1. ABSTRACT

Over the last 15 years neonatal morbidity and mortality have changed little for very low birth weight (VLBW) babies despite significant technological and therapeutic advances. While clinical trials and animal models have until recently improved outcomes in this gestational age group, further productivity from these traditional sources are not likely. A recent study of monozygotic and dizygotic twins shows that the main determinants of neonatal morbidity and mortality in VLBW babies – bronchopulmonary dysplasia, necrotizing enterocolitis, and intraventricular hemorrhage – have significant genetic components. Incremental improvements in the future, therefore, will likely depend on identification of these genetic components for targeting specific therapies. Cost-effective methods and resources, fueled by the Human Genome and HapMap Projects and recent successes in identifying genes for a small number of complex genetic diseases, are available now and through creative planning and timely implementation would likely yield useful results.

2. INTRODUCTION

Advances in perinatal-neonatal medicine over the past 30 years have produced remarkable achievements in lowering neonatal morbidity and mortality due to complications of pregnancy, labor, delivery, and the consequences of premature birth. Supplementary oxygen therapy, assisted mechanical ventilation, invasive physiologic monitoring, total parenteral nutrition, surfactant replacement, and extra-corporeal membrane oxygenation are all examples of noteworthy milestones in the short history of one of medicine’s youngest subspecialties. Yet, despite these remarkable achievements, in some areas - notably in complications from prematurity – improvements have produced little change over the past 15 years. For example, between 1987 and 1994 the NICHD Neonatal Research Network reported a decrease in mortality in neonates with birthweights between 500 and 1500 grams, from 23% to 17%. Between 1994 and 2000, however, mortality in these very low birth weight (VLBW) newborns, mainly from intraventricular hemorrhage (IVH), necrotizing enterocolitis (NEC), and
bronchopulmonary dysplasia (BPD), changed little (1). What comprises this seemingly impenetrable statistical barrier? Is it because over the past 10 years neonatal medicine has approached some technological obstacle, or physiologic chasm (i.e., the distance from terminal bronchiole to pulmonary capillary), or an unrecognized and virulent acquired neonatal infection, or perhaps an unrecognized fundamental physiologic property unique to fetuses at mid-trimester? Certainly any and all of these are possible explanations, and several are subjects of current clinical trials.

Alternatively, consider the transformation of medical genetics – focusing on rare mendelian disorders and chromosomal aberrations – to genetic medicine, which reflects the pervasiveness of genetic factors in all parts of clinical medicine and, that genetic factors are involved in all diseases. Over the past 30 years, largely through heritability studies of twins, we have learned that genetic factors contribute to as much as 54% of coronary artery disease, 30% of hypertension, 42% of cancer, and 60 - 80% of psychiatric disorders (2-7). This short list comprising the most common diseases of adults, were until recently, considered to be mostly sporadic in origin, perhaps familial, but infrequently genetic. Figure 1A is a conventional depiction of the relationships of onset of genetic disease and age from fetal life through adulthood (reviewed in Childs, 2002) (8). Prenatally, chromosomal aberrations, such as aneuploidy, are responsible for many spontaneous abortions and miscarriages (Figure 1A, peak from conception to 3 months on the left side of the graph). While inborn errors of metabolism can also result in abortion, the placenta frequently can support the pregnancy even with profound errors, which then manifest at birth.

Figure 1. Continuity of disease across the lifetime. (With permission from Elsevier) (8). A. Conventional depiction of genetic disease from conception to adulthood. B. Recognition of complex diseases in the premature VLBW neonate born between 24 weeks and term. White peaks are chromosomal aberrations. Grey peaks are mendelian single-gene disorders. Hatched peaks are complex disorders.
shortly after fetal-placental separation (Figure 1A, grey peak at birth). These are primarily the rare single gene recessive disorders (ie, mendelian) such as methylmalonic aciduria. In contrast, complex diseases arising from a combination of genetic factors and environmental effects, are the most common of human diseases, and conventionally are depicted as manifesting in adulthood (Figure 1A, hatched peak after puberty rising through adulthood). Complex disorders include the short list described above as well as type I and type II diabetes, dyslexia, obesity, eating disorders, and congenital heart disease. Screening for hyperlipidemia, hypercholesterolemia, and hypertension are routinely performed by pediatricians, who are increasingly recognizing complex disorders in adolescents and children. Studies of twins in the neonatal period now suggest that complex disorders should be considered even earlier.

In Figure 1B the mode of complex disorders is modified to include a peak between 6 months and term gestation corresponding with the birth and hospital course of the typical VLBW neonate. This peak is a summation of the variance due to tolerance in the VLBW neonate for positive pressure ventilation, oxygen toxicity, parenteral nutrition, opportunistic infections, swings in blood pressure from fluid resuscitation, gastric feedings, and the myriad of other therapies necessary to sustain the life of a VLBW neonate. In this paradigm BPD is a consequence of prematurity and positive pressure ventilation and oxygen therapy, as well as the genetic factors that enable tolerance to their inherent toxicities. For example, hypothesize that inheritance of a particular allele of the superoxide dismutase gene (SOD) creates susceptibility to oxygen free radicals in terminal bronchioles at 24 through 28 weeks of gestation. If a neonate is delivered at 38 weeks (term) or perhaps if we could identify a specific genetic factors for IVH, BPD, NEC, and retinopathy of prematurity (ROP), perhaps they would suggest novel drug therapies. Or perhaps if we could discern the operative genetic factors for IVH, BPD, NEC, and retinopathy of prematurity (ROP), perhaps they would suggest novel drug therapies. Or perhaps if we could identify a specific genetic risk factor for NEC we might consider alternative feeding regimens. Fundamentally however, consideration of the diseases of prematurity as complex disorders permits alternative approaches to understanding the basic pathophysiology – beyond ‘global immaturity’ – that takes advantage of the explosion of genetic information, tools, and resources that have been developed over the past two decades. These alternative genetic approaches, increasingly successful in identifying the key pathophysiologic factors of diseases we now routinely accept as complex in origin in adults, could also be used to address complex disorders in the neonate.

3. GENE IDENTIFICATION STRATEGIES

Recognition of genetics as a cause of human disease began in earnest in the early 1900’s with the rediscovery of Mendel’s work on peas 35 years earlier, the description of orderly inheritance of disease, the chromosomal theory of heredity and description of the first autosomal recessively-inherited mendelian disorder called alkaptonuria, and definition of the “gene” as the chromosomal unit of heredity (9-11). Yet, between 1931 and 1980, while the list of recognized inherited disorders in humans grew into the thousands (OMIM, Online Mendelian Inheritance in Man, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), only about 100 genes capable of causing human disease were known. Most of these were identified by an aberrant protein – some abnormal protein biophysical property or function that could be measured. The sickle variant of the human beta globin protein is the prototypical example. In sickle cell disease abnormal beta globin protein folding produces a signature migration change visible by starch gel electrophoresis, which is caused by a glutamic acid to valine substitution in the beta globin gene at amino acid number 6 (12). This amino acid change, Glu6Val, is encoded in nucleotide number 70 of the mRNA, GAG>GUG, transcribed from the beta globin gene encoded on chromosome 11p15.

Map-based approaches, which are currently the most popular methods of gene-discovery, actually originated in the 1920’s with the identification of the ABO and Rh RBC surface antigens, and their roles in determining blood transfusion compatibility (reviewed in McKusick, 2002) (13). Since these antigens, and the WBC cell surface antigens HLA-A, -B, and -C, are encoded by unique nucleotide sequence of genes on chromosomes 9, 1, and 6, they served as proxies of chromosomes transmitted from each parent – the very first chromosomal markers - decades before DNA-based genetic markers were widely available. While limited to just 3 chromosomes, these early markers were used to map the general location (locus) of ankylosing spondylitis (AS in linkage with B27) and hemochromatosis (HFE in linkage with A3/B7) nearby the HLA genes on 6p21.3, phosphoglucomutase 1 (PGM1) on 1p31 (a fair distance from Rh on 1p36.2 -1p34), and the adenylyl-kinase deficiency (AK1) locus close to ABO on 9q34. Perhaps more importantly, they proved the usefulness of genetic linkage analysis in mapping the location of human disease genes and inspired the search for denser marker maps that could be used to localize diseases and traits anywhere on the 22 autosomes and 2 sex chromosomes of the human genome.

The description by Botstein, White, Skolnick, and Davis of restriction fragment length polymorphisms (RFLPs or “riflips”) as linkage markers in family studies, ushered in the modern era in map-based gene-discovery by providing a new and potentially dense source of genetic
Genetics and prematurity

Markers (14). Descriptions of other types of heritable DNA polymorphisms, such as variable number tandem repeats (VNTRs) and CA-dinucleotides soon followed so that by 1996 the number of available genetic markers rose to more than 5,000 (15).

3.1. Positional cloning

Identification of a gene on the X chromosome for chronic granulomatous disease in 1986 heralded a new strategy for identifying disease genes based on chromosomal mapping, later dubbed “positional cloning.” The positional cloning approach (Figure 2A) has generally involved the iterative processes of genetic linkage analysis, frequently abbreviated as ‘linkage,’ and association analysis, frequently abbreviated as ‘association,’ but also known as linkage disequilibrium analysis. The approaches differ in study design, power, and sensitivity.

3.1.1. Genetic Linkage Analysis

Genetic linkage analysis identifies alleles that co-segregate with a disease (or phenotype) within families. Historically, multi-generational families, typified by Huntington’s disease and cystic fibrosis, were used to assign the general location (locus) of a disease gene by genetic linkage analysis (Figure 2A). Collecting DNA samples across multiple generations, however, can be difficult due to dead or missing family members and concerns for confidentiality of personal medical information. Fortunately, genetic linkage information is additive across families, and most contemporary linkage analyses are performed on collections of many nuclear (two-generation) families. The smallest nuclear family is the triad consisting of both biological parents and a single offspring, and by convention the power of a linkage study is described in terms of the required number of triads to achieve statistical significance. In genetic linkage analysis the number of inter-generational transmissions of a genetic marker concurrent with the defined phenotype, BPD for example, is tallied. The likelihood that the marker and phenotype are transmitted to effected offspring to those that are not is described by Pritchard et al (2000) and Devlin and Roeder (1999) and structured association described by Pritchard et al (2000) (16, 17). Genomic control methods use the independent markers (unlinked to the phenotype) to adjust the distribution of a standard test statistic. Structured association methods use the marker loci to infer the details of population substructure and allows for the possibility that the alleles at a test locus may have different effects in

While linkage is more sensitive and can localize a disease gene to between 5 to 10 million basepairs anywhere in the genome with a panel of just 400 to 600 markers, association mapping presumes that the allele and the causative genetic mutation have remained in close proximity on the same chromosomal segment through historical recombinant events. Association is an extraordinarily powerful approach to gene discovery, since depending on the genomic location association intervals are generally small, and frequently less than 100,000 bases. While linkage is more sensitive and can localize a disease gene location to between 5 to 10 million bases anywhere in the genome with a panel of just 400 to 600 markers, association is more precise. However, to achieve this precision across the entire genome of 3 billion bases would require 100,000 to 300,000 markers, which until recently, was a technological and economical restraint. Therefore many gene discovery projects begin with linkage to identify large genomic locations (loci), then attempt to identify the disease gene with association.

3.1.3. Family-Based Control

While genetic association remains a powerful approach to identify disease genes, it is vulnerable to hidden population stratification. This occurs when subjects and controls are not precisely matched and can yield false positive and false negative association. One approach to correct for hidden stratification is the transmission disequilibrium test (TDT). By assessing association only in the presence of linkage, the TDT eliminates confounding due to population stratification. While a test of association, the TDT interrogates families – typically small nuclear families - and compares the frequency of parental alleles transmitted to effected offspring to those that are not transmitted. Other methods of correction include genomic control described by Devlin and Roeder (1999) and structured association described by Pritchard et al (2000) (16, 17). Genomic control methods use the independent markers (unlinked to the phenotype) to adjust the distribution of a standard test statistic. Structured association methods use the marker loci to infer the details of population substructure and allows for the possibility that the alleles at a test locus may have different effects in
Figure 2. Transition from linkage to association to candidate-gene approaches to disease gene-discovery, 1992 to 2005. (Reproduced with permission from Nature Genetics) (30). A. Positional Cloning: original depiction of positional cloning approach from Collins (1992), starting with recruitment of multi-generational families for linkage analysis and single-tube genotyping. This proved a useful paradigm mostly for rare single-gene mendelian-inherited disorders (30). Broad localization to within 5-20 million bases was initially determined by genetic linkage analysis followed by positional refinement to less than 1 million bases by association analysis. Both linkage and association could be supplanted if a disease phenotype could be found to segregate with a gross chromosomal abnormality; for example dystrophin (DMD) was localized by rearrangements of Xp21. Positional cloning preceded the Human Genome Project so that genes had to be identified by tedious physical mapping and transcript identification techniques. Genomic DNA sequencing identified mutations. B. Genome-Wide Association: with the availability of large numbers of genetic markers distributed through the entire genome and concomitant increases in throughput and lowering of genotyping costs, by-passing linkage analysis and starting directly with association is now feasible. Family-based control approaches to association necessitate collections of many small nuclear families, replacing the multi-generational families that are more advantageous for linkage. High throughput genotyping arrays replace single-tube genotyping (top right hand corner). Contemporary physical mapping means looking up available transcript maps on publicly available databases; cloning is no longer necessary. C. Candidate-Gene Association: with the expansion of the human proteome database, many studies skip linkage and association analyses and start at candidate-genes. However, the reproducibility and ultimately the biologic relevance.
Genetics and prematurity

different subpopulations (for example, the test locus might be in strong linkage disequilibrium with a disease mutation in some subpopulations, but not in others).

The official birth date of the U.S. Human Genome Project was on October 1, 1990 (13). James Watson served as director for the first three years and oversaw the assignment of international chromosomal sequencing centers, centers of linkage analysis, gene discovery, and even a center that focused on the ethics of genetic analysis. Ultimately, these efforts culminated in the sequencing of all 3 billion bases comprising the entire human genome, discovery of 35,000 genes, and cataloguing of more than 10 million markers distributed through all 22 autosomes and the X and Y chromosomes (http://www.ncbi.nlm.nih.gov/SNP/snps_summary.cgi). These resources are maintained in databases through the National Center of Biotechnology and Information (NCBI), and freely accessible to anyone with a computer, internet connection, and web browser. In 2002 the International HapMap Project was founded with the goal of providing a resource that would facilitate the design of efficient genome-wide association studies by characterizing patterns of genetic variation and linkage disequilibrium in a sample of 269 subjects from four geographical populations (18, 19). Phase 1 of HapMap, containing more than one million SNPs is now complete, and the data with associated summaries and query-based tools are available online at http://www.hapmap.org. This important resource allows selection of reference panels containing just 250,000 – 500,000 SNPs that reliably capture a high fraction (70 – 80%) of all common variants (those with an allele frequency above 5%) (20). The availability of these dense marker maps, the HapMap, and the concurrent technological advancements in genotyping and incremental drops in genotyping costs, have provided the means for genetic linkage analyses of large family collections, and more recently, fueled the possibility of foregoing general linkage analysis and going straight to genome-wide association studies directly (Figure 2B). While still in early development, genome-wide association has been validated by comparisons to linkage results (21), and have had remarkable success identifying genes for a small number of complex genetic disorders such as Hirschsprung disease (22), myocardial infarction (23), and age-related macular degeneration (21-24).

3.1.4. Candidate-Gene

Completion of the human gene map comprised of just around 35,000 genes, has also spurred candidate-gene approaches to discovery. A candidate-gene approach can by-pass initial localization or fine mapping by linkage and association, by directly suggesting candidates based on function or metabolic pathways (Figure 2C). For example, one hypothesis suggests that BPD and ROP are caused in large part by tissue vulnerability to damage by oxygen free-radicals. One could screen the entire proteome (the database of all gene functions) for genes related to free-radical scavenging, then test each one for mutations or polymorphisms (variations of the gene sequence) in a case-control study design. Historically this approach was constrained by inadequate information on gene function and related pathways, and technical constraints of high-throughput mutation analysis. But as the proteome database has exponentially expanded in numbers and information content over the past decade, and the cost of sequencing has decreased by orders of magnitude, increasingly candidate-gene studies are becoming more popular. Candidate-gene association studies are also highly susceptible to type I errors (false positive association) due to chance, ascertainment bias, and population stratification, which may ultimately limit widespread application (25, 26).

4. STUDY DESIGNS FOR COMPLEX DISORDERS OF THE PREMATURE NEONATE

To date there have been a few dozen studies of genetic factors that contribute to the common disorders of the VLBW neonate. Most noteworthy are the surfactant apoprotein deficiencies that cause respiratory failure. But this small family of diseases act as rare mendelian recessive traits, and as such are inherently powered for studies in few families. Far more challenging will be the complex disorders, IVH and periventricular leukomalacia, BPD, NEC, and ROP, where several genes and environmental factors are likely involved. Published studies of these disorders, have until recently, employed exclusively candidate-gene approaches (reviewed in Shetty et al, in-press) (27). While several achieve p-values <.05, they are typically underpowered and would not meet the minimal criteria of a test for false positive report probability (26).

A recent report by Bhandari et al (2006) presents the first formal genetic heritability study of BPD, NEC, and IVH in VLBW monozygotic and dizygotic twins, showing all three diseases are familial, and that over 50% of the variance in BPD is from genetic factors (28). Hopefully, additional prospective heritability studies will follow. There is sufficient justification now for at least a large-scale collection for genetic linkage analysis, and perhaps genome-wide association. Depending upon estimates of the number of contributing genes, allele frequencies, and allele effects, an efficient cost-effective genetic linkage study design may be adequate to identify susceptibility loci with as few as 1,000 triads (3,000 total subjects). For a genome-wide association study, a sample of 1,000 cases and 1,000 controls (2,000 total subjects) may be adequate to identify associations of moderate genotype relative risk (~1.5) in a variety of realistic settings (29). Microarray-based and other contemporary methods of genotyping require less than a microgram of genomic DNA, which is well within the average yield of buccal swabs (generally 3 – 10 micrograms) and obviating the need for generating expensive cell lines. Large-scale high-throughput genotyping is available now at less than 10 cents per genotype. Given the likelihood that any single subject will express more than one complex phenotype, there would be considerable efficiency in subject collection if more than one disorder were to be collected in the same study (i.e, BPD and IVH). Furthermore, since the frequency of the diseases are well-known and proportional to birthweight and gestational age, the collection could be well-planned and targeted for racial and ethnic distributions, inter-
Genetics and prematurity

institutional differences in care, and socio-economic considerations.

In conclusion, neonatal morbidity and mortality for VLBW neonates have changed little over the last 15 years despite significant advances in health care and investment of healthcare dollars. While empiric studies, general technical advancements, and animal models have until recently improved neonatal outcomes in this gestational age group, further productivity from these traditional sources are not likely. At least one recent study shows that in addition to global immaturity of enzymatic systems, the common problems of prematurity - BPD, NEC, and IVH - have significant genetic components. Incremental improvements in the outcome of VLBW neonates in the future will likely depend on our ability to identify these genetic components and to target specific therapies. We can start this process now by recognizing that the common diseases of prematurity have significant genetic susceptibility components, and by launching robust discovery projects to identify these gene targets. Cost-effective methods and resources, fueled by the Human Genome and HapMap Projects, are available now, and through creative planning and timely implementation will yield useful results.

5. ACKNOWLEDGEMENTS

Support for JRG and HM is from NIH R01 NS43530. Additional support for JRG and from R01 AA11330, R01 DA12690, and R01 DA12849.

6. REFERENCES


**Key Words:** Prematurity, Genetics, VLBW, RDS, BPD, IVH, NEC, ROP, Linkage, Association, Genome, Review

**Send correspondence to:** Dr Jeffrey R. Gruen, Department of Pediatrics, Yale Child Health Research Center, 464 Congress Avenue, New Haven, CT 06520-8081, Tel: 203-737-2202, Fax: 203-737-5972, Email: jeffrey.gruen@yale.edu

http://www.bioscience.org/current/vol12.htm