



Published in final edited form as:

Birth Defects Res. 2019 December 01; 111(20): 1618–1632. doi:10.1002/bdr2.1554.

Exome sequencing of family trios from the National Birth Defects Prevention Study: Tapping into a rich resource of genetic and environmental data

Mary M. Jenkins¹, Lynn M. Almlı^{1,2}, Faith Pangilinan³, Jessica X. Chong⁴, Elizabeth E. Blue⁵, Stuart K. Shapira¹, Janson White⁴, Daniel McGoldrick⁶, Joshua D. Smith⁶, James C. Mullikin³, Christopher J. Bean¹, Wendy N. Nembhard⁷, Xiang-Yang Lou⁸, Gary M. Shaw⁹, Paul A. Romitti¹⁰, Kim Keppler-Noreuil¹¹, Mahsa M. Yazdy¹², Denise M. Kay¹³, Tonia C. Carter¹⁴, Andrew F. Olshan¹⁵, Kristin J. Moore¹⁵, Nanette Nascone-Yoder¹⁶, Richard H. Finnell¹⁷, Philip J. Lupo¹⁸, Marcia L. Feldkamp¹⁹, NISC Comparative Sequencing Program²⁰, The University of Washington Center for Mendelian Genomics²¹, Deborah A. Nickerson⁶, Michael J. Bamshad^{4,6}, Lawrence C. Brody³, Jennita Reefhuis¹, The National Birth Defects Prevention Study

¹National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, Atlanta, Georgia ²Carter Consulting Incorporated, Atlanta, Georgia ³National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland ⁴Department of Pediatrics, University of Washington, Seattle, Washington ⁵Department of Medicine, University of Washington, Seattle, Washington ⁶Department of Genome Sciences, University of Washington, Seattle, Washington ⁷Fay W Boozman College of Public Health, University of Arkansas for Medical Sciences, Little Rock, Arkansas ⁸College of Medicine, University of Arkansas for Medical Sciences, Little Rock, Arkansas ⁹Stanford University School of Medicine, Department of Pediatrics, Stanford, California ¹⁰Department of Epidemiology, University of Iowa, Iowa City, Iowa ¹¹Children's National Medical Center, George Washington University, Washington, District of Columbia ¹²Massachusetts Department of Public Health, Boston, Massachusetts ¹³Division of Genetics, Wadsworth Center, New York State Department of Health, Albany, New York ¹⁴Marshfield Clinic Research Institute, Marshfield, Wisconsin ¹⁵Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, North Carolina ¹⁶Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, North Carolina ¹⁷Center for Precision Environmental Health, Departments of Molecular & Cellular Biology and Medicine, Baylor College of Medicine, Houston, Texas ¹⁸Department of Pediatrics, Section of Hematology-Oncology, Baylor College of Medicine, Houston, Texas ¹⁹Division of Medical Genetics, Department of Pediatrics, University of Utah School of Medicine, Salt Lake City, Utah ²⁰NIH Intramural Sequencing Center, National Human Genome Research Institute, Bethesda, Maryland ²¹University of Washington, Seattle, Washington

Correspondence: Mary M. Jenkins, National Center on Birth Defects and Developmental Disabilities, Centers for Disease Control and Prevention, 4770 Buford Highway, MS S106-3, Atlanta, GA 30341., mqj2@cdc.gov.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

Abstract

Background: The National Birth Defects Prevention Study (NBDPS) is a multi-site, population-based, case-control study of genetic and nongenetic risk factors for major structural birth defects. Eligible women had a pregnancy affected by a birth defect or a liveborn child without a birth defect between 1997 and 2011. They were invited to complete a telephone interview to collect pregnancy exposure data and were mailed buccal cell collection kits to collect specimens from themselves, their child (if living), and their child's father. Over 23,000 families representing more than 30 major structural birth defects provided DNA specimens.

Methods: To evaluate their utility for exome sequencing (ES), specimens from 20 children with colonic atresia were studied. Evaluations were conducted on specimens collected using cytobrushes stored and transported in open versus closed packaging, on native genomic DNA (gDNA) versus whole genome amplified (WGA) products and on a library preparation protocol adapted to low amounts of DNA.

Results: The DNA extracted from brushes in open packaging yielded higher quality sequence data than DNA from brushes in closed packaging. Quality metrics of sequenced gDNA were consistently higher than metrics from corresponding WGA products and were consistently high when using a low input protocol.

Conclusions: This proof-of-principle study established conditions under which ES can be applied to NBDPS specimens. Successful sequencing of exomes from well-characterized NBDPS families indicated that this unique collection can be used to investigate the roles of genetic variation and gene-environment interaction effects in birth defect etiologies, providing a valuable resource for birth defect researchers.

Keywords

birth defects; gene-environment interaction; genetics; intestinal atresia; sequence analysis

1 | INTRODUCTION

Overall, birth defects are common, affecting approximately one in 33 babies born in the United States each year (Centers for Disease Control and Prevention, 2008); however, individual types of birth defects are relatively rare. Birth defects are the leading cause of infant death in the United States (Kochanek, Murphy, Xu, & Arias, 2017) and contribute substantially to morbidity and disability. The estimated annual cost of birth defect-associated hospitalizations in the United States in 2013 was \$22.9 billion (Arth et al., 2017). This economic burden is accompanied by substantial stress and disruption of family life. Despite its public health significance, the causes of most birth defects remain unknown, although there is evidence for multifactorial etiology possibly involving many genetic risk factors and environmental, lifestyle, and demographic factors (Feldkamp, Carey, Byrne, Krikov, & Botto, 2017; Krauss & Hong, 2016; Maslen, 2018).

Several population-based studies have evaluated genetic and nongenetic risk factors for birth defects in other countries (e.g., the Norwegian mother and child cohort study [Magnus et al., 2016]; the Danish National Birth Cohort [Olsen et al., 2001]; and the European

Collaboration on Craniofacial Anomalies [Mossey et al., 2017]). However, in the United States, the National Birth Defects Prevention Study (NBDPS) is one of the only multistate population-based birth defect risk factor studies that collected both nongenetic exposure data and biological specimens (Reefhuis et al., 2015; Yoon et al., 2001). Other studies have assessed either nongenetic pregnancy exposures (e.g., Klebanoff, 2009; Louik, Lin, Werler, Hernandez-Diaz, & Mitchell, 2007; Mitchell, 1988) or genetic risk factors (e.g., Gelb et al., 2013; Melvin et al., 2000; Yu et al., 2012), but few have had the ability to combine both. Data and specimens collected as part of the NBDPS provide unique opportunities to identify gene–environment interaction effects that could elucidate risk factors for birth defects.

To date, NBDPS genetic association or gene–environment interaction studies have largely focused on testing candidate genes in multiple pathways (e.g., xenobiotic metabolism, one-carbon metabolism, and DNA synthesis and repair; Cleves et al., 2011; Hoang et al., 2019; Hobbs et al., 2014; Jenkins et al., 2014; Lupo, Canfield, et al., 2012; Lupo, Chapa, et al., 2012; Nembhard et al., 2017; Schmidt et al., 2010; Tang, Cleves, et al., 2015; Tang, Hobbs, et al., 2015). In contrast to these candidate gene approaches, we are poised to examine genetic variation using massively parallel DNA sequencing technologies. Here, we describe the optimization of sequencing methods using archived NBDPS specimens and demonstrate the feasibility of using these optimized methods for variant discovery and analysis.

1.1 | NBDPS population

The NBDPS was a population-based case–control study of risk factors for birth defects conducted by 10 Centers for Birth Defects Research and Prevention (CBDRP; Reefhuis et al., 2015; Yoon et al., 2001). The study included women with pregnancies affected by one of more than 30 major structural birth defects and mothers of liveborn children without major structural birth defects. A woman was eligible for inclusion if her pregnancy ended on or after October 1, 1997, and if her estimated delivery date was on or before December 31, 2011. Cases were ascertained from existing birth defect surveillance systems in each of 10 states (Arkansas, California, Georgia, Iowa, Massachusetts, New Jersey, New York, North Carolina, Texas, or Utah) and could be liveborn, stillborn, or pregnancy terminations. Liveborn control children were ascertained from vital records or birth hospital logs in the same geographic areas as cases. Clinical geneticists reviewed medical records using standardized case definitions to determine eligibility (Rasmussen et al., 2003). Cases with known syndromes, chromosomal or single-gene disorders were excluded.

After obtaining verbal consent, computer-assisted telephone interviews were conducted in English or Spanish with over 40,000 eligible women between 6 weeks and 24 months after their estimated dates of delivery. During the maternal interview, data were collected on pregnancy history, family history, and sociodemographic factors and on the following exposures from 1 year before conception through the end of pregnancy: maternal chronic diseases; infections; fever; use of medications, illicit drugs, tobacco, caffeine, alcohol, and vitamins; diet and nutrition; weight; physical activity; injuries; stress; occupations; drinking water use; occupational and environmental exposures; and other factors.

Following the completed interview, each woman was mailed cytobrushes (two per participant) to collect buccal cell specimens from herself, her child (if living), and her

child's biological father (if available). DNA specimens from more than 23,000 families were collected to evaluate genetic risk factors and to examine the role that genetic variants might play in modifying risks associated with environmental factors (Rasmussen et al., 2002; Reefhuis et al., 2015). Between 1997 and 2003, cytobrushes were mailed back to Centers in closed plastic tubes that did not allow air drying (hereafter referred to as "wet brushes" [Cyto-Pak Cytosoft Brushes CP-5B, Medical Packaging Corporation, Camarillo, CA]). In 2003, materials and procedures were modified. Specifically, after sampling, cytobrushes were instead packaged in open paper-backed peel pouches (hereafter referred to as "dry brushes" [Cytology Brush Pack CYB-1, Medical Packaging Corporation]; Gallagher et al., 2011). This new packaging mitigated bacterial and fungal growth by allowing specimens to dry during transport. Among the cytobrushes returned during the course of the study, 22% were wet brushes and 78% were dry brushes.

Institutional review boards at the Centers for Disease Control and Prevention (CDC) and each CDBRP approved the NBDPS protocol.

1.2 | Collaborative working group

With the rapid advances in technologies used to perform high-throughput DNA sequencing and the accompanying reduction in sequencing costs, NBDPS investigators convened a working group in 2014 to determine the best ways to leverage NBDPS data and specimens with sequencing technologies so as to improve the efficiency and likelihood of discovery of genetic risk factors and gene–environment interaction effects for birth defects. The working group was tasked with making recommendations about the technologies and study designs and how to prioritize birth defects for sequencing.

The collaborative working group recommended exome sequencing (ES) instead of whole genome sequencing (WGS), because although both ES and WGS have comparable sensitivity to detect single nucleotide variants in coding regions, the substantially lower cost of ES at present allows more specimens to be sequenced and hence, increases statistical power for analyses. Full trios (mother, father, and child) were prioritized, a study design that enables identification of de novo variants causing disease (Chesi et al., 2013; Hunt et al., 2014). The trio study design also reduces the impact of population admixture in the assessment of inherited variants and increases flexibility of analyses, allowing investigators to conduct hybrid log-linear analyses (Weinberg, Wilcox, & Lie, 1998); transmission disequilibrium tests (Spielman, McGinnis, & Ewens, 1993); analyses of maternal genetic effects, maternal-fetal gene interaction effects, and parent-of-origin effects (Ainsworth, Unwin, Jamison, & Cordell, 2011); analyses of gene–environment interaction effects (Tai et al., 2015; Umbach & Weinberg, 2000); and other study designs. A case-only study design was chosen to maximize the number of specimens that could be included with limited funds and because use of publicly available controls and unaffected parents as controls were options for association testing. Family trios reporting use of egg, embryo, or sperm donors were excluded.

1.3 | Case groups

By design, NBDPS included some very rare birth defects, which were prioritized for the ES projects for several reasons. First, the specimens had not been used previously as power was low for candidate gene-based association studies. Thus, DNA in quantities sufficient for ES and follow-up studies were available. Second, methods used to analyze these trios would be similar to those used to analyze families with Mendelian traits, and ES has been used successfully to uncover the molecular basis of multiple Mendelian disorders (Bamshad et al., 2011). Third, this provided an opportunity to have an impact on families with less common phenotypes.

Birth defects that were prioritized for this initial round of ES included the following: colonic atresia or stenosis, anterior segment dysgenesis eye defects, primary congenital glaucoma, transverse limb reduction defects, split hand/foot malformation, cloacal exstrophy, bladder exstrophy, anophthalmos or microphthalmos, sacral agenesis, biliary atresia, tricuspid atresia, Ebstein anomaly, hypoplastic left heart syndrome, and heterotaxy (Table 1). Each case child's medical records were reviewed by clinical geneticists and classified as having (a) one major structural birth defect or a complex sequence of birth defects that occurs in a well-defined pattern (isolated) or (b) more than one major unrelated birth defect in separate organ systems or a complex sequence occurring in a well-defined pattern with one or more unassociated major birth defects (multiple). A clinician trained in pediatric cardiology also classified each congenital heart defect (CHD) as "simple" (a well-defined specific defect), "associations" (a common, uncomplicated combination of two or three cardiac defects that often occur together), or "complex" (a combination or pattern of independent defects in multiple cardiac structures; Botto, Lin, Riehle-Colarusso, Malik, & Correa, 2007; Rasmussen et al., 2003). Each child with a CHD was classified based on both cardiac and noncardiac defects. The detailed classification protocol used by NBDPS clinicians allows for stratified analyses well-suited to identify potential causal variants (i.e., precise phenotypes reduces etiologic heterogeneity and allows assessment of children by defect severity). Results from children with more than one of the included birth defects will be analyzed with each case group separately.

2 | METHOD OPTIMIZATION

2.1 | Specimen type and library preparation protocol

As there is currently no "standard" ES protocol, and NBDPS specimens were collected up to two decades ago using less than optimal methods, we first sought to optimize the method for sequencing exomes from self-collected cytobrush-derived DNA specimens. This method optimization would allow the development of a workflow that could be used across phenotypes, thereby ensuring consistency. Pilot studies were conducted using 9 wet brush-derived and 51 dry brush-derived DNA specimens from 20 children with rare intestinal anomalies, colonic atresia or stenosis, and their parents (60 specimens total). The 20 case children were ascertained from surveillance systems in 8 of 10 Centers (New Jersey and New York were not represented).

DNA extraction from wet brushes was completed at each Center's laboratory using several protocols (Rasmussen et al., 2002). DNA extracted from one wet brush per participant was mailed to the NBDPS biorepository; DNA extracted from the second wet brush per participant was retained at the Center's laboratory. DNA extraction from dry brushes was modified such that the two brushes collected per participant were processed separately (Gallagher et al., 2011). DNA extraction from one dry brush per participant continued to be completed at each Center's laboratory using several protocols and was retained locally. DNA extraction from the second dry brush per participant was completed at a CDC laboratory using Puregene (Gentra, Germantown, MD) or a phenol chloroform method (Garcia-Closas et al., 2001), modified by separating the aqueous and organic phases using Phase Lock Gel barrier tubes according to the manufacturer's instructions (Fisher Scientific, Pittsburgh, PA); the resulting DNA was mailed to the NBDPS bio-repository. DNA specimens included in the ES projects were stored at the NBDPS biorepository (1997–2011: CDC and ATSDR Specimen Packaging, Inventory and Repository in Lawrenceville, GA; 2011-present: Fisher BioServices, Rockville, MD).

DNA specimens were shipped from the NBDPS bio-repository to a National Human Genome Research Institute laboratory (Dr. Lawrence Brody, Bethesda, MD) for processing and initial quality control (QC). Modest DNA yields prompted an evaluation of whole genome amplification (WGA) of these archived specimens and use of WGA product as input for ES. Genomic DNA (gDNA, 50 ng) from all 60 specimens was shipped to QIAGEN (Germantown, MD) for WGA using REPLI-g[®] WGA reagents. The QIAGEN lab reported that 56% (5 of 9) of the wet brush specimens and 100% (51 of 51) of the dry brush specimens were successfully amplified.

Fifty-six of 60 specimens passed QC at QIAGEN and were sequenced at the National Institutes of Health (NIH) Intramural Sequencing Center (NISC, Rockville, MD; <https://www.nisc.nih.gov/>). Briefly, libraries were prepared from 1 µg of each WGA product using the KAPA DNA Library Preparation Kit (KAPA Biosystems, Boston, MA). A 25-base pair (bp) sequencing run was completed to assess unique alignment to the human reference sequence. Four of the five wet brush specimens that passed initial QC yielded 40% or fewer reads that aligned uniquely to the human reference sequence (range 0.5–39.3%). These results suggested that these specimens were contaminated with high levels of nonhuman DNA. Among the 51 dry brush specimens, all WGA products sequenced were “usable” but had limitations, such as low library diversity (a measure of unique reads), high mismatch rate in comparison to the human reference, or poor concordance based on mapping the paired reads from the ends of the sequenced WGA fragment to the reference genome (i.e., the mapped distance between the two reads was not within the expectations for the size range of the fragment).

Exomes were captured using SeqCap EZ Human Exome + UTR kit v3.0 (Roche NimbleGen, Madison, WI), and sequencing was completed on Illumina's HiSeq 2500 system (Illumina, San Diego, CA), which covered 96 Mb and employed 126-bp paired-end read sequencing. Image analysis and base calling were performed using Illumina Genome Analyzer Pipeline software (version 1.18.64.0) with default parameters.

Sequence reads were mapped to hg19 using Illumina's efficient large-scale alignment of nucleotide databases for machine QC and were realigned using NovoAlign. Binary alignment map (BAM) files were generated. Genotypes were called using a probabilistic Bayesian algorithm, most probable genotype (MPG; Teer et al., 2010), and annotated using ANNOVAR (Wang, Li, & Hakonarson, 2010). An MPG score ≥ 10 for at least 85% of the target exome sequence was considered minimum coverage (an MPG score of 10 corresponds roughly with $10\times$ to $20\times$ depth of coverage), and bases with PHRED quality scores <20 were excluded. These metrics were applied only to evaluate unamplified gDNA versus WGA product.

To determine the quality of ES data from WGA products, 1 μg of unamplified gDNA from 10 of the 56 specimens was sequenced, and each pair of results derived from the same specimen was compared (Table 2). Overall, dry-brush-derived DNA specimens yielded higher quality DNA and performed better than wet-brush-derived DNA specimens. Additionally, ES quality metrics were better when input was native gDNA versus WGA product. Although some WGA products performed reasonably well when compared to unamplified DNA (specimens 7–10), there was no correlation between gDNA yield and WGA product performance; however, the percent of precapture reads that aligned to human reference sequence correlated with mean depth of coverage and MPG scores. Based on these results, the decision was made to move forward using unamplified gDNA from dry brush specimens as input for ES.

Given the low quantity of gDNA present in these specimens, a low input library preparation protocol (Accel-NGS[®] 2S Plus DNA Library Kit, Swift BioSciences, Ann Arbor, MI) was evaluated. This method was applied to 100 ng gDNA from six dry-brush-derived DNA specimens of children with colonic atresia or their parents and six replicates of HapMap control specimen NA12878 (Coriell Institute, Camden, NJ). All other steps were performed as described above. The percent of on-target bases with MPG scores ≥ 10 ranged from 95.9 to 96.4% for high-quality control replicates and 93.7 to 96.7% for NBDPS specimens. Total variants observed compared to the reference genome ranged from 86,243 to 87,583 for controls and 83,269 to 89,264 for NBDPS specimens. Average depth of coverage in the target region ranged from $58.2\times$ to $67.0\times$ for controls and $38.1\times$ to $65.7\times$ for NBDPS specimens. Based on the similarity of results from high-quality control replicates and NBDPS specimens using the low input protocol, as well as their similarity to results from 10 unamplified gDNA specimens sequenced using the standard library preparation protocol (requiring 1 μg DNA), all specimens derived from NBDPS dry brush trios and included in one of the ES projects were sequenced using unamplified gDNA and a low input library preparation protocol.

2.2 | Proof-of-principle study

Following method optimization, we used specimens from children with colonic atresia and their parents to complete a proof-of-principle study to demonstrate the feasibility of using the optimized methods for variant discovery and analysis.

The prevalence of colonic atresia is estimated at 1 in 20,000 (Mirza, Iqbal, & Ijaz, 2012) to 1 in 66,000 live births (Davenport, Bianchi, Doig, & Gough, 1990). Approximately one-half of

children with colonic atresia have additional defects (El-Asmar, Abdel-Latif, El-Kassaby, Soliman, & El-Behery, 2016; Etensel et al., 2005). If isolated colonic atresia is identified and treated early, the prognosis is good with an overall mortality of <26%; however, mortality increases considerably if there is a delay in treatment beyond 3 days (Etensel et al., 2005). Colonic atresia cases are classified according to the continuity of the bowel and mesentery (Bland-Sutton, 1889) into three types (Types I, II, and III, in order of severity from least to most severe) and a catch-all category of “not otherwise specified.” It is estimated that 75% of colonic atresias are proximal to the splenic flexure in the ascending colon (Winters, Weinberger, & Hatch, 1992). Furthermore, most cases have Type III (Powell & Raffensperger, 1982), characterized by a “V” shape in the mesentery with proximal and distal blind ends, which can have more postoperative repair complications compared to the other types of colonic atresia.

The etiology of colonic atresia is unknown. The prevailing theory involves a vascular insult (accident) to the mesenteric vessels during development (Louw & Barnard, 1955). Another theory is genetic; there is one report of a familial occurrence of colonic atresia with three isolated cases among first-degree relatives (Benawra, Puppala, Mangurten, Booth, & Bassuk, 1981). To our knowledge, there are no published genetic studies that include isolated human colonic atresia cases; however, mouse models suggest three putative candidate genes for human colonic atresia: fibro-blast growth factor 10 (*Fgf10*) and its receptor (*Fgfr*; Fairbanks et al., 2005) and *Cdx2* (Gao, White, & Kaestner, 2009). Most reports of colonic atresia are case/surgical reports so environmental exposure and pedigree information are sparse, although several studies have reported cases of colonic atresia associated with intrauterine varicella infection (Alexander, 1979; Hitchcock, Birthistle, Carrington, Calvert, & Holmes, 1995; Sauve & Leung, 2003). Using NBDPS data, several other exposures have been associated with colonic atresia, including periconceptual cold/flu with fever (Waller et al., 2018) and periconceptual genitourinary infections (Howley et al., 2018).

Following the methods outlined in the Appendix, BAM files from nine trios (nine probands and their unaffected parents) with sufficient dry-brush-derived DNA that were successfully sequenced at NISC were transferred to the University of Washington Center for Mendelian Genomics (UW-CMG) for reprocessing and analysis. Freemix was used to estimate contamination levels of specimens (Jun et al., 2012). One trio was removed for evidence of contamination (i.e., a proband with freemix estimate >6%), resulting in eight useable trios. Using GEMINI v0.20.2 (Paila, Chapman, Kirchner, & Quinlan, 2013), variant filters included a depth 6, genotype quality 20, allele frequency 0.005 across populations represented in reference databases (1,000 Genomes phase 3 [1000 Genomes Project Consortium et al., 2015], Exome Aggregation Consortium [ExAC v0.3; Lek et al., 2016], Exome Sequencing Project 6,200 [v2; NHLBI Exome Sequencing Project, n.d.], or the UK10K February 15, 2016 release ([UK10K Consortium et al., 2015]), and a predicted medium to high impact severity on the gene/protein (GEMINI, 2017). These filters yielded 156,716 total single nucleotide polymorphisms (21,570 singletons) with a 2.4 transition/transversion (Ti/Tv) ratio and 20,374 indels (3,921 singletons) for the eight full trios. Summary statistics for each individual are provided in Table 3.

Given that little is known about genetic risk for colonic atresia, two independent analyses were conducted. The first analysis considered all Mendelian inheritance models, except for autosomal dominant. After variant filtration, no gene had prioritized variants from more than one family, and the following plausible modes of inheritance were represented: de novo, compound heterozygous, X-linked de novo, and X-linked recessive. After excluding variants with evidence of mismapping or sequencing errors, 31 genes were identified with possible pathogenic variants for the eight trios and are shown in the Supporting Information Table S1 by modes of inheritance. Each variant was visualized in Integrative Genomics Viewer (Robinson et al., 2011) and interrogated for predicted functional impact, conservation of alignment position, low frequency in control populations, and biological plausibility. There were 13 predicted protein-truncating or loss/gain of function variants (e.g., splice site, stop-gain, frameshift, and in-frame indel) and an additional 3 potentially pathogenic variants (>15 PHRED-scaled Combined Annotation Dependent Depletion [CADD] score [Kircher et al., 2014]; these 16 variants are bolded in the Supporting Information Table S1).

For the next analysis, we assessed all variants in the three candidate genes implicated in colonic atresia from animal studies: *FGF10*, *FGFR2*, and *CDX2*. Prior to filtering for predicted pathogenic coding variants, we identified 2 variants in *FGF10*, 30 in *FGFR2*, and 6 in *CDX2*. All variants had PHRED-scaled CADD scores ≤ 15 , indicating little evidence of conservation or predicted pathogenicity (data not shown). Moreover, the variants were not very rare (population frequency > 0.005) and some occurred in many parents and probands, which suggest that they are unlikely to be causal for the rare defect of colonic atresia. Further follow-up was not conducted.

Given the paucity of genetic studies on colonic atresia, the small number of trios analyzed, and that no genes had rare candidate variants across multiple families, we were unable to make definitive conclusions on the likely mode of inheritance or potential risk variants in the NBDPS probands with colonic atresia who could be studied. One variant in *SCARA3* was submitted to Matchmaker Exchange (Philippakis et al., 2015) because of its high PHRED-scaled CADD score (23.8) and biological plausibility. Given the sporadic occurrence, colonic atresia could result from environmental effects, genetic effects, or a combination of genetic and environmental effects; mosaicism; or possible somatic mutations.

2.3 | Processing of additional case groups

Following the proof-of-principle study, exomes of case children with one of nine other birth defects and their parents were selected for ES at NISC using the optimized methods (i.e., dry-brush-derived gDNA and a low input library preparation protocol). The additional case groups included the following: anterior segment dysgenesis eye defects, primary congenital glaucoma, transverse limb reduction defects, split hand/foot malformation, cloacal exstrophy, bladder exstrophy, anophthalmos or microphthalmos, sacral agenesis, and biliary atresia. UW-CMG sequenced exomes of selected CHD trios (tricuspid atresia, Ebstein anomaly, hypoplastic left heart syndrome, and heterotaxy with and without CHDs) using dry-brush-derived gDNA and a low input library preparation protocol optimized in their laboratory (ThruPLEX DNA-seq Kit, Rubicon Genomics, Ann Arbor, MI).

As in the proof-of-principle study, BAM files from each case group sequenced at NISC were transferred to UWCMG so that each case group was processed separately using the same pipeline (details in Appendix). UW-CMG used peddy [Pedersen & Quinlan, 2017] to check sex, ancestry (using Principal Components Analysis), and pedigrees/relationships and annotated the variant call format files with the ENSEMBL Variant Effect Predictor (v89; McLaren et al., 2016). A summary variant report was prepared for each case group. This report included a list of genes identified using variant filtration in GEMINI (Paila et al., 2013) under each mode of inheritance (homozygous recessive, compound heterozygous, de novo, X-linked recessive, and X-linked de novo) except autosomal dominant, in multiple families and for each family. In addition, UWCMG provided a report for each case group describing copy number variants (CNVs) identified using CoNIFER (Krumm et al., 2012). Upon project completion, these data will be shared broadly in public repositories; as examples, aggregate variant and broad phenotype data will be shared through dbGaP and Geno2MP (Chong et al., 2015); likely pathogenic and pathogenic variants through ClinVar (Landrum et al., 2014); and candidate genes with the MatchMaker Exchange (Philippakis et al., 2015) via MyGene2 (Chong et al., 2016); and the CMG website.

When enough specimens are available per defect, rare variant association testing, such as burden and kernel-based testing, will be conducted within ancestry groups. Rare variants will be validated by Sanger sequencing and potentially included in functional studies. Additionally, the rich environmental exposure data collected from NBDPS participants can be mined to assess exposures that might modify genetic effects, although small numbers limit the robustness of such an assessment for some defects. We plan to publish all results, including negative findings, as for some of these phenotypes, these might be the only current exome-sequenced cohorts with numbers large enough to conduct these analyses, making the results important to include in the peer-reviewed literature.

3 | DISCUSSION

We optimized ES methods for NBDPS specimens and completed a proof-of-principle study using DNA specimens from families with children affected by colonic atresia. Use of unamplified dry-brush-derived gDNA as input for ES yielded the best quality metrics, and sample sizes were enhanced by employing a low input library preparation protocol. We successfully exome sequenced archived DNA specimens collected up to two decades ago and demonstrated the feasibility of using these methods for discovery of potentially pathogenic variants. Using these optimized methods, ES was completed, data were processed, and initial summary reports were created for all selected case groups. These data were shared with NBDPS investigators who focus on specific birth defects. These lead investigators will manage all aspects of their respective defect-specific analyses, including replication and possible functional analyses. Lead investigators and collaborators will be encouraged to combine the rich exposure data collected from NBDPS participants through the computer-assisted telephone interview with exome data to assess potential gene–environment interaction effects.

During a joint NIH/CDC workshop, *Opportunities and Public Health Priorities for Genetics Research on Birth Defects of Complex Etiology*, the potential of the NBDPS to be used to

investigate genetic factors as well as gene-environment interaction effects was highlighted (Olshan, Hobbs, & Shaw, 2011). Attendees discussed promising approaches for studying genetic risk factors for birth defects, including ES. Attendees suggested that NBDPS could not carry out this work in isolation. The rarity of the conditions under study and the need for replication requires that multiple research groups join forces. Other groups have acknowledged the need for more large-scale and interdisciplinary efforts to characterize genetic susceptibility to birth defects (Khokha, Mitchell, & Wallingford, 2017), and the implementation of WGS of selected structural birth defect cohorts through the Gabriella Miller Kids First Pediatric Research Program (<https://commonfund.nih.gov/KidsFirst>) provides another opportunity to realize this need. ES data from the selected birth defects included in the NBDPS will be an important and complementary source of genomic data.

3.1 | Challenges and opportunities

NBDPS sequencing projects face several challenges. First, the modest DNA amounts can be of suboptimal quality from self-collected buccal cells (Gallagher et al., 2011), potentially leading to false negative results. Second, collection and storage of specimens for up to two decades can negatively affect DNA stability and integrity (Madisen, Hoar, Holroyd, Crisp, & Hodes, 1987; Visvikis, Schlenck, & Maurice, 1998). Third, sample sizes are modest due to the rarity of the outcome under study, limiting power for analysis of gene-environment interaction effects. Other analytic challenges remain, such as distinguishing noise from signal and assessing background levels of de novo variants.

However, the NBDPS does provide a unique resource for genomic studies of birth defects with several strengths. NBDPS specimens are ethnically diverse: over 35% of participants self-identified as a race or ethnicity other than non-Hispanic White. This is in contrast to the majority of publicly available genomic reference data that are from populations of European origin (i.e., non-Hispanic White). Because of our family trio study design, we will be able to use exome data from parents of all race-ethnicities who have no family history of the specific birth defect under study as controls. Analysis of exomes of all race-ethnicities will allow improved understanding of risk factors for these defects and possible preventive measures for all populations.

Our strategy is to focus initially on identifying highly penetrant de novo variants and other candidate variants under all possible Mendelian models of inheritance; however, many of these birth defects will likely be the result of multiple moderate effect variants working together (i.e., oligogenic model). Due to the collaborative nature of these projects, we will be able to assess potential pleiotropy across case groups. As we will have limited opportunity to identify more modest effect variants with our sample sizes, we seek opportunities to collaborate with investigators worldwide who have specimens from families with these defects. We are also poised to expand beyond our initial set of phenotypes.

The NBDPS sequencing projects exemplify collaborative efforts between Centers in multiple regions of the United States. This ensures that diverse populations are ascertained. The study is characterized by collaboration between government agencies and academia, as well as between interdisciplinary scientists (i.e., epidemiologists, geneticists, clinicians, and biostatisticians). The combination of data available from these NBDPS trios (e.g., pregnancy

exposure information, genetic data, birth defects clinical information, and family histories) is a unique resource to help advance our understanding of the biological pathways involved in fetal development and to elucidate possible gene–environment interaction effects associated with birth defect risks. Such advances could inform effective birth defect prevention strategies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We would like to thank the many families who completed interviews and provided biologic specimens for the NBDPS. We also thank the many NBDPS scientists and staff, especially members of the Genetics Collaborative Working Group. Additionally, the provision of birth defects surveillance data from the following public health programs was greatly appreciated: Arkansas Department of Health; California Department of Public Health Maternal Child and Adolescent Health Division; Georgia Department of Public Health and the Metropolitan Atlanta Congenital Defects Program; Iowa Department of Public Health (Iowa Registry for Congenital and Inherited Disorders); Massachusetts Department of Public Health; North Carolina Department of Health and Human Services; New Jersey Department of Health; New York State Department of Health (Congenital Malformations Registry); Texas Department of State Health Services (Birth Defects Epidemiology and Surveillance Branch); and Utah Department of Health (Utah Birth Defect Network). This project was supported through CDC cooperative agreements under PA 96043, PA 02081, FOA DD09–001, FOA DD13–003, and NOFO DD18–001 to the CBDRP participating in the NBDPS and/or the Birth Defects Study To Evaluate Pregnancy exposureS (BD-STEPS). Data processing and exome sequencing of congenital heart defects were provided by UW-CMG and were funded by National Human Genome Research Institute (NHGRI) and National Heart, Lung and Blood Institute grants UM1 HG006493 and U24 HG008956. This work was also supported by the NIH’s Division of Intramural Research of the NHGRI. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the CDC, the NIH, or the California Department of Public Health.

Funding information

Centers for Disease Control and Prevention, Grant/Award Numbers: PA 96043, PA 02081, FOA DD09–001, FOA DD13–003, NOFO DD18–001; National Human Genome Research Institute, Grant/Award Numbers: U24 HG008956, UM1 HG006493; National Heart, Lung, and Blood Institute; NIH’s Division of Intramural Research of the NHGRI

APPENDIX

EXOME DATA PROCESSING AND ANALYSIS AT THE UW-CMG

After ES at NISC, BAM files were sent to UW-CMG for reprocessing. BAM files were aligned to human reference (hg19hs37d5) using BWA-MEM (Burrows-Wheeler Aligner; v0.7.10; Li & Durbin, 2009). Read data from a flow-cell lane were treated independently for alignment and QC purposes in instances where merging of data from multiple lanes was required (e.g., for DNA sample multiplexing). Read-pairs not mapping within $\pm 2 SD$ of the average library size ($\sim 150 \pm 15$ bp for exomes) were removed. All aligned read data were subject to the following steps: (a) “duplicate removal” (Picard MarkDuplicates; v1.111); (b) indel realignment (GATK IndelRealigner; v3.2–2); and (c) base quality recalibration (GATK BaseRecalibrator; v3.2–2). Variant detection and genotyping were performed using the HaplotypeCaller tool from GATK (v3.2). Following GATK best practices, variant quality score recalibration was performed. Variants flagged as low quality or potential false positives (quality score ≤ 50 , long homopolymer run >4 , quality by depth < 5 , or within a cluster of SNPs) were excluded from further analysis. Specimen relationships and ancestry

were verified using peddy (v0.2.9; Pedersen & Quinlan, 2017), whereas evidence for contamination was measured using freemix estimates from verifyBamID (Jun et al., 2012). Specimens with contamination rates $\leq 3\%$ were assessed prior to inclusion in further analyses; variants identified in these families will only be considered as candidates after validation in an independent aliquot.

Variant filtration was conducted under standard Mendelian inheritance models with the following parameters using GEMINI 0.20.2 (Paila et al., 2013). Variant filters included depth ≥ 6 , genotype quality ≥ 20 , and alternate allele frequency ≥ 0.005 across populations represented in reference databases (1000 Genomes phase 3; 1000 Genomes Project Consortium et al., 2015), Exome Aggregation Consortium (ExAC v0.3; Lek et al., 2016), Exome Sequencing Project 6200 (v2; NHLBI Exome Sequencing Project, n.d.), or the UK10K February 15, 2016 release (UK10K Consortium et al., 2015), and a predicted medium to high impact severity on the gene/protein (GEMINI, 2017).

CNV DISCOVERY

CNVs were called using CoNIFER v0.2.2 (Krumm et al., 2012). Raw CNV calls were filtered to exclude those primarily in duplicated or repetitive regions of the genome (using a 50% reciprocal overlap mask for segmental duplications and nondiploid genomic regions), as well as for duplicated processed pseudogenes. Calls with low signal strength (dependent on the size of the call) were filtered to reduce the number of false positives while still retaining high sensitivity (absolute SVD-ZRPKM cutoff values: ≥ 1.5 for 2 exon calls, ≥ 1.0 for 3–5 exon calls, and ≥ 0.5 for calls with >5 exons).

VARIANT PRIORITIZATION AND FOLLOW-UP

Although these steps might have varied depending upon the birth defect under study, in general, variants were prioritized based on predicted functional impact, predicted pathogenicity (e.g., with CADD; Kircher et al., 2014), rarity in population control databases, such as ExAC or gnomAD, and biological plausibility. The functional impact of variants was estimated by the type of variant (protein-truncating or loss/gain of function variants, including splice site, stop-gain, frameshift, and in-frame insertions/deletions) or from in silico tools, such as PolyPhen-2 (Adzhubei et al., 2010), SIFT (Vaser, Adusumalli, Leng, Sikic, & Ng, 2016), and CADD (Kircher et al., 2014). Criteria for prioritization of variants included PHRED-scaled CADD >15 . Variant interpretation was facilitated by consulting ClinVar, locus-specific databases, and the literature. Genes were also prioritized if human and/or animal model data aligned with the phenotype.

Web Resources

The URLs for data presented herein are as follows:

Burrows-Wheeler Aligner, <http://bio-bwa.sourceforge.net/>

CADD, <http://cadd.gs.washington.edu/>

ClinVar, <http://www.ncbi.nlm.nih.gov/clinvar/>

CoNIFER, <http://conifer.sourceforge.net/>
 dbGAP, <http://www.ncbi.nlm.nih.gov/gap>
 Ensembl, <http://www.ensembl.org/index.html>
 ExAC, <http://exac.broadinstitute.org/>
 gnomAD, <http://gnomad.broadinstitute.org/>
 GATK, <https://software.broadinstitute.org/gatk/>
 GEMINI, <https://gemini.readthedocs.io/en/latest/>
 MyGene2, <https://mygene2.org/>
 Geno2MP, <http://geno2mp.gs.washington.edu/>
 MPG, <https://research.nhgri.nih.gov/software/bam2mpg/index.shtml>
 NHLBI Exome Sequencing Project, <http://evs.gs.washington.edu/EVS/>
 peddy, <http://annovar.openbioinformatics.org/en/latest/>
 Picard, <https://broadinstitute.github.io/picard/javadoc/picard/overview-summary.html>
 SAMtools, <http://www.htslib.org/>
 1000 genomes, <https://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/>
 UK10K, <https://www.uk10k.org/>
 UW-CMG, <http://uwcmg.org/#/>

REFERENCES

- 1000 Genomes Project Consortium, Auton A., Brooks LD., Durbin RM., Garrison EP., Kang HM., ... Abecasis GR. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. 10.1038/nature15393 [PubMed: 26432245]
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, ... Sunyaev SR (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, 7(4), 248–249. 10.1038/nmeth0410-248 [PubMed: 20354512]
- Ainsworth HF, Unwin J, Jamison DL, & Cordell HJ (2011). Investigation of maternal effects, maternal-fetal interactions and parent-of-origin effects (imprinting), using mothers and their offspring. *Genetic Epidemiology*, 35(1), 19–45. 10.1002/gepi.20547 [PubMed: 21181895]
- Alexander I (1979). Congenital varicella. *British Medical Journal*, 2 (6197), 1074.
- Arth AC, Tinker SC, Simeone RM, Ailes EC, Cragan JD, & Grosse SD (2017). Inpatient hospitalization costs associated with birth defects among persons of all ages—United States, 2013. *MMWR. Morbidity and Mortality Weekly Report*, 66(2), 41–46. 10.15585/mmwr.mm6602a1 [PubMed: 28103210]
- Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, & Shendure J (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews. Genetics*, 12(11), 745–755. 10.1038/nrg3031
- Benawra R, Puppala BL, Mangurten HH, Booth C, & Bassuk A (1981). Familial occurrence of congenital colonic atresia. *The Journal of Pediatrics*, 99(3), 435–436. [PubMed: 7264804]
- Bland-Sutton JD (1889). Imperforate ileum. *The American Journal of the Medical Sciences*, 98(5), 457–462.

- Botto LD, Lin AE, Riehle-Colarusso T, Malik S, & Correa A (2007). Seeking causes: Classifying and evaluating congenital heart defects in etiologic studies. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 79(10), 714–727. 10.1002/bdra.20403 [PubMed: 17729292]
- Centers for Disease Control and Prevention. (2008). Update on overall prevalence of major birth defects—Atlanta, Georgia, 1978–2005. *MMWR. Morbidity and Mortality Weekly Report*, 57(1), 1–5. [PubMed: 18185492]
- Chesi A, Staahl BT, Jovicic A, Couthouis J, Fasolino M, Raphael AR, ... Gitler AD (2013). Exome sequencing to identify de novo mutations in sporadic ALS trios. *Nature Neuroscience*, 16(7), 851–855. 10.1038/nn.3412 [PubMed: 23708140]
- Chong JX, Buckingham KJ, Jhangiani SN, Boehm C, Sobreira N, Smith JD, ... Bamshad MJ (2015). The genetic basis of Mendelian phenotypes: Discoveries, challenges, and opportunities. *American Journal of Human Genetics*, 97(2), 199–215. 10.1016/j.ajhg.2015.06.009 [PubMed: 26166479]
- Chong JX, Yu JH, Lorentzen P, Park KM, Jamal SM, Tabor HK, ... Bamshad MJ (2016). Gene discovery for Mendelian conditions via social networking: de novo variants in KDM1A cause developmental delay and distinctive facial features. *Genetics in Medicine*, 18(8), 788–795. 10.1038/gim.2015.161 [PubMed: 26656649]
- Cleves MA, Hobbs CA, Zhao W, Krakowiak PA, MacLeod SL, & National Birth Defects Prevention Study. (2011). Association between selected folate pathway polymorphisms and nonsyndromic limb reduction defects: A case-parental analysis. *Paediatric and Perinatal Epidemiology*, 25(2), 124–134. 10.1111/j.1365-3016.2010.01160.x [PubMed: 21281325]
- Davenport M, Bianchi A, Doig CM, & Gough DC (1990). Colonic atresia: Current results of treatment. *Journal of the Royal College of Surgeons of Edinburgh*, 35(1), 25–28. [PubMed: 2342005]
- El-Asmar KM, Abdel-Latif M, El-Kassaby AA, Soliman MH, & El-Behery MM (2016). Colonic atresia: Association with other anomalies. *Journal of Neonatal Surgery*, 5(4), 47. 10.21699/jns.v5i4.422 [PubMed: 27896155]
- Etensel B, Temir G, Karkiner A, Melek M, Edirne Y, Karaca I, & Mir E (2005). Atresia of the colon. *Journal of Pediatric Surgery*, 40(8), 1258–1268. 10.1016/j.jpedsurg.2005.05.008 [PubMed: 16080929]
- Fairbanks TJ, Kanard RC, Del Moral PM, Sala FG, De Langhe SP, Lopez CA, ... Burns RC (2005). Colonic atresia without mesenteric vascular occlusion. The role of the fibroblast growth factor 10 signaling pathway. *Journal of Pediatric Surgery*, 40(2), 390–396. 10.1016/j.jpedsurg.2004.10.023 [PubMed: 15750935]
- Feldkamp ML, Carey JC, Byrne JLB, Krikov S, & Botto LD (2017). Etiology and clinical presentation of birth defects: Population based study. *BMJ*, 357, j2249. 10.1136/bmj.j2249 [PubMed: 28559234]
- Gallagher ML, Sturchio C, Smith A, Koontz D, Jenkins MM, Honein MA, & Rasmussen SA (2011). Evaluation of mailed pediatric buccal cytobrushes for use in a case-control study of birth defects. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 91(7), 642–648. 10.1002/bdra.20829 [PubMed: 21630425]
- Gao N, White P, & Kaestner KH (2009). Establishment of intestinal identity and epithelial-mesenchymal signaling by Cdx2. *Developmental Cell*, 16(4), 588–599. 10.1016/j.devcel.2009.02.010 [PubMed: 19386267]
- Garcia-Closas M, Egan KM, Abruzzo J, Newcomb PA, Titus-Ernstoff L, Franklin T, ... Rothman N (2001). Collection of genomic DNA from adults in epidemiological studies by buccal cytobrush and mouthwash. *Cancer Epidemiology, Biomarkers & Prevention*, 10(6), 687–696.
- Gelb B, Brueckner M, Chung W, Goldmuntz E, Kaltman J, Kaski JP, ... Rosenberg E (2013). The congenital heart disease genetic network study: Rationale, design, and early results. *Circulation Research*, 112(4), 698–706. 10.1161/circresaha.111.300297 [PubMed: 23410879]
- GEMINI. (2017). The GEMINI database schema. Retrieved from https://gemini.readthedocs.io/en/latest/content/database_schema.html?highlight=table%20schema#the-variant-impacts-table.
- Hitchcock R, Birthistle K, Carrington D, Calvert SA, & Holmes K (1995). Colonic atresia and spinal cord atrophy associated with a case of fetal varicella syndrome. *Journal of Pediatric Surgery*, 30(9), 1344–1347. [PubMed: 8523241]

- Hoang TT, Lei Y, Mitchell LE, Sharma SV, Swartz MD, Waller DK, ... Agopian AJ (2019). Maternal lactase polymorphism (rs4988235) is associated with neural tube defects in off-spring in the National Birth Defects Prevention Study. *The Journal of Nutrition*, 149(2), 295–303. 10.1093/jn/nxy246 [PubMed: 30689919]
- Hobbs CA, Cleves MA, Macleod SL, Erickson SW, Tang X, Li J, ... National Birth Defects Prevention Study. (2014). Conotruncal heart defects and common variants in maternal and fetal genes in folate, homocysteine, and transsulfuration pathways. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 100(2), 116–126. 10.1002/bdra.23225 [PubMed: 24535845]
- Howley MM, Feldkamp ML, Papadopoulos EA, Fisher SC, Arnold KE, Browne ML, & National Birth Defects Prevention Study. (2018). Maternal genitourinary infections and risk of birth defects in the National Birth Defects Prevention Study. *Birth Defects Research*, 110(19), 1443–1454. 10.1002/bdr2.1409 [PubMed: 30402975]
- Hunt D, Leventer RJ, Simons C, Taft R, Swoboda KJ, Gawne-Cain M, ... Baralle D (2014). Whole exome sequencing in family trios reveals de novo mutations in PURA as a cause of severe neurodevelopmental delay and learning disability. *Journal of Medical Genetics*, 51(12), 806–813. 10.1136/jmedgenet-2014-102798 [PubMed: 25342064]
- Jenkins MM, Reefhuis J, Gallagher ML, Mulle JG, Hoffmann TJ, Koontz DA, ... National Birth Defects Prevention Study. (2014). Maternal smoking, xenobiotic metabolizing enzyme gene variants, and gastroschisis risk. *American Journal of Medical Genetics. Part A*, 164A(6), 1454–1463. 10.1002/ajmg.a.36478 [PubMed: 24668907]
- Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, ... Kang HM (2012). Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American Journal of Human Genetics*, 91(5), 839–848. 10.1016/j.ajhg.2012.09.004 [PubMed: 23103226]
- Khokha MK, Mitchell LE, & Wallingford JB (2017). An opportunity to address the genetic causes of birth defects. *Pediatric Research*, 81(2), 282–285. 10.1038/pr.2016.229 [PubMed: 27925620]
- Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, & Shendure J (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3), 310–315. 10.1038/ng.2892 [PubMed: 24487276]
- Klebanoff MA (2009). The collaborative perinatal project: A 50-year retrospective. *Paediatric and Perinatal Epidemiology*, 23(1), 2–8. 10.1111/j.1365-3016.2008.00984.x [PubMed: 19228308]
- Kochanek KD, Murphy S, Xu J, & Arias E (2017). Mortality in the United States, 2016 In NCHS Data Brief, no. 293 (pp. 1–8). Hyattsville, MD: National Center for Health Statistics.
- Krauss RS, & Hong M (2016). Gene-environment interactions and the etiology of birth defects. *Current Topics in Developmental Biology*, 116, 569–580. 10.1016/bs.ctdb.2015.12.010 [PubMed: 26970642]
- Krumm N, Sudmant PH, Ko A, O’Roak BJ, Malig M, Coe BP, ... Eichler EE (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Research*, 22(8), 1525–1532. 10.1101/gr.138115.112 [PubMed: 22585873]
- Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, & Maglott DR (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(Database issue), D980–D985. 10.1093/nar/gkt1113 [PubMed: 24234437]
- Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, ... Exome Aggregation Consortium. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616), 285–291. 10.1038/nature19057 [PubMed: 27535533]
- Lewis CJ, Hedberg-Buenz A, DeLuca AP, Stone EM, Alward WLM, & Fingert JH (2017). Primary congenital and developmental glaucomas. *Human Molecular Genetics*, 26(R1), R28–R36. 10.1093/hmg/ddx205 [PubMed: 28549150]
- Li H, & Durbin R (2009). Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14), 1754–1760. 10.1093/bioinformatics/btp324 [PubMed: 19451168]
- Lin AE, Krikov S, Riehle-Colarusso T, Frias JL, Belmont J, Anderka M, ... National Birth Defects Prevention Study. (2014). Laterality defects in the National Birth Defects Prevention Study (1998–2007): Birth prevalence and descriptive epidemiology. *American Journal of Medical Genetics, Part A*, 164A(10), 2581–2591. 10.1002/ajmg.a.36695 [PubMed: 25099286]

- Louik C, Lin AE, Werler MM, Hernandez-Diaz S, & Mitchell AA (2007). First-trimester use of selective serotoninreuptake inhibitors and the risk of birth defects. *The New England Journal of Medicine*, 356(26), 2675–2683. 10.1056/NEJMoa067407 [PubMed: 17596601]
- Louw JH, & Barnard CN (1955). Congenital intestinal atresia; observations on its origin. *Lancet*, 269(6899), 1065–1067. [PubMed: 13272331]
- Lupo PJ, Canfield MA, Chapa C, Lu W, Agopian AJ, Mitchell LE, ... Zhu H (2012). Diabetes and obesity-related genes and the risk of neural tube defects in the National Birth Defects Prevention Study. *American Journal of Epidemiology*, 176(12), 1101–1109. 10.1093/aje/kws190 [PubMed: 23132673]
- Lupo PJ, Chapa C, Noursome D, Duhon C, Canfield MA, Shaw GM, ... National Birth Defects Prevention Study. (2012). A GCH1 haplotype and risk of neural tube defects in the National Birth Defects Prevention Study. *Molecular Genetics and Metabolism*, 107(3), 592–595. 10.1016/j.ymgme.2012.09.020 [PubMed: 23059057]
- Madisen L, Hoar DI, Holroyd CD, Crisp M, & Hodes ME (1987). DNA banking: The effects of storage of blood and isolated DNA on the integrity of DNA. *American Journal of Medical Genetics*, 27(2), 379–390. 10.1002/ajmg.1320270216 [PubMed: 3605221]
- Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, ... Stoltenberg C (2016). Cohort profile update: The Norwegian mother and child cohort study (MoBa). *International Journal of Epidemiology*, 45(2), 382–388. 10.1093/ije/dyw029 [PubMed: 27063603]
- Mai CT, Isenburg J, Langlois PH, Alverson CJ, Gilboa SM, Rickard R, ... National Birth Defects Prevention Network. (2015). Population-based birth defects data in the United States, 2008 to 2012: Presentation of state-specific data and descriptive brief on variability of prevalence. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 103(11), 972–993. 10.1002/bdra.23461 [PubMed: 26611917]
- Maslen CL (2018). Recent advances in placenta-heart interactions. *Frontiers in Physiology*, 9, 735 10.3389/fphys.2018.00735 [PubMed: 29962966]
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, ... Cunningham F (2016). The Ensembl variant effect predictor. *Genome Biology*, 17(1), 122 10.1186/s13059-016-0974-4 [PubMed: 27268795]
- Melvin EC, George TM, Worley G, Franklin A, Mackey J, Viles K, ... Speer MC (2000). Genetic studies in neural tube defects. NTD collaborative group. *Pediatric Neurosurgery*, 32(1), 1–9. 10.1159/000028889 [PubMed: 10765131]
- Mirza B, Iqbal S, & Ijaz L (2012). Colonic atresia and stenosis: Our experience. *Journal of Neonatal Surgery*, 1(1), 4. [PubMed: 26023363]
- Mitchell AA (1988). Slone epidemiology unit birth defects study. *Genetic Resources*, 4, 31–32.
- Mossey PA, Little J, Steegers-Theunissen R, Molloy A, Peterlin B, Shaw WC, ... Rubini M (2017). Genetic interactions in nonsyndromic orofacial clefts in Europe-EUROCRAN study. *The Cleft Palate-Craniofacial Journal*, 54(6), 623–630. 10.1597/16-037 [PubMed: 27996298]
- Nembhard WN, Tang X, Hu Z, MacLeod S, Stowe Z, & Webber D (2017). Maternal and infant genetic variants, maternal periconceptional use of selective serotonin reuptake inhibitors, and risk of congenital heart defects in offspring: Population based study. *BMJ*, 356, j832 10.1136/bmj.j832 [PubMed: 28264803]
- NHLBI Exome Sequencing Project. (n.d.). Exome variant server. Retrieved from <http://evs.gs.washington.edu/EVS/>.
- NORD. (2004). Split hand/split foot malformation. Retrieved from <https://rarediseases.org/rare-diseases/split-handsplit-foot-malformation/>.
- NORD. (2016). Caudal regression syndrome. Retrieved from <https://rarediseases.org/rare-diseases/caudal-regression-syndrome/>.
- Olsen J, Melbye M, Olsen SF, Sorensen TI, Aaby P, Andersen AM, ... Sondergaard C (2001). The Danish National Birth Cohort—Its background, structure and aim. *Scandinavian Journal of Public Health*, 29(4), 300–307. [PubMed: 11775787]
- Olshan AF, Hobbs CA, & Shaw GM (2011). Discovery of genetic susceptibility factors for human birth defects: An opportunity for a national agenda. *American Journal of Medical Genetics. Part A*, 155A(8), 1794–1797. 10.1002/ajmg.a.34103 [PubMed: 21739590]

- Paila U, Chapman BA, Kirchner R, & Quinlan AR (2013). GEMINI: Integrative exploration of genetic variation and genome annotations. *PLoS Computational Biology*, 9(7), e1003153 10.1371/journal.pcbi.1003153 [PubMed: 23874191]
- Parker SE, Mai CT, Canfield MA, Rickard R, Wang Y, Meyer RE, ... National Birth Defects Prevention Network. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 88(12), 1008–1016. 10.1002/bdra.20735 [PubMed: 20878909]
- Pedersen BS, & Quinlan AR (2017). Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *American Journal of Human Genetics*, 100(3), 406–413. 10.1016/j.ajhg.2017.01.017. [PubMed: 28190455]
- Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, ... Rehm HL (2015). The match-maker exchange: A platform for rare disease gene discovery. *Human Mutation*, 36(10), 915–921. 10.1002/humu.22858 [PubMed: 26295439]
- Powell RW, & Raffensperger JG (1982). Congenital colonic atresia. *Journal of Pediatric Surgery*, 17(2), 166–170. [PubMed: 7077498]
- Rasmussen SA, Lammer EJ, Shaw GM, Finnell RH, McGehee RE Jr., Gallagher M, ... Murray JC (2002). Integration of DNA sample collection into a multi-site birth defects case-control study. *Teratology*, 66(4), 177–184. 10.1002/tera.10086 [PubMed: 12353214]
- Rasmussen SA, Olney RS, Holmes LB, Lin AE, Keppler-Noreuil KM, & Moore CA (2003). Guidelines for case classification for the National Birth Defects Prevention Study. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 67(3), 193–201. 10.1002/bdra.10012
- Reefhuis J, Gilboa SM, Anderka M, Browne ML, Feldkamp ML, Hobbs CA, ... National Birth Defects Prevention Study. (2015). The National Birth Defects Prevention Study: A review of the methods. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 103(8), 656–669. 10.1002/bdra.23384
- Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, & Mesirov JP (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1), 24–26. 10.1038/nbt.1754
- Sauve RS, & Leung AK (2003). Congenital varicella syndrome with colonic atresias. *Clinical Pediatrics (Phila)*, 42(5), 451–453. 10.1177/000992280304200512
- Schmidt RJ, Romitti PA, Burns TL, Murray JC, Browne ML, Druschel CM, ... National Birth Defects Prevention Study. (2010). Caffeine, selected metabolic gene variants, and risk for neural tube defects. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 88(7), 560–569. 10.1002/bdra.20681
- Sharafieh R, Child AH, & Sarfarazi M (2012). Molecular genetics of primary congenital glaucoma In Traboulsi EI (Ed.), *Genetic disease of the eye* (2nd ed, pp. 295–305). New York, NY: Oxford University Press.
- Spielman RS, McGinnis RE, & Ewens WJ (1993). Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics*, 52(3), 506–516. [PubMed: 8447318]
- Tai CG, Graff RE, Liu J, Passarelli MN, Mefford JA, Shaw GM, ... Witte JS (2015). Detecting gene-environment interactions in human birth defects: Study designs and statistical methods. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 103(8), 692–702. 10.1002/bdra.23382
- Tang X, Cleves MA, Nick TG, Li M, MacLeod SL, Erickson SW, ... National Birth Defects Prevention Study. (2015). Obstructive heart defects associated with candidate genes, maternal obesity, and folic acid supplementation. *American Journal of Medical Genetics. Part A*, 167(6), 1231–1242. 10.1002/ajmg.a.36867 [PubMed: 25846410]
- Tang X, Hobbs CA, Cleves MA, Erickson SW, MacLeod SL, Malik S, & National Birth Defects Prevention Study. (2015). Genetic variation affects congenital heart defect susceptibility in offspring exposed to maternal tobacco use. *Birth Defects Research. Part A, Clinical and Molecular Teratology*, 103(10), 834–842. 10.1002/bdra.23370 [PubMed: 26033827]
- Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, Swift AJ, ... Biesecker LG (2010). Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Research*, 20(10), 1420–1431. 10.1101/gr.106716.110 [PubMed: 20810667]

- UK10K Consortium, Walter K., Min JL., Huang J., Crooks L., Memari Y., ... Soranzo N. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571), 82–90. 10.1038/nature14962 [PubMed: 26367797]
- Umbach DM, & Weinberg CR (2000). The use of case-parent triads to study joint effects of genotype and exposure. *American Journal of Human Genetics*, 66(1), 251–261. 10.1086/302707 [PubMed: 10631155]
- Vaser R, Adusumalli S, Leng SN, Sikic M, & Ng PC (2016). SIFT missense predictions for genomes. *Nature Protocols*, 11(1), 1–9. 10.1038/nprot.2015.123 [PubMed: 26633127]
- Visvikis S, Schlenck A, & Maurice M (1998). DNA extraction and stability for epidemiological studies. *Clinical Chemistry and Laboratory Medicine*, 36(8), 551–555. 10.1515/cclm.1998.094 [PubMed: 9806458]
- Waller DK, Hashmi SS, Hoyt AT, Duong HT, Tinker SC, Gallaway MS, ... National Birth Defects Prevention Study. (2018). Maternal report of fever from cold or flu during early pregnancy and the risk for noncardiac birth defects, National Birth Defects Prevention Study, 1997–2011. *Birth Defects Research*, 110(4), 342–351. 10.1002/bdr2.1147 [PubMed: 29094488]
- Wang K, Li M, & Hakonarson H (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16), e164 10.1093/nar/gkq603 [PubMed: 20601685]
- Weinberg CR, Wilcox AJ, & Lie RT (1998). A log-linear approach to case-parent-triad data: Assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. *American Journal of Human Genetics*, 62(4), 969–978. 10.1086/301802 [PubMed: 9529360]
- Winters WD, Weinberger E, & Hatch EI (1992). Atresia of the colon in neonates: Radiographic findings. *AJR. American Journal of Roentgenology*, 159(6), 1273–1276. 10.2214/ajr.159.6.1442400 [PubMed: 1442400]
- Yoon PW, Rasmussen SA, Lynberg MC, Moore CA, Anderka M, Carmichael SL, ... Edmonds LD (2001). The National Birth Defects Prevention Study. *Public Health Reports*, 116(Suppl 1), 32–40. 10.1093/phr/116.S1.32
- Yu L, Wynn J, Ma L, Guha S, Mychaliska GB, Crombleholme TM, ... Chung WK (2012). De novo copy number variants are associated with congenital diaphragmatic hernia. *Journal of Medical Genetics*, 49(10), 650–659. 10.1136/jmedgenet-2012-101135 [PubMed: 23054247]

TABLE 1

Family trios included in exome sequencing projects by birth defect, National Birth Defects Prevention Study 1997–2011

Birth defect	Estimated national prevalence per 10,000 live births	Reference	Family trios ^a
Colonic atresia/stenosis	4.7	Parker et al. (2010)	20
Transverse limb reduction	3.5 (upper), 1.7 (lower)	Parker et al. (2010)	190
Anterior segment dysgenesis	0.05–0.16	Lewis et al. (2017)	24
Congenital glaucoma	1.0	Sharafieh, Child, and Sarfarazi (2012)	38
Split hand/foot malformation	0.6	NORD (2004)	26
Cloacal exstrophy	2.2	Mai et al. (2015)	21
Bladder exstrophy	0.3	Mai et al. (2015)	27
Anophthalmos/microphthalmos	1.1	Mai et al. (2015)	68
Sacral agenesis	0.1–0.5	NORD (2016)	34
Biliary atresia	0.6	Mai et al. (2015)	59
Hypoplastic left heart syndrome	2.5	Mai et al. (2015)	140
Tricuspid atresia	1.4	Mai et al. (2015)	46
Ebstein anomaly	0.7	Mai et al. (2015)	49
Heterotaxy	0.8	Lin et al. (2014)	101

^aCounts include family trios who had sufficient genomic DNA from dry brushes (> 200 ng).

TABLE 2
 Comparison of exome sequencing quality metrics from WGA product or gDNA, National Birth Defects Prevention Study 1997–2011

Specimen number	Brush type	Percent of reads aligning with human reference		Mean depth of coverage per exome		Percent on-target (MPG 10)		Total variants	
		WGA	gDNA	WGA	gDNA	WGA	gDNA	WGA	gDNA
1	Wet	8.1	58.2	11.1	105.9	40.6	93.0	27,392	86,129
2	Wet	39.3	55.6	54.6	122.4	85.7	93.1	86,586	86,586
3	Dry	4.8	38.5	3.3	103.0	3.1	93.2	2,985	85,912
4	Dry	15.5	44.7	14.2	137.8	51.0	93.2	34,689	87,231
5	Dry	22.2	62.1	23.6	101.6	66.6	92.2	50,458	84,227
6	Dry	41.7	70.9	64.1	110.6	82.0	92.4	68,710	83,621
7	Dry	60.1	70.8	51.9	115.0	79.2	92.5	67,856	87,139
8	Dry	64.1	74.1	51.5	99.3	81.8	92.1	67,601	83,149
9	Dry	67.1	70.7	52.4	93.1	83.7	92.5	70,596	83,426
10	Dry	70.1	73.2	61.4	117.2	85.5	92.5	72,422	83,470
Average		39.3	64.5	38.8	110.6	65.9	92.6	54,929	85,089

Abbreviations: gDNA, genomic DNA; MPG, most probable genotype; WGA, whole genome amplified.

TABLE 3 Proof-of-principle study results: Characteristics of each trio and exome sequencing summary statistics per individual, National Birth Defects Prevention Study, 1997–2011

Trio	Family member	Self-reported race/ethnicity	Colonic Atresia ^a	Coverage ^b		Total variants ^c
				Mean	Proportion at 20×	
T1	Proband	Non-Hispanic White	Type III	47.0	0.87	60,427
	Mother	Non-Hispanic White	Unaffected	60.5	0.94	62,597
	Father	Non-Hispanic White	Unaffected	62.9	0.94	63,064
T2	Proband	Hispanic	NOS	56.9	0.91	63,407
	Mother	Hispanic	Unaffected	64.0	0.93	64,183
	Father	Hispanic	Unaffected	73.3	0.94	63,261
T3	Proband	Hispanic	Type III	70.3	0.95	64,678
	Mother	Hispanic	Unaffected	66.7	0.94	64,185
	Father	Hispanic	Unaffected	55.9	0.89	63,316
T4	Proband	Non-Hispanic White	Type III	85.9	0.97	63,245
	Mother	Non-Hispanic White	Unaffected	70.5	0.94	62,667
	Father	Non-Hispanic White	Unaffected	76.6	0.95	62,717
T5	Proband	Hispanic	NOS	75.4	0.95	64,603
	Mother	Hispanic	Unaffected	93.8	0.96	65,470
	Father	Hispanic	Unaffected	78.7	0.90	63,642
T6	Proband	Non-Hispanic White	NOS	56.9	0.93	62,375
	Mother	Non-Hispanic White	Unaffected	60.4	0.94	61,776
	Father	Non-Hispanic White	Unaffected	73.9	0.95	63,442
T7	Proband	Hispanic	NOS	76.3	0.96	63,633
	Mother	Hispanic	Unaffected	64.7	0.94	62,758
	Father	Hispanic	Unaffected	66.4	0.94	64,007
T8	Proband	Hispanic	NOS	42.8	0.87	63,097
	Mother	Hispanic	Unaffected	45.0	0.88	60,733
	Father	Hispanic	Unaffected	48.1	0.89	63,087
Average				65.5	0.93	63,182

Abbreviation: NOS, not otherwise specified.

^aColonic atresia classification: Type III, characterized by a “V” shape in the mesentery with proximal and distal blind ends.

^bCoverage per specimen from Picard.

^cTotal number of variants from GEMINI v0.20.2.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript