Genetic Epidemiology and Nonsyndromic Structural Birth Defects: From Candidate Genes to Epigenetics

Dr. Charlotte A. Hobbs, MD, PhD, Dr. Shimul Chowdhury, PhD, Dr. Mario A. Cleves, PhD, Dr. Stephen Erickson, PhD, Dr. Stewart L. MacLeod, PhD, Dr. Gary M. Shaw, DPH, Dr. Sanjay J. Shete, PhD, Dr. John S. Witte, PhD, and Dr. Benjamin Tycko, MD, PhD

Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, AR (Drs. Hobbs, Chowdhury, Cleves, Erickson, MacLeod); Department of Pediatrics, Stanford University School of Medicine, Stanford, CA (Dr. Shaw); Department of Epidemiology, MD Anderson Cancer Center, Houston, TX (Dr. Shete); Institute for Cancer Genetics and Department of Pathology and Cell Biology, Columbia University Medical Center, New York, NY (Dr. Tycko); Department of Epidemiology and Biostatistics, University of California in San Francisco, San Francisco, CA (Dr. Witte)

Abstract

Birth defects are a leading cause of infant morbidity and mortality worldwide. The vast majority of birth defects are nonsyndromic, and although their etiologies remain mostly unknown, evidence supports the hypothesis that they result from the complex interaction of genetic, epigenetic, environmental, and lifestyle factors. Since our last review published in 2002 describing the basic tools of genetic epidemiology used to study nonsyndromic structural birth defects, many new approaches have become available and have been used with varying success. Through rapid advances in genomic technologies, investigators are now able to interrogate large portions of the genome at a fraction of previous costs. With next generation sequencing (NGS), research has progressed from assessing a small percentage of single nucleotide polymorphisms (SNPs) to assessing the entire human protein-coding repertoire (exome) – an approach that is starting to uncover rare but informative mutations associated with nonsyndromic birth defects. Here we report on the current state of genetic epidemiology of birth defects and comment on future challenges and opportunities. We consider issues of study design, and we discuss common variant approaches including candidate gene studies and genome-wide association studies (GWAS). We also discuss the complexities embedded in exploring gene-environment interactions. We complete our review by describing new and promising NGS technologies and examining how the study of...
epigenetic mechanisms could become the key to unraveling the complex etiologies of nonsyndromic structural birth defects.

Research on birth defects in the post-genomic era

On April 14th, 2003, the Human Genome Project (HGP) was completed. To celebrate the 10th year anniversary, JAMA published a thematic issue on genomics on April 10th, 2013. In 2002, in a commentary published in the Archives of Pediatrics and Adolescent Medicine, we described the basic tools of genetic epidemiology as applied to the study of nonsyndromic structural birth defects. Since that time, many new approaches have become available to discover genetic factors leading to birth defects, and epigenetics has also come to the forefront as an important contributing factor. Here we present an updated review of genetic epidemiology and birth defects, focusing on these new approaches.

The prevalence of structural birth defects varies globally, ranging from approximately 3% to 6% of all live births. These birth defects are a leading cause of infant mortality, and are more prevalent than most chronic diseases of childhood such as autism, pediatric cancers, and Type I diabetes. Most structural birth defects develop early in embryogenesis, during the first 10 weeks of pregnancy, and the majority of these defects occur in isolation affecting only one organ system. Despite recent progress in finding highly informative mutations in some cases a majority of nonsyndromic birth defects still does not appear to be accounted for by a single gene or chromosomal abnormality. The most prevalent defects are orofacial clefts, heart, neural tube, and limb defects. When birth defects are not associated with known multi-organ syndromes; they are referred to as nonsyndromic defects. The etiologies of most nonsyndromic structural birth defects remain incompletely understood, and most are thought to result from a complex interplay between genetic, epigenetic, environmental, and lifestyle factors. Maternal lifestyle factors, such as smoking, can alter developmental processes and expression of key developmental genes, such as GATA4, and the impact of environmental exposures and lifestyle on the developing fetus can in turn be influenced by maternal and fetal genetic susceptibilities.

While prior research on nonsyndromic birth defects has focused largely on the independent roles of environmental and/or lifestyle exposures and genetics, epigenetic causes have now begun to receive increased attention in this and many other human diseases. As early as 1940, Waddington, a British embryologist, geneticist, and philosopher of science, defined epigenetics as “…the interactions of genes with their environment which bring the phenotype into being”. Although studies on epigenetics and birth defects are still limited, they are very important as they may help to establish the molecular basis for gene-environment interactions.

Embryogenesis requires an intricate coordination of cell migration, proliferation, and death that ultimately determines three-dimensional events in embryo formation and development. The complexity of embryonic processes requires that multiple genes and biological pathways are involved in an intricate series of events that are susceptible to perturbations due to environmental exposures or maternal conditions. In the following sections we will (a) review current approaches to identify genetic factors associated with birth defects, (b)
introduce potential epigenetic approaches, and (c) provide a perspective on the challenges and opportunities for future studies.

**Candidate Gene Studies**

Studies to identify etiologies of nonsyndromic birth defects in humans are for the most part limited to demonstrating association between genetic variants and birth defect phenotypes. Such association studies are most readily compared with traditional case-control epidemiologic designs. Population-based samples of cases and controls or small nuclear families (mother, father and child) or both are enrolled in a study in which the investigator compares the frequencies of specific genetic variants, polymorphisms, in affected and unaffected individuals. Such comparisons could include the index children as well as one or both parents.

Once it has been established that genetic variations impact the occurrence of a particular birth defect, studies are designed to identify and evaluate candidate genes. Candidate genes are those that have been proven or are believed to be associated with the malformation from animal studies, known or suspected developmental and signaling pathways, or from studying a small number of human pedigrees, and thus need to be confirmed in large, carefully selected, and well characterized human population samples. The selection of appropriate candidate genes is a key step in such studies. Knowledge of the pathogenesis of the malformation and the function of one or more proteins implicated in the disease can facilitate the identification of a suitable candidate gene. Genes that are known to be associated with variations in biologically relevant metabolic processes also may be selected as candidate genes. For example, several studies have demonstrated the protective effect of maternal periconceptional intake of folic acid on the occurrence of orofacial clefts\(^\text{11}\), neural tube defects\(^\text{12}\), and cardiac defects.\(^\text{13}\) Subsequently several studies have investigated the association between genetic polymorphisms in the folate metabolic pathway and the risk of these three defect groups.\(^\text{14-16}\)

Single nucleotide polymorphisms (SNPs) are the most common sources of variation in human genomes.\(^\text{17}\) Each SNP is a difference in a single nucleotide at a specific site within the genome. For example, a SNP may form by substituting the nucleotide cytosine with the nucleotide thymine at a specific genomic location, annotated as C>T. If both of these variants (alleles) are compatible with life, then they can be present at detectable frequencies in the general population. In fact, we now know that there are approximately 7 million SNPs with a minor allele frequency (MAF) of over 5%.\(^\text{18}\) Until completion of the International HapMap Project in 2009, population- or family-based studies of candidate genes were limited to the detection of associations between phenotypes of interest and only a small number of functional SNPs within candidate genes. By identifying a dense map of common variants and their correlations with each other (i.e. linkage disequilibrium), investigations such as the International HapMap Project and the 1000 Genomes Project (Figure 1) have made it possible to test for associations between birth defects and common (MAF>5% in the population) or even somewhat rarer genetic variants within each gene.\(^\text{19,20}\) In the past 5 years, multiple reviews of candidate gene studies and birth defects have been published. Key pathways that have been implicated in the development of orofacial clefts and neural tube,
heart, and kidney defects include the Wnt signaling pathway, the BMP signaling pathway, the Hedgehog signaling pathway and variants in genes coding for key enzymes in the folate/homocysteine and oxidative stress pathways 21–26.

**Genome Wide Association Studies (GWAS) and Copy Number Variants (CNVs)**

Candidate gene or pathway approaches are limited by their reliance on preexisting knowledge of relevant pathways and/or mechanisms that may adversely impact embryogenesis. With the completion of the HapMap Project and the advances in technology that provide platforms to genotype large number of SNPs efficiently, GWAS was introduced as a popular method for studying common but genetically complex human diseases.27 In contrast to candidate-gene studies, GWAS are not limited by prior knowledge but rather take an agnostic approach in which no SNP is considered, *a priori*, to have a higher likelihood of being associated with the relevant phenotype than any other SNP in the genome.28 Between January 2005 and December 2012, over 1350 GWAS studies were reported. With varying success, GWAS have identified common SNPs associated with risks for specific pediatric or adult diseases, but only a few have been completed on birth defects, including one on hypospadias29 and several on nonsyndromic cleft lip with or without cleft palate (NSCL/P). Mangold et al. identified two loci, in chromosome bands 17q22 and 10q25.3, which were associated with this type of birth defect,30 while the other three studies identified strong associations between NSCL/P and a locus in 8q24.21.31–33 These findings support the utility of GWAS for identifying novel chromosomal regions associated with birth defects. However, most genetic studies of nonsyndromic birth defects continue to rely on candidate SNPs, leaving most of the genome unexplored.34,35 Thus, there is a need to comprehensively explore the genome to identify new regions harboring genes associated with birth defects.36 In addition to SNPs, which affect only a single nucleotide base, multiple lines of evidence indicate that copy number variants (CNVs) can play an important role in the etiology of some cases of birth defects. CNVs are defined as DNA sequences, ranging from kilobases to megabases in length, which are present in variable copy number in comparison to a reference genome.37 In the past decade, molecular techniques such as array-based comparative genomic hybridization (aCGH), genotyping microarrays, and high throughput DNA sequencing have given a much richer picture of this form of genetic variation, and CNVs have increasingly been discovered to be associated with birth defects. For example, a survey of CNVs in 114 subjects with tetralogy of Fallot38 and their unaffected parents identified 11 *de novo* CNVs that were absent or extremely rare in >2,000 controls, and pathogenic CNVs affecting the *GATA4* and *NODAL* genes have been found in more than one study of congenital heart disease.39 Rare and/or *de novo* CNVs have also been implicated in microphthalmia,40 congenital diaphragmatic hernia,41 cleft lip and/or palate,42 other craniofacial defects43 and renal defects.44

For most GWAS, replication and validation of findings is necessary to separate true relationships from chance findings (Type 1 statistical errors). Genetic variants identified by GWAS are theoretically in linkage disequilibrium with functional variants that may be causal. To identify causal variants, targeted resequencing of genomic regions in close
proximity to selected SNPs, and whole-exome sequencing are new technological advances that hold considerable promise. In contrast to SNPs, which are often simply regional markers, CNV associations with birth defects are more likely to be directly causal of the phenotype being studied - via gene dosage increase or decrease, direct gene disruption, and cis-acting effects via disruption of gene regulatory sequences.45

**Gene-environment Interactions**

Genetic predispositions in conjunction with environmental influences are thought to be implicated in most birth defects. However, investigations into birth defects and interactions between functional or marker SNPs and environmental or lifestyle exposures are few.8 This paucity of studies may in part reflect the difficulty of obtaining robust genetic and environmental measures on the same samples, and the increased statistical power needed to detect gene-environment interactions. Nonetheless, there have been some indications of an interaction between functional SNPs in the methyl donor pathway and the beneficial effects of folic acid supplementation against NTDs and other birth defects.35

**Next Generation Sequencing (NGS) Technologies**

The first human genome was sequenced over a period of 10 years using classical Sanger (dideoxy terminator) sequencing, at a cost of almost $3 billion.46 “Next-generation” massively parallel sequencing (NGS) has exploded into the research and clinical genetics arena since its availability in 2005, with major advantages including markedly reduced sequencing time, reduced cost per nucleotide base, and substantial increases in data output.47 With new technologies, a genome can be sequenced within days at a current cost less than $10,000. Reviews detailing NGS chemistries and the practical advantages of specific platforms have been published.48–50 Data storage and protection, data analysis, clinical interpretation, and ethical issues, such as whether to report incidental, potentially adverse, findings to the patient or family, remain major challenges in the incorporation of NGS into the clinical arena.51,52

For the discovery and clinical testing of genes implicated in common birth defects such as CHDs, neural tube defects (NTDs), and cleft lip and palate (CL/P), different NGS-based approaches may be utilized. One approach is to select candidate genes for targeted deep resequencing. For conditions in which multiple genes with strong evidence of disease pathogenicity exist, such as those identified through GWAS and CNV studies, a targeted resequencing approach may be warranted. Although targeted NGS gene panels are offered clinically for certain conditions such as mitochondrial disorders,51 they are not yet offered clinically for the major types of non-syndromic birth defects. Another approach is the interrogation of the exome (all gene exons, that is, all protein coding regions) of the genome. Although the human exome consists of only 1–2% of the entire genome, it is estimated that up to 85% of disease-causing mutations are harbored within it. The merits of sequencing the exome as a diagnostic tool have been delineated and exome sequencing is starting to aid in the diagnosis and treatment of inherited conditions.48,52 Whole genome sequencing is also available, but to date the sequencing cost and the difficulties associated with data warehousing, analysis and interpretation remain prohibitive.
NGS technology holds promise for non-syndromic birth defects research and genetic diagnosis because 1) it allows the simultaneous analysis of the many candidate genes that have been identified so far; and 2) the technology is capable of detecting rare genetic variation. Although GWAS have identified many loci associated with complex traits, common genetic variation accounts for only a small percentage of heritability. Thus, identification of rare genetic variation may yield larger effect sizes in complex diseases, including isolated birth defects.

The use of NGS for discovering causal mutations in non-syndromic birth defects is still in its early stages. NGS was used recently to show that de-novo mutations contribute to approximately 10% of severe CHD. Strikingly, several genes encoding histone modifying and chromatin remodeling enzymes, such as MLL2, CHD7, and KDM5-A and –B, were found mutated in that study, implicating epigenetic alterations in CHD pathogenesis. Likewise, NGS approaches may become fruitful for finding mutations, such as those already known in the HNF1B and PAX2 genes, which can underlie congenital kidney malformations. Genome scanning methods for CNVs continue to identify lesions underlying CHD, some predisposing to it generally, and others with lesion specificity and it can be expected that additional recurrent CNVs will detected in nonsyndromic birth defect cases using NGS.

**Clinical Testing using NGS**

NGS targeted resequencing of candidate genes, whole exome sequencing, and even whole genome sequencing are being offered clinically in a growing number of laboratories. The GeneTests website and the NIH genetic testing registry provide valuable resources for clinicians to identify clinical genetic testing laboratories. Recently, the merits, considerations, and challenges of using NGS technologies in a clinical diagnostic setting have been discussed in publications through the American College of Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP). Although resequencing assays are clinically available for complex diseases including cardiomyopathy, autism spectrum disorder and intellectual disability, targeted assays using NGS are not yet available for non-syndromic birth defects such as CHDs, NTDs and CL/P. Whole exome sequencing is a potential genetic testing option for these conditions, but large research studies are only now being conducted in non-syndromic birth defects. Nonetheless, the apparent multigenic nature of non-syndromic birth defects makes NGS experimental approaches ideal for gene discovery and further pathogenic mutation characterization. Overall, the evidence accumulated through many dedicated research studies suggests that non-syndromic CHDs, NTDs, and CL/P may be caused by multiple rare, familial, genetic mutations, interacting with maternal genotype and exposures. Large population studies utilizing NGS may delineate genetic causes of common birth defects, and may aid in preconception risk and guide future therapeutic interventions.

**Epigenetic Alterations in Non-syndromic Birth Defects**

Despite the tremendous advances in human genetics enabled by the HGP and brought to fruition with GWAS and NGS, many aspects of human embryology and biology still cannot
be adequately explained by genetics alone. Normal embryogenesis requires the specification of a multitude of cell types/organs that depend on transcriptional regulation programmed by epigenetic mechanisms, namely modifications to DNA and its associated proteins that define the distinct gene expression profiles for individual cell types at specific developmental stages. Disruption of such control mechanisms is associated with a variety of diseases with behavioral, endocrine or neurologic manifestations, and quite strikingly with disorders of tissue growth, which will in all likelihood include structural birth defects. As a precedent, several well-studied syndromic birth defects, including Prader-Willi syndrome, Angelman syndrome, Beckwith-Wiedemann syndrome and Russell-Silver syndrome are known to be caused by loss of imprinting, uniparental disomy, or deletion/mutation of epigenetically regulated genes.\textsuperscript{64}

An epigenetic trait can be defined as a “stably (somalically) heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence”.\textsuperscript{65} Several classes of epigenetic phenomena have been identified and recently reviewed.\textsuperscript{8,66} Epigenetic patterns, essential for controlling gene expression in normal growth and development, are established by a number of mechanisms including DNA methylation at cytosine residues in CpG dinucleotides and covalent modifications of histone proteins, as well as by less well understood mechanisms controlling long-range chromatin architecture within the cell nucleus. DNA methylation involves the transfer of a methyl group to cytosine in a CpG dinucleotide, catalyzed by DNA methyltransferase enzymes that establish and maintain these patterns through cell division. Importantly DNA methylation has been shown to be essential for normal development.\textsuperscript{66,67} DNMT1, the major maintenance methylase, has a high affinity for hemimethylated DNA\textsuperscript{68} and it therefore acts to propagate methylation patterns in somatic cell divisions, while other enzymes, such as DNMT3A-DNMT3L, are responsible for initiating the epigenetic patterns. Importantly, there is already some evidence for an interaction of environment with epigenetics. Research on mice with a mutation in the \textit{Agouti} gene has provided an excellent example of how maternal diet and epigenetics may affect fetal phenotypes.\textsuperscript{69} This model has shown that variations in maternal dietary constituents affecting the methyl donor pool, such as folic acid, can result in alterations in coat color in the offspring, as a result of differential CpG methylation.\textsuperscript{69,70} As another example, with clinical implications, a high folate diet given to \textit{H. felis}-infected gastric cancer-prone mice at weaning prevented the development of gastric dysplasia and cancer.\textsuperscript{71}

Many techniques, which we reviewed recently,\textsuperscript{64} have been developed to study DNA methylation. The gold standard for comprehensive analysis of the methylation status of CpG sites, that is, DNA methylation patterns, is sodium bisulfite chemical conversion of DNA.\textsuperscript{72} This procedure deaminates non-methylated deoxycytidine (dC) to deoxyuracil (dU) residues; during subsequent PCR amplification, the latter are converted to A/T base pairs. However, if the C is methylated, the DNA sequence obtained after PCR does not change. Methylation-specific PCR or Pyrosequencing using bisulfite converted DNA provide quantitative measurements of DNA methylation levels, while another approach, which has now been made high throughput via NGS, involves amplification of bisulfite PCR products followed by sequencing of clones. This more thorough approach permits DNA methylation levels of a large number of contiguous CpG sites to be quantified, and the precise patterns of
methylation, including clonal heterogeneity and allele-specificity, to be displayed. By combining sodium bisulfite conversion and microarrays or NGS, genome-wide DNA methylation patterns can be determined essentially genome-wide. A number of initiatives have now been implemented to define human epigenetic patterns at high resolution with complete genomic coverage, with the goal of integrating epigenetics into the study of common but complex human diseases.\textsuperscript{73} These new technologies and approaches may provide keys to unravel genetic and environmental factors that impinge on epigenomes to affect normal processes in embryological development and lead to human malformations when these processes go awry. In searching for evidence of “epigenome x environment” interactions in datasets from such studies, it will be crucial to take the cell types being sampled and the age of the subjects into account as methylation patterns are highly cell type-specific and, for some genes, can change with age.\textsuperscript{74} It will also be important to remember that the genetic makeup of an individual exerts a strong influence on his/her epigenome: specifically, multiple studies have now shown a strong influence of human haplotypes, that is, clusters of SNPs in a given chromosomal region, on the patterns of DNA methylation in that region.\textsuperscript{75}

The Future: linking genome, epigenome and environment in non-syndromic birth defects

We concluded our 2002 article by stating that identification of genes associated with birth defects does not lead to an immediate understanding of the relation between the gene and the birth defect with which it is associated. Identification “is only the first step in a long path to understanding the cause of the condition and ultimately to find preventive or corrective strategies.\textsuperscript{1} As new technologies are made available, genetic epidemiologists are quick to utilize these new platforms to generate databases that seem to be ever-increasing in size. Our increased understanding of the importance of epigenetics in the development of birth defects suggests that an approach that simultaneously investigates genome-wide genetic and epigenetic variation in participants for whom environmental exposure data has been obtained may be a major step forward. Such studies may help to establish the mechanistic link between genetic variants and environmental exposures.

Acknowledgments

Funding/Support: This research was supported by the Centers for Disease Control and Prevention (5U01 DD000491) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01 HD039054 to CAH and P01 HD035897 to BT) and the Arkansas Biosciences Institute.

References


JAMA Pediatr. Author manuscript; available in PMC 2015 April 01.


56. Gelb BD. Recent advances in understanding the genetics of congenital heart defects. Curr Opin Pediatr. 2013
59. Genomics ACoMGa. ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing.


Figure 1.
Stepping stones: Projects that made genome-wide association studies (GWAS) possible.