of delayed puberty, especially idiopathic hypogonadotrophic hypogonadism (IHH) and Kallmann syndrome [reviewed in 19–21]. While there has been much success in this field, common variants in the majority of genes underlying IHH and Kallmann syndrome have not as of yet been found to affect timing of puberty in the general population [22]. Recently, a meta-analysis did find an association between variants in TAC3R, which encodes a neurokinin B receptor and is a known cause of IHH, and ESR1, the estrogen receptor gene, with age at menarche [23].

In 2009, four GWA studies were published which identified 2 loci associated with age at menarche in the general population [24–27]. The first locus, on chromosome 6q21, is near the LIN28B gene, a regulator of the let-7 class of microRNAs [28]; the second locus, on chromosome 9q31.2, has no associated gene nearby. These 2 loci explain 0.6% of the variation in the timing of menarche [24] and each allele causes a change of ~1.2 months in pubertal timing [26]. One of these studies confirmed that the variant in LIN28B is associated both with timing of breast development in girls as well as pubic hair stage and tempo of growth in boys [25]. Interestingly, a different variant in LIN28B was identified in 1 of the GWA studies for stature [15] indicating that individual loci and genes may have effects on multiple phenotypes. A transgenic mouse which overexpresses Lin28a, a homolog of LIN28B, has been shown to have delayed puberty and increased size, supporting a role for LIN28B in both growth and the timing of puberty [29].

Recently, a large meta-analysis including over 80,000 women identified an additional 30 loci associated with age of menarche [23]. These loci highlighted a few candidate genes including INHBA which encodes the protein subunit of inhibin-βA. This subunit is part of inhibin A and activin A, both of which play important roles in the regulation of the hypothalamic-pituitary-gonadal axis. Additional candidates include RXRG which encodes a nuclear hormone receptor and PCSK2 which is the gene for the proprotein convertase that cleaves proleukinizing hormone-releasing hormone. Additionally, this study identified 4 previously known obesity loci and 3 loci with genes involved in energy homeostasis supporting the known clinical association between obesity and timing of puberty. Finally, gene pathway analysis highlighted coenzyme A and fatty acid biosynthesis as processes related to timing of menarche. These findings will open up many new avenues for research of the underlying mechanisms of pubertal timing.

Obesity

Obesity is a growing concern for the pediatric population. According to the National Health and Nutrition Examination Survey 2007–2008 (NHANES), 16.9% of children ages 2 through 9 were at or above the 95th percentile for BMI and 31.7% were at or above the 85th percentile, the cutoffs for obesity and overweight, respectively [30]. Although the recent increase in the obesity epidemic is clearly due to environmental influences and not a rapid change in the genetic background, there is a strong genetic predisposition to obesity with heritability estimates ranging from 40 to 70% [31]. The GIANT consortium recently completed a GWA study meta-analysis for BMI which included 249,796 individuals [32]. They identified 32 loci that reached the level of genome-wide significance, 14 of which had been previously found to be associated with BMI or other obesity-related traits. These 32 loci only explained 1.45% of the variance in BMI in their samples and accordingly had poor predictive value. The majority of these loci were confirmed to have directionally consistent associations in children with early-onset severe obesity and with BMI in children and adolescents. Genes implicated by these loci include POMC, MC4R, BDNF, and SH2B1 which are involved in the neuro-regulatory pathways of energy homeostasis and appetite regulation. The authors also highlight the GIPR gene, located near the 19q13.2 locus, which encodes the receptor for gastric inhibitory polypeptide (GIP). GIP is an incretin hormone that affects insulin response to a glucose load. They hypothesize that there may be a connection between incretins, insulin secretion, and risk of obesity.
consistent with other suggestions [33]. Additional pathway analyses found enrichment of genes in multiple biological pathways, which will be targets for future research. Most of the implicated genes and pathways represent biological processes that are not known to be involved in obesity, highlighting our limited molecular understanding of the pathogenesis of this disease. Furthermore, there are currently no safe and effective medical therapies for the treatment of obesity, so these GWA results represent an opportunity to revolutionize our understanding and therefore treatment of obesity.

In addition to examining genetic variants associated with BMI and generic obesity, work has been done looking at genetic causes of extreme obesity. We and others performed a GWA study of a cohort of 972 individuals presenting for gastric bypass surgery with a mean BMI of 50.6 [34]. We identified convincing evidence for association with the FTO locus and supportive evidence implicating approximately half of the BMI loci that had previously been identified in the general population. Furthermore, individuals with higher BMI had a greater number of risk alleles when compared to individuals with a lower BMI. Prior to this analysis, it was unknown whether extreme obesity simply included individuals on the extreme ends of the spectrum of the polygenic variation in BMI or if those individuals had distinct rare genetic etiologies that are not influenced by common genetic variants, essentially defining novel monogenic or oligogenic conditions. This study supports the notion that the genetic causes of extreme obesity include the polygenic variants that contribute to milder obesity, with the more extreme individuals having an overall higher burden of risk alleles. However, other studies of severe obesity have identified loci not seen in studies of the general population [35], and rare functional mutations in MC4R are more frequent contributors to early-onset severe obesity than of milder obesity in adults [36], suggesting that additional genetic factors may play a role in severe obesity but have less of an impact in milder population variation in obesity. Thus, severe obesity may lie on a spectrum between rare monogenic diseases that have obesity as a feature and an accumulation of polygenic variants that leads to higher BMI.

**Vitamin D Deficiency**

Vitamin D deficiency is highly prevalent in children and adolescents around the world [reviewed in 37]. While the environmental influences on vitamin D, i.e. sun exposure and diet, are well known, studies estimate that vitamin D levels have a heritability of up to 53% [38]. Two recent GWA studies independently discovered the same 3 loci associated with vitamin D levels and deficiency [38, 39]. These loci include the genes for the vitamin D binding protein (GC), for the hepatic 25-hydroxylase gene (CYP2R1), and for 7-dehydrocholesterol reducetase as well as NAD synthetase (DHCR7/NADSYN1). These genes are strong biological candidates involved in the metabolism of vitamin D. The investigators designed a genotypic risk score based on the presence or absence of a risk allele at each locus. Individuals within the top quartile of genotypic risk score had over twice the risk of vitamin D deficiency compared with those in the lowest quartile based on these three SNPs alone [38]. Additionally, the effect on vitamin D levels of being homozygous for one of these risk alleles was similar to the effect of having received vitamin D supplementation [38]. It is unknown whether these risk variants modify the response to vitamin D supplementation or sun exposure, and further research is needed in this area.

**Conclusions and Future Directions**

Significant progress has been made in recent years in unraveling the genetic roots of complex diseases. GWA studies have played a major role in this effort as they have allowed for the discovery of novel genetic associations that are not based on prior knowledge. The studies reviewed above provide a synopsis of the major advances in several topics related to pediatric endocrinology.

There are a number of key concepts that can be gleaned from these studies. As can be seen from the studies of height and obesity, very large sample sizes are needed to detect variants with small effect sizes. Despite these large collaborative efforts, the predictive value of variants identified through GWA studies is generally quite low. It is unlikely that GWA studies of common variants will result in definitive risk prediction for the individual patient [for statistical considerations in T1D, see 5]. The question remains whether adding genotype information to known traditional predictive risk factors will prove to be clinically useful. As genetic studies uncover additional loci with common and rare variants, the predictive power will improve, but clinical utility will typically depend on the availability of an intervention to prevent the disease and the risk-benefit ratio and cost-effectiveness of that intervention. However, as seen with IFIHI [9], individual patients with rare functional variants in genes, in-
cluding those implicated by GWA studies, may have a much higher or lower risk of developing disease, and in this case, genotyping or sequencing may be quite useful for disease prediction. With the recent advancements in DNA sequencing technologies, future studies will likely sequence candidate genes within the loci identified by GWA studies and eventually throughout the genome, with hopes of identifying additional rare causal variants. Such rare variants may fall on a spectrum between the typically quite rare variants with large functional effects that have been shown to underlie monogenic forms of disease (including monogenic forms each of the diseases/traits discussed above), and the common variants with subtle functional implications that have thus far been identified by GWA studies. Because of the high background rate of rare variants that will be identified by extensive resequencing, functional studies to identify those rarer variants with larger functional effects may be particularly useful in this setting.

It should be clear that there will not be a DNA chip any time soon which can tell you precisely how tall your short stature patient will be, but below we discuss the more likely, if long-term, benefits of GWA studies. The true promise of GWA studies is that they allow for the discovery of novel biology. GWA studies open up new avenues for research such as the let-7 microRNA pathway in puberty and GIP in obesity. These pathways may point to novel biological mechanisms underlying these disorders as well as identify potential new therapeutic targets.

In addition to understanding the underlying biology of an individual’s predisposition to a disease, GWA studies can also identify genes involved in disease control and complications as described above with T1D. Although not discussed here, GWA studies are increasingly being used in pharmacogenetics studies to investigate genetic predictors of response to therapy. These studies have not yet been performed in the field of pediatric endocrinology.

Although current examples of immediate clinical impact from GWA studies are limited, we strongly believe that GWA studies have begun to revolutionize our understanding of the genetic underpinnings of human disease. As noted above, there is still a tremendous amount of work necessary to move from finding the associated locus to finding the causal gene, pathogenic variants, and biological insights. Fine mapping studies need to be done to further identify which gene is causing the association signal and then functional biological studies need to be performed to understand the underlying biology. As seen in stature and diabetes, GWA studies highlight genes that encode targets of currently known therapies and thus there is good reason to believe that other loci encompass future therapeutic targets. Importantly, the timeline for moving from identifying the locus to finding the causal gene to understanding the biology and finally to developing a novel therapy is quite long (likely at least 10–20 years). However, we believe that starting with GWA results will increase the likelihood of successful biological and therapeutic discoveries in the years to come.

References