The Gabriella Miller Kids First Pediatric Research Program (Kids First) Poster Session at ASHG
Accelerating Pediatric Genomics Research through Collaboration
October 15th, 2019

Background

The [Gabriella Miller Kids First Pediatric Research Program](#) (Kids First) is a trans-NIH Common Fund program initiated in response to the [2014 Gabriella Miller Kids First Research Act](#). The program's vision is to alleviate suffering from childhood cancer and structural birth defects by fostering collaborative research to uncover the etiology of these diseases and support data sharing within the pediatric research community. This is implemented through developing the Gabriella Miller Kids First Data Resource (Kids First Data Resource) and populating this resource with whole genome sequence datasets and associated clinical and phenotypic information. Both childhood cancers and structural birth defects are critical and costly conditions associated with substantial morbidity and mortality. Elucidating the underlying genetic etiology of these diseases has the potential to profoundly improve preventative measures, diagnostics, and therapeutic interventions.

Purpose

During this evening poster session, attendees will gain a broad understanding of the utility of the genomic data generated by Kids First, learn about the progress of Kids First X01 cohort projects, and observe demonstrations of the tools and functionalities of the recently launched [Kids First Data Resource Portal](#). The session is an opportunity for the scientific community and public to engage with Kids First investigators, collaborators, and a growing community of researchers, patient foundations, and families. Several other NIH and external data efforts will present posters and be available to discuss collaboration opportunities as we work together to accelerate pediatric research. NIH Kids First program staff will be available to share information about the program as well as current funding opportunities, answer questions, and receive feedback.

Project/Program (links to abstracts)	Presenter	Affiliation	Poster Board Key (see map below)
Demonstrations @ Screen (see schedule below)	*Meet the DRC Experts*	Kids First Data Resource Center (https://kidsfirstdrc.org)	DRP
NIH Kids First Program	James Coulombe	Kids First, NICHD, NIH	NIH
Kids First X01 "mini" posters (for PIs who could not attend)	Valerie Cotton	Kids First, NICHD, NIH	NIH-miniX01s
Kids First Data Resource Center Posters			
Kids First DRC – Phenotype Harmonization	Deanne Taylor	The Children's Hospital Of Philadelphia /UPenn	DRC-P
Kids First DRC – Genomic Harmonization	Yuankun Zhu	The Children's Hospital Of Philadelphia	DRC-G
Kids First DRC – Outreach Strategies	Jena Lilly	The Children's Hospital Of Philadelphia	DRC-O
Kids First Sequencing Centers			
Broad	Stacey Gabriel	Broad Institute	Seq-B
HudsonAlpha-St. Jude	Shawn Levy	HudsonAlpha Institute for Biotechnology	Seq-H
Kids First X01 Cohort Investigators			
Orofacial Clefts Overview	Elizabeth Leslie	Emory University	X01-1
Orofacial Clefts	Madison Bishop	Emory University	X01-2
Orofacial Clefts	Nandita Mukhopadhyay	Univ. of Pittsburgh	X01-3
Disorders of Sex Development	Emmanuele Delot	Children's National Medical Center & George Washington University	X01-4
Kidney & Urinary Tract Defects	Hila Milo Rasouly	Columbia University	X01-5
Familial Hodgkin Lymphoma	Jamie Flerlage	St. Jude Children's Research Hospital	X01-6

Return to top ↰

Ewing Sarcoma	Gareth Mitchell	Huntsman Cancer Institute	X01-7
Neuroblastoma	Zalman Vaksman	The Children's Hospital Of Philadelphia	X01-8
Intersection of childhood cancer and structural birth defects	Patrick Sleiman	The Children's Hospital Of Philadelphia	X01-9
Down syndrome, CHD, and ALL	Philip Lupo	Baylor College of Medicine	X01-10
Rhabdomyosarcoma	Philip Lupo	Baylor College of Medicine	X01-11
Cancer Susceptibility	Owen Hirschi	Baylor College of Medicine	X01-12
Craniosynostosis	Paul Romitti	Kids First X01 Craniosynostosis	X01-13
CHARGE Syndrome	Donna Martin	University of Michigan	X01-14
Structural Heart Defects	Bruce Gelb	Pediatric Cardiac Genomics Consortium	X01-15
Hearing Loss	Cynthia Morton	Brigham and Women's Hospital	X01-16
Bladder extrophy, Epispadias	Angie Jelin	Johns Hopkins	X01-17
Adolescent idiopathic scoliosis	Carol Wise	TX Scottish Rite Hospital	X01-18
Cranial Dysinnervation	Arthur Lee	Boston Children's Hospital	X01-19
	Julie Jurgens	Boston Children's Hospital	
Congenital Diaphragmatic Hernia	Yufeng Shen	Columbia University	X01-20
Enchondromas (cartilage tumors)	Nara Sobreira	Johns Hopkins University	X01-21, 22, 23
Cornelia de Lange	Ian Krantz	The Children's Hospital Of Philadelphia	X01-24
Collaborators			
Newborn Screening Translational Research Network (NBSTRN)	Amy Brower	ACMG	C1

Sequence Data Delivery Pilot	Stephen Sherry	National Center for Biotechnology Information	C2
INCLUDE	Melissa Parisi	NICHD	C3
NHLBI TOPMed	Matthew P Conomos & Sarah Nelson	TOPMed Data Coordinating Center, University of Washington	C4
NHLBI's New Data Platform	Stan Ahalt	RENCI	C5
NCI CRDC	Erika Kim	NCI	C6
St. Jude Cloud	Alex Gout	St. Jude Children's Research Hospital	C7
Common Fund Data Ecosystem	C. Titus Brown	UC Davis	C8
KOMP2	Steve Murray	The Jackson Laboratory	C9
Gene Expression Database	Jacqueline Finger	Gene Expression Database	C10
Rare Diseases ilClinical Research Network	Peter White	Cincinnati Children's Hospital Medical Center	C11
Undiagnosed Diseases Network (UDN)	Cecilia Esteves	Undiagnosed Diseases Network (UDN)	C12
Children's Brain Tumor Tissue Consortium (CBTTC)	Adam Resnick	The Children's Hospital of Philadelphia, CBTTC	C13
CBTTC: pediatric brain cancer genomic analysis	Yiran Guo	The Children's Hospital Of Philadelphia	C14
(Non-KF) Neuroblastoma	Zhenqiu Liu	Penn State University	C15
Monarch	Peter Robinson	The Jackson Laboratory	C16

Poster Schedule:
- **6-7pm**: Kids First Investigators at their posters (X01s, sequencing centers, DRC)
- **7-8pm**: Collaborators at their posters (labeled with "C")
- **8-9pm**: Continued Discussion and Networking (presenters are welcome to continue standing at their poster or visit others!)

Data Resource Portal (DRP) Demonstrations:
- <u>6:00-6:15pm</u>: *Explore Data* Tool Demo
- <u>6:30-6:45pm</u>: Cavatica Demo, the X01 Journey & *Meet the Expert* with **Allison Heath, PhD**
- <u>7:00-7:15pm</u>: Clinical and Phenotypic Harmonization & *Meet the Expert* with **Deanne Taylor, PhD**
- <u>7:30-7:45pm</u>: Genomic Harmonization & *Meet the Expert* with **Yuankun Zhu**
- <u>8:00-8:15pm</u>: Data Resource Portal Overview & *Meet the Expert* with **Adam Resnick, PhD**

<div style="border: 1px solid black; text-align: center;">
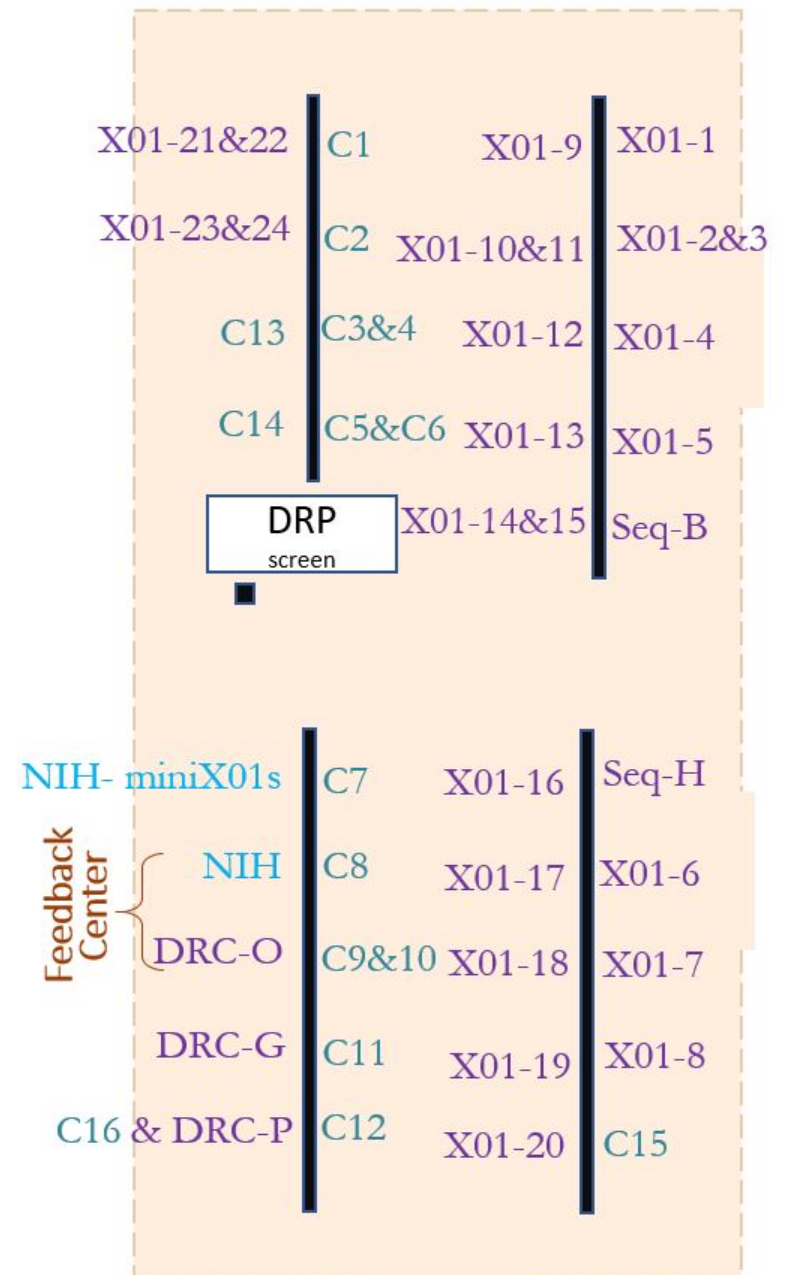
ABSTRACTS

</div>

The Gabriella Miller Kids First Pediatric Research Program

NIH Gabriella Miller Kids First Working Group, presented by **Valerie Cotton** and **James Coulombe**
National Institutes of Health

Childhood cancers and structural birth defects have profound, life-long effects on patients and their families. A risk factor for developing childhood cancer is being born with a birth defect, suggesting there are shared genetic pathways underlying these conditions. Additionally, these conditions tend to be relatively rare; therefore, studying sufficiently powered datasets requires coordinated efforts, collaboration, and data sharing. In 2015, the NIH Common Fund established the Gabriella Miller Kids First Pediatric Research Program (Kids First), a data sharing initiative focused on generating genomic data and developing tools and resources to help researchers better understand the genetic mechanisms that underlie structural birth defects and childhood cancers. In 2018, Kids First launched the Gabriella Miller Kids First Data Resource Portal (https://portal.kidsfirstdrc.org/), a web-based platform which serves as the primary entry point to large amounts of genetic and associated clinical and phenotypic data from patients with structural birth defects and childhood cancers and their families, and cloud-based workspace to empower analyses of these data. To date, Kids First has selected 33 patient cohorts for whole genome sequencing, representing 36,000 genomes and 14,000 patients affected with the following conditions:

- Adolescent Idiopathic Scoliosis
- Bladder Extrophy, Epispadias Complex
- Cancer Susceptibility
- Childhood Acute Myeloid Leukemia
- Congenital Anomalies of the Kidney and Urinary Tract
- Cornelia de Lange Syndrome
- Congenital Diaphragmatic Hernia
- Congenital Heart Defects and Acute Lymphoblastic Leukemia in Children with Down Syndrome
- Craniofacial Microsomia
- Disorders of Sex Development
- Enchondromatoses
- Esophageal Atresia and Tracheoesophageal Fistulas
- Ewing Sarcoma
- Familial Leukemia
- Fetal Alcohol Spectrum Disorders
- Hearing Loss
- Infantile Hemangiomas
- Laterality Birth Defects

- CHARGE Syndrome (Coloboma of the eye, Heart Defects, Atresia of the choanae, Retardation of growth and development, Genital abnormalities including pubertal delay and infertility, Ear abnormalities with deafness and vestibular disorders)
- Kidney and Urinary Tract
- Neural Tube Defects (Spina Bifida)
- Neuroblastomas
- Nonsyndromic Craniosynostosis
- Microtia
- Orofacial Clefts
- Osteosarcoma
- Patients with both childhood cancer & birth defects
- Pediatric Extracranial Germ Cell Tumors
- Pediatric Intracranial Germ Cell Tumors
- Pediatric Rhabdomyosarcoma
- Structural Brain Defects in Children
- Structural Heart & Other Defects
- Syndromic Cranial Dysinnervation Disorders
- T-cell Acute Lymphoblastic Leukemia
- Vascular Anomaly Syndromes; PHACE, LUMBAR, CLOVES and Overgrowth syndromes

The Kids First program's vision is to alleviate suffering from structural birth defects and childhood cancer by fostering collaboration and supporting data sharing to accelerate research toward more effective preventative measures, diagnostics, and treatments for these pediatric conditions.

Quantifying phenotype similarity for complex harmonized disease cohorts in the Gabriella Miller Kids First Data Resource Center

Lilia Becker[1], **Deanne M. Taylor**[1,2] Lei Qian[1], Nicole Vasilevsky[3], Shahim Essaid[3], Melissa Haendel[3,4], Avital Kelman[5], Natasha Singh[5], Parimala Killada[5], Meen Chul Kim[5], Bailey K. Farrow[5], Brian Furner[6], Luca Graglia[6],Samuel Volchenbaum[6,7], Allison Heath[5], Adam Resnick[5], and all contributors to the Gabriella Miller Kids First Data Resource Center.

(1) Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA; (2) Department of Pediatrics, Perelman School of Medicine, The University of Pennsylvania, Philadelphia PA; (3) Translational and Integrative Sciences Lab, Oregon Health & Science University, Portland, OR; (4) Oregon Health & Science University, Portland, OR (5) The Center for Data Driven Discovery in Biomedicine, The Children's Hospital of Philadelphia, Philadelphia, PA; (6) Center for Research Informatics, The University of Chicago, Chicago, IL; (7) Department of Pediatrics, Biological Sciences Division, The University of Chicago, Chicago, IL

The NIH Common Fund's Gabriella Miller Kids First Pediatric Research Program (KF) is a collaborative initiative focused on providing large-scale clinically-annotated genomic data for pediatric cancer and structural birth defect cohorts, including trio germline whole genome sequencing (WGS) and tumor WGS and RNA-seq. The KF Kids First Data Resource Center (KFDRC; https://kidsfirstdrc.org/) is charged with empowering collaborative research and discovery through integration of KF cohorts and external data. As of mid-2019, there are 33 GMKF unique disease cohorts funded through KF: 23 structural birth defect (SBD) and 10 pediatric cancer (PC) cohorts, with one mixed SBD & PC cohort. Approximately 7300 participants from 2300 SBD and PC families have been harmonized into the KFDRC since the launch of the portal in June 2018, making the KFDRC one of the largest pediatric data resources of its kind (http://www.kidsfirstdrc.org).

Currently, over 23,000 individuals (probands and families) have been funded by KF for sequencing and harmonization, representing approximately 7000 individual probands, all of whom have disease-specific clinical and phenotype data that must be harmonized against common classification sets for cohort analysis intra- and inter-operability.

We show how the Kids First Data Resource is working to harmonize clinical and outcome data across phenotypes and diseases using the Human Phenotype Ontology (HPO) for phenotype classification. Notably, we present work on quantification of individual and cohort-level phenotype content and pairwise similarities using metrics such as the cosine measure and principal component analyses. Quantification of phenotype similarity is useful for identifying not only cohort groups that may be more analytically interoperable in genetic analyses, but will also help inform contributing investigators as to the quality and interoperability of their own clinical phenotypic datasets with the rest of KF cohorts.

Kids First Data Resource Center: Genomic Harmonization

Yuankun Zhu
Children's Hospital of Philadelphia

To empower collaborative discoveries across the Gabriella Miller Kids First Pediatric Research Program and other integrated datasets, the Kids First Data Resource Center (DRC) is tasked to build infrastructure and workflows for data intaking, harmonization, integration and data access. As part of the DRC data harmonization effort, the DRC Genomic Harmonization team has developed and deployed cloud based workflows following the GATK Best Practice recommendations with the goal of being functionally equivalent with other current large genomic research efforts. The data processing is done via the Cavatica platform within an Amazon Web Services (AWS) environment. The workflow is featured with scatter-gather parallelization and AWS resources optimization. Over 10,000 WGS and 1,000 RNA-Seq has been harmonized using this framework. We present challenges and opportunities in analysis, and integration of the genomic data on a large scale.

Working together to put kids first: Outreach strategies driving collaborative research, data sharing and cross-disease analysis to accelerate discoveries in pediatric cancer and structural birth defects

Tatiana Patton (4), Robert Moulder (4), Erin Alexander (3,4), Donna Vito (1), Jonathan Waller (4), Colleen Gaynor (1,4), Sarah Thomas (4), Bailey Farrow (4), Joseph Yamada (5), Kim Cullion (5), Danyelle Winchester (6), Angela Waanders (2,4), Allison Heath (3,4), Pichai Raman (1,4), Adam Resnick (1,2,3,4), **Jena Lilly** (3,4)

(1)Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia, Philadelphia, PA
(2)Division of Oncology & Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, PA
(3)Division of Neurosurgery, The Children's Hospital of Philadelphia, Philadelphia, PA
(4)Center for Data-Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA
(5) Ontario Institute of Cancer Research
(6) Division of Program Coordination, Planning, and Strategic Initiatives, Office of the Director, National Institutes of Health (NIH)

The Gabriella Miller Kids First Program launched the Kids First Pediatric Data Resource Center (DRC) in 2017 as a collaborative, pediatric research initiative with the goal of understanding the genetic causes of and the potential links between childhood cancer and structural birth defects that can advance . The DRC is charged developing data-driven platforms that integrate large amounts of genomic and clinical data, empowering the collaborative discovery, engagement, and necessary partnerships that are crucial for progress in our biological understanding of diseases, enabling rapid translation to personalized treatments for patients and accelerating discovery of genetic causes and shared biologic pathways within and across these conditions. The DRC is comprised of 3 cores including the Data Resource Portal Core, Data Coordination Core and the Administrative & Outreach Core (AOC).

The AOC brings together researchers, physicians, and patient and foundation advocates to support collaborative research and data sharing to accelerate discoveries. The AOC specific aims are to employ outreach strategies including print, web, social media, in-person presentations, conferences, videos, webinars, e-newsletters, surveys, communication strategies, and reports to support accelerated discoveries.

The AOC is committed to learning from the childhood cancer and birth defect communities. By capturing, synthesizing, and prioritizing unmet needs for development of the Kids First DRC portal, website, and materials, the AOC engages with researchers, clinicians, foundations, and patient advocates in

the childhood cancer and structural birth defect communities. In its first year, the AOC partnered with 32 foundations to launch the Kids First DRC Portal and support data sharing throughout the research community Key findings during the first six months of requirements gathering revealed the following unmet needs: a) Increase understanding of the disease types, research projects, and the investigators that are a part of the Kids First community b) Highlight the need for cross-disease analyses including structural birth defects and childhood cancers and c) Promote education on the data sharing, agreements, data availability and accessibility.

The AOC, gathered pertinent user requirements, conducted educational activities, and engaged prospective users of the researcher community resulting in over 200 users and 22,000 portal views since launch and will continue to use feedback from the research community to further inform the development of the Kids First DRC tools and materials to meet the goals of the program.

Whole Genome Sequencing in Support of the Gabriella Miller Kids First Pediatric Research Program

STACY MANO[1], STACEY GABRIEL[1], DANIEL MACARTHUR[1,2], MICHAEL TALKOWSKI[1,2]
[1]Broad Institute of MIT and Harvard, Cambridge, MA, [2]Massachusetts General Hospital, Boston, MA

The Gabriella Miller Kids First (GMKF) program presents the opportunity to provide a rich genomic data resource to propel pediatric disease research. Key elements to a successful program will be the provision of high quality genome sequence data on well-phenotyped patients and their families; the collection and accessibility of data to the research community in an intuitive manner; and the integration of genetic data with phenotypic information in the context of this program and comparison to other large data resources. The ultimate goal is to assemble a complete catalogue of genes that underlie structural birth defects and pediatric cancer and to enable the use of this information to better understand disease mechanisms, diagnostic opportunities and therapeutic direction.

The sequencing center at the Broad Institute serves as a resource for the GMKF Research Program, as we have done in support of other large flagship NIH genome projects. Our center brings the domain expertise in high throughput data generation, processing and analysis, and disease gene discovery required to meet the objectives of the GMKF Program. We apply deep, high-quality phased whole genome sequencing data on selected samples. We are prepared to apply our well-tested methods for extraction of DNA from a range of sample types, most importantly saliva samples and paraffin-embedded material which are key to pediatric and cancer research.

Over the three year period we have processed just under 8,000 samples pushing the boundary on new data types and lower cost. We are flexible to a mix of cohort types, whether they are trio based (for structural birth defects) or quads (in cancer studies). We also work with investigators to perform follow up and functional validation as needed. A key feature of our center is our implementation of a robust analytical framework for variant assessment and disease gene discovery, which takes advantage of Broad investigators' world-leading roles in statistical genetics, functional annotation, and clinical variant interpretation as well as access to exome and genome data from over 250,000 reference samples. This has enabled us to build a systematic pipeline for gene discovery that will be made freely available to the GMKF program and collaborators.

For year four, we were awarded 5021 samples with a flexible mix of cohorts. We are committed to the continued generation of high quality data on the awarded samples.With data produced and processed in a consistent way, we can offer seamless integration of GMKF data into our analytic framework. For many of the diseases targeted by pediatric research community, confident discovery of causal genes will require aggregation of cases across centers around the world. We offer to enable a new standard for data sharing in clinical genomics by rapidly releasing genetic and phenotype data, accelerating collaboration and facilitating robust disease gene discovery.

The overall goal of this project is to generate high quality sequence data to help researchers understand the underlying mechanisms of disease, leading to more refined diagnostic capabilities and ultimately more targeted therapies or interventions.

HudsonAlpha-St Jude Genome Sequencing Center

Shawn Levy & John Easton

HudsonAlpha Institute for Biotechnology & St. Jude Children's Research Hospital

Expanding our understanding of the genetic contributions and etiologies of birth defects and childhood cancer will have a significant and direct impact on individuals and families affected by those conditions. The HudsonAlpha-St Jude Genome Sequencing Center (HASJ-GSC) has expands its capabilities to provide an efficient and experienced genomics resource for the generation and analysis of high-quality sequence and variant data to support the Gabriella Miller Kids First Research Program. We are looking forward to collaborating with the entire program and present here the innovation, capabilities, and experience of the HASJ-GSC to produce data for the Gabriella Miller Kids First Research Program that will be accessible and available to the research community and leveraged to its maximal impact for years to come. The HASJ-GSC will generate exceptional quality whole-genome sequence and variant data for all samples and for the pediatric cancer samples, RNASeq and whole-exome data. The combined genome, exome and RNASeq will provide as much resolution as possible to understanding the genetic and functional genomic changes observed in pediatric cancer. The HASJ-GSC will also provide support for a comprehensive collection of additional methodologies such as long read, linked-long read and RNAseq. These will expand the resolution and type of investigation that can be supported for X01 investigators. The HASJ-GSC will also provide a reliable and efficient data storage and data access capability that provides fast, reliable and efficient data access, sharing and reporting for the project/sample Program Directors/Principal Investigators (PDs/PIs) (X01 PDs/PIs). Finally, the HASJ-GSC will support efficient submission of sequence and variant data to the Gabriella Miller Kids First Data Resource Center and help facilitate submission to the appropriate public databases. Together, HudsonAlpha and St Jude Children's Research Hospital have created a sequencing resource that leverages the respective experience and capabilities of the two institutions and makes those capabilities available for the Gabriella Miller Kid's First Research Program.

Orofacial Clefts: Whole Genome Sequencing Resources from Kids First

Mary Marazita[1], Eleanor Feingold[1], Azeez Butali[2], Terri Beaty[3], Jeff Murray[2], **Elizabeth Leslie[4]**

      1-University of Pittsburgh
      2-University of Iowa
      3-Johns Hopkins University
      4-Emory University

Orofacial clefts (OFCs) are one of the most recognizable of all human birth defects and represent a substantial personal, familial, economic, and public health burden. OFCs are the most common craniofacial birth defect in humans (about 1 in 700 live births) and are a heterogeneous group of disorders affecting the lip (CL), the lip and palate (CLP), or the palate alone (CP). Individuals with CL and CLP are often grouped together and referred to as CL with or without CP (CL/P). There are notable ethnic and gender differences in the birth prevalence rate of OFCs with Asian populations having higher rates of OFCs and African populations the lowest. The majority of OFCs (70% of CL/P cases; 50% of CP cases) occur as isolated (i.e. nonsyndromic) anomalies with no other major cognitive or structural anomalies and are regarded as a complex disorder with an etiology reflecting the action of multiple genetic and environmental risk factors. During childhood, affected OFC individuals suffer from feeding difficulties, speech, hearing, and dental problems that require multiple craniofacial and dental surgeries and ongoing therapies that result in lifetime healthcare costs in excess of several hundred thousand dollars in high resource settings but are a substantial cause of mortality in low resource settings. Although the long-term prognosis is excellent for most individuals with OFCs, they can experience lifelong psychosocial effects, increased mortality rates from all causes, and a higher risk of various cancer types. Whole genome sequencing resources from Kids First provide an opportunity to elucidate the genetic architecture of OFCs in diverse populations. Understanding the etiologies of this disorder is important not only for increasing our knowledge of craniofacial biology and how clefts arise, but ultimately for improved prevention, treatment, and prognosis for individuals affected by OFCs.

The contribution of de novo mutations identified by WGS of 643 trios with orofacial clefts

Madison Bishop[1], Pankaj Chopra[1], Nandita Mukhopadhyay[2], David J. Cutler[1], Michael P. Epstein[1], George L. Wehby[3], Luz Consuelo Valencia-Ramirez[4], Claudia Restrepo[4], Lina Moreno-Uribe[3], Jacqueline T. Hecht[5], Frederic Deleyiannis[6], Seth M. Weinberg[2], Terri H. Beaty[7], Jeffrey C. Murray[3], Eleanor Feingold[2], Mary L. Marazita[2], Elizabeth J. Leslie[1]

[1]Emory University, Atlanta, GA, [2]University of Pittsburgh, Pittsburgh, PA, [3]University of Iowa, Iowa City, IA, [4]Fundación Clinica Noel, Medellin, Colombia, [5]McGovern Medical School UT Health at Houston, Houston, TX, [6]UC Health, Colorado Springs, CO, [7]Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD

Orofacial clefts (OFCs) are the most common craniofacial malformation in humans with a substantial genetic component to etiology. About 15% of all OFC cases have a family history of OFC, supporting a contribution of inherited genetic variants. However, the remaining cases occur sporadically, suggesting a role for *de novo* mutations (DNMs). A handful of DNMs have been reported, but their overall contribution to OFCs has not been thoroughly assessed in a large sample and on a genome-wide scale. Therefore, we performed whole-genome sequencing on 643 European or Latino case-parent trios with nonsyndromic

OFCs (dbGaP accession numbers: phs001168.v2.p2 and phs001420.v1.p1, respectively) and developed a stringent filtering pipeline to generate a list of high confidence DNMs. Of the 58,735 DNMs identified genome-wide, 757 were exonic (~1.2 exonic DNMs per trio). Genes carrying DNMs in OFCs were significantly enriched for terms related to relevant biological processes (embryo development, embryo morphogenesis), mouse phenotypes (abnormal embryo morphology), and human disease (cleft). Several genes showed multiple DNMs, including known OFC risk genes (*IRF6*, N=3; *CTNND1*, N=2) and novel candidates (*ZFHX4*, N=2, *RAP1GAP*, N=2). Notably, predicted loss-of-function DNMs were found in multiple genes mutated in Mendelian OFC syndromes, but individuals in this cohort carrying these mutations lacked additional features required for clinical diagnosis of a syndrome (e.g., *CHD7*, *COL2A1*, *CTNND1*, *TFAP2A*, and *RPL5*). Additionally, we found missense DNMs in OFC candidate genes previously identified through genome-wide association studies or animal models: *FAF1*, *THADA, and SHROOM3*. Cumulatively, this study provides the strongest evidence to date for a role for DNMs in the etiology of OFC.

Whole genome sequencing of orofacial cleft trios from the Gabriella Miller Kids First Pediatric Research Consortium identifies a new locus on chromosome 21

Nandita Mukhopadhyay[1], Madison Bishop[2], Michael Mortillo[3], Pankaj Chopra[2], Jacqueline B. Hetmanski[4], Margaret A. Taub[5], Lina M. Moreno[6], Luz Consuelo Valencia-Ramirez[7], Claudia Restrepo[7], George L. Wehby[8], Jacqueline T. Hecht[9], Frederic Deleyiannis[10], Azeez Butali[11], Seth M. Weinberg[1,12], Terri H. Beaty[4], Jeffrey C. Murray[13], Elizabeth J. Leslie[2,§], Eleanor Feingold[1,12,14,§], Mary L. Marazita[1,12,15,§ *]

Orofacial clefts (OFCs) are one of the most common birth defects worldwide. Non-syndromic forms of OFCs are the most common, and their genetic component has been only partially determined. Here, we analyze whole genome sequence (WGS) data for association with risk of OFCs in European and Colombian families selected from a multicenter family-based OFC study. Part of the Gabriella Miller Kids First Pediatric Research Program, this is the first large-scale WGS study of OFC in parent-offspring trios. WGS provides more precise and richer genetic data than imputation on single nucleotide polymorphic (SNP) marker panels. Here, association analysis of genome-wide single nucleotide variants (SNV) and short insertions and deletions (indels) identified a new locus on chromosome 21 in Colombian families, within a region known to be expressed during craniofacial development. This study reinforces the ancestry differences seen in the genetic etiology of OFCs, and the need for larger samples when for studying OFCs and other birth defects in ethnically and geographically diverse populations.
dbGaP accession numbers:
European trios: phs001168.v2.p2
Colombian trios: phs001420.v1.p1

Overcoming challenges in identification, annotation, and interpretation of variants in DSD genes.

E.C. Delot [1,2]; H. Barseghyan [1,2]; A. Parivesh [1]; H. Lee [3]; E. Vilain [1,2]; the DSD Translational Research

1) Center for Genetic Medicine Research, Children's National Medical Center, Washington, D.C., USA; 2) Genomics and Precision Medicine, School of Medicine and Health Sciences, George Washington University Washington, D.C., USA; 3) UCLA Depts of Human Genetics & Pathology and Laboratory Medicine, Los Angeles, CA, USA

Whole-genome sequencing (WGS) of 94 individuals with a wide spectrum of Disorders/Differences of Sex Development (DSD) revealed challenges of variant calling and classification for several well-known DSD genes. Using two different analysis pipelines, pathogenic variants were identified in AR, SRY, WT1, NR5A1/SF1 and SRD5A2, with a 31% diagnosis rate in the non-syndromic 46,XY DSD subcohort.

Sequencing of AR proved challenging with low gene coverage and read depth, in particular next to the polyQ or polyG repeats. We were able to identify a complex frameshift variant missed by commercial clinical testing. WGS established a firm diagnosis of 5a-reductase deficiency in 4 patients with homozygous or compound heterozygous variants in SRD5A2. Two of these had undergone clinical exome sequencing, which had missed the diagnosis. We found that SRD5A2 is incorrectly annotated in ENSEMBL (but not RefSeq) as a non-coding transcript, resulting in ENSEMBL-based annotators filtering out SRD5A2 variants. We recommend reanalysis of negative exomes for patients with the appropriate phenotypes. Identification of the V89L polymorphism in SRD5A2 was also difficult because of an error of alignment to Hg19 by some platforms. Availability of larger, more ethnically diverse reference databases (such as ExAC/gnomAD) also prompted reclassification as likely benign of previously published pathogenic variants (e.g. in MAP3K1).

To assess evidence available in the databases used by laboratories to determine variant pathogenicity, we examined ClinVar variants in 69 DSD genes. For several of the long-known DSD genes, available proof was obvious, with the highest numbers of variants reported for AR, NR0B1/DAX1, CYP21A2, SRD5A2, SRY, NR5A1/SF1, or WT1. However, historical variants, with pathogenicity strongly supported by published evidence from in vitro or animal models, were often missing, resulting in gross under-estimation of available evidence for several genes (e.g. AMHR2, SRY or CYP21A2). The difficulty of interpreting whether a ClinVar-reported variant is causative of the DSD phenotype is increased for genes causing syndromic disorders where the genital phenotype is incompletely penetrant (e.g. DHCR7 in CHARGE) or where XY and XX individuals are differentially affected (e.g. ATRX). These challenges highlight the need for standardization of variant annotation tools and expert human interpretation of sequencing data in patients with rare disorders of complex etiology such as DSD.

Analysis of de-novo coding variants highlights new candidate genes for kidney malformations.

Hila Milo Rasouly, Maddalena Marasa, Byum Hee Kil, Landino Allegri, Anna Materna-Kiryluk, Giuseppe Masnata, Velibor Tasic, Marijan Saraga, Gian Marco Ghiggeri, Ali G. Gharavi

BACKGROUND: Renal hypodysplasia (RHD) is one of the most common causes of pediatric kidney failure. It includes complete absence of a kidney (agenesis) or small and dysplastic kidneys (hypodysplasia). Although multiple causative genes have been identified, they only account for 10-15% of cases. A search for de-novo variants, has led to the identification of numerous novel genes for congenital heart defects and neurodevelopmental disorders. We hypothesized that de-novo analysis can similarly identify new RHD-causing genes.
METHODS: Whole-exome sequencing was performed on 141 RHD trios without family history of kidney disorders, including 88 males and 53 females. In addition, 522 RHD singletons with were sequenced and 10,525 matched controls available through the Columbia Institute of Genomic Medicine were analyzed. The sequences were annotated using an in-house software, ATAV, and all variants were annotated using an in-house clinical pipeline. De-novo variants were identified and potential enrichment for de-novo variants was

analyzed with the denovolyzer package in R. Case-control gene-level enrichment was analyzed by including only variants well covered in both cases and controls.

RESULTS: The quality control (QC) filters used for clinical diagnosis led to the identification of a mean rate of 15 de-novo variants, including 2.8 de-novo synonymous variants. However, for the purpose of enrichment analysis for de-novo variants, the QC filters were strengthen using the QC levels of inherited variants, leading to a mean rate of 1.5 de-novo variants, including 0.33 de-novo synonymous variants. There was no significant enrichment of de-novo synonymous variants, on the other hand we observed a significant enrichment of protein altering de-novo variants (Fold Enrichment (FE) = 1.62, p-value=5.9×10^{-9}). Interestingly, the enrichment for loss-of-function variants was observed only in males (FE_{males} = 2.1, p-value=5.8×10^{-3} versus $FE_{females}$ = 1.1, p-value= 0.49). Of the 211 genes with de-novo variants, 11 genes had more than one protein altering de-novo variant. Based on the Gudmap public database, only one of those 11 genes is highly expressed in the developing kidney: *EPHB1*. Of those 11 genes, *EPHB1* is also the only one to be under constraint based on the Gnomad pLi and missense Z-scores. The identification of 2 de novo missense variants in *EPHB1* in 141 trios reached a p-value of 7.4×10^{-5} (denovolyzer). The ephrin-Eph signaling has been previously associated with kidney development (*EPHA4, EPHB2, EFNB2*). We therefore compared the prevalence of rare predicted deleterious *EPHB1* variants in cases and controls (maximum 2 heterozygotes in Gnomad; REVEL>0.5, MTR domain centile <30). We identified 3 additional conserved missense variants in the 522 additional singleton cases. In 10,525 controls, we identified 12 rare missense variants (p-value=0.03).

CONCLUSIONS: Despite limited sample size, we detected an excess of de-novo mutations in RHD and identifyed an important pathogenetic mechanism of disease. Combining trio analysis with a larger cohort of cases and controls, we identified *EPHB1* as a novel candidate gene for RHD.

Whole genome sequencing of multiplex classic Hodgkin Lymphoma pedigrees

Jamie E. Flerlage, MD; Evadnie Rampersaud, PhD; Michael Wang, PhD; Gang Wu, PhD; Mary L. McMaster, MD; Lynn Goldin, PhD; Charles Mullighan, MBBS, MSc, MD; Kim Nichols, MD; Kayla Hamilton; Jamie Maciaszek, PhD; Max Qian, PhD; Monika L. Metzger, MD1; Jun J. Yang, PhD.

Classical pediatric Hodgkin lymphoma (cHL) is rare, with documented familial aggregation and reported 2-6 fold increased risk of developing cHL for first-degree relatives of probands. Responsible pathogenetic mechanisms are largely unknown and few genomic aberrations have been described to date. In the largest sequenced cohort of multiplex clinically-homogeneous cHL families (Rotunno et al. Haematologic. 2016), one potentially causal coding variant in the KDR gene was identified in two families, leading us to speculate that additional coding variants as well as structural and noncoding variants could explain some of the genetic cause in the remaining families. Thus, through collaboration with the NCI as part of the Gabriella Miller Kids First Program, we sequenced this cohort (229 individuals in the 17 most informative families with previous WES) using a whole genome sequencing (WGS) approach. Families were eligible if one person was diagnosed with cHL ≤ 21 years of age and had another affected first-degree relative. WGS was performed by Hudson Alpha Sequencing Core with resulting ~30X coverage. All samples passed sequencing QC and mendelian inheritance checks. Data were analyzed using GATK v4, ANNOVAR, slivar, Merlin, and a suite of CNV/SV callers (e.g., CNVnator, Delly, CREST) for mapping, variant calling, annotation, filtering, multipoint linkage analysis, and identification of structural variations. Potential cHL risk alleles were identified on the basis of co-segregation with the disease phenotype, frequency in controls, predicted functional consequences, and recurrence at the variant and gene level across pedigrees. Noncoding variants were prioritized based on co-localization with putative enhancer and other regulatory regions using B-lymphocyte derived DNAse1 ATAC-seq and ChromHMM public datasets. Consistent with WES results, we identified the KDR c.3193G>A variant in 2 families and many private coding variants which could be of

interest if validated in independent studies. No additional known HL gene coding variant was identified. We also found 98 additional coding variants in 65 genes (6 loss of function) that were shared across ≥ 2 families which warrant further investigation. Among these are variants at HLA loci, a region that has been heavily and consistently implicated by GWAS of HL. Additionally, non-coding variants around known HL gene regions were discovered and are currently undergoing in-silico QC validation. CNV and SV analyses are currently underway to enable a comprehensive picture of HL genomics in families.

Expanded Ewing sarcoma cohort for tumor genomics and association with DNA repair deficiencies, clinical presentation, and outcome

Young, E., O'Brien, S., Gardiner, J., Maese, L., Robinson, R., Yandel, M., Cannon-Albright, L., Lessnick, S., Randall, L., Jones, K., Putnam, A., Wright, J., Spraker-Perlman, H., Spector, L., Lupo, P., Hayes, L., **Mitchell, G**., Schiffman, JD.

In our initial Gabriella Miller Kids First X01 award, we studied a cohort of germline DNA samples from the Project GENESIS cancer epidemiology study (Genetics of Ewing Sarcoma International Study, COG AEPI10N5), which included whole genome sequencing on 329 ES trios (patient-mother-father) plus 123 individual ES patients, with 327 ES patients from the Children's Oncology Group (COG). Our results from this study increase the number of variants in DNA repair genes associated with ES. Additionally, family history in patients with ES suggest both known and novel cancer syndromes associated with ES risk, including identifying rare, but definitive cases of familial ES. Further analysis of microsatellite and GWS results are ongoing, in addition to combined analyses with family history and other collected epidemiological variables. However, our data thus far suggest ES cases and family members could benefit from germline cancer predisposition testing.

For this expanded study to explore ES tumor genomics, we will utilize the COG ES biobank and our colleagues at HudsonAlpha and St. Jude to request and submit for sequence analysis the available tumor pairs to the 185 COG ES germline trios plus an additional set of paired COG germline-tumor samples to analyze the largest cohort of approximately 500 ES germline-tumor pairs for deep sequencing.

We will test the hypotheses that germline DNA repair deficiencies (as determined by DNA repair gene variants and germline rates of de novo alterations) will contribute to specific tumor genomic features, that these DNA repair deficiencies reflected in the germline will correlate with clinical features of, as well as test for any correlation between our measured genomic features and clinical presentation, and that the same germline and tumor genomic features will correlate with clinical outcome of the ES patients as reflected in event-free survival and overall survival.

This expanded GMKF X01 will build on our initial findings and will allow us to determine the prevalence and clinical significance of ES-like tumors that previously were included in ES biological and clinical trials. This study represents the largest and most comprehensive ES genomic analysis of its kind, and builds 00upon the successful sequencing of previous ES germline samples through the GMKF X01 program.

Discovering the genetic basis of neuroblastoma initiation and progression

Zalman Vaksman[1], Emily Blauel[1], Alex L. Lee[1], Laura E. Egolf[1], Rebecca Kaufman[1], Komal S. Rathi[1], Apexa Modi[1], Khusbhu Patel[1], Alvin Farrel[1], Jo Lynne Rokita[1], Patricia V. Basta[2], Andrew F. Olshan[2], John M. Maris[1], Sharon J. Diskin[1]

[1] Children's Hospital of Philadelphia, Philadelphia, PA, USA
[2] University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

Neuroblastoma (NB) remains one of the deadliest childhood cancers. Our long term goal is to improve the outcome of children diagnosed with NB by characterizing the germline and somatic events driving tumorigenesis so that rational, evidence-based therapies can be developed. Our objective here is to perform a primary integrative analysis of the Gabriella Miller Kids First (GMKF) X01 HL136997 NB cohort to identify both inherited and acquired genetic alterations and mutational signatures in NB that may be exploitable for risk prediction and/or therapeutic intervention. Through this research program, we have sequenced 2,277 individual samples and generated 2,655 DNA/RNA sequences. These data include robustly-annotated patient-parent triads/dyads (n=593) and matched tumor DNA (n=366) and RNA (n=228) sequencing. Our central hypothesis is that both inherited and acquired coding and non-coding mutations influence NB initiation and progression. Here, we will test our hypothesis and accomplish our objective in two specific aims: Aim 1) Identify and assess heritability of pathogenic coding and non-coding variants in NB triads and dyads. Rare coding and non-coding single nucleotide variants (SNVs), indels, and structural variants (SVs) will be investigated through a well-developed computational pipeline which we will now extend to the study of triad and dyad data to determine heritability. Pathogenicity will be assessed using multiple algorithms along with the integration of non-coding RNA and epigenetic data from in-house ChIP-seq, ATAC-seq and Capture C data together with public (e.g. ENCODE, Epigenomic Roadmap) to evaluate non-coding variation. Gene-based gTDT, burden and enrichment testing will be performed. Findings will be further evaluated in our parallel pan-childhood cancer germline study of 4,573 children spanning major cancer subtypes and also in children with NB and structural birth defects included in X01 HL140554. Aim 2) Perform integrative tumor-normal analyses to elucidate functional relevance of genetic risk factors. Tumor DNA and RNA sequencing data will be processed through our pipelines designed to evaluate SNV, indels, SVs, and copy number in DNA and account for the study of isoforms, fusions and other novel transcripts in RNA. Next, we will perform an integrative in silico evaluation of recurrent germline events using matched tumor DNA and RNA sequencing and extending to the 1,180 paired germline-tumor samples profiled in the NCI-TARGET project. Mutational signatures will be evaluated, and epistatic interactions assessed. By integrating the NB X01 cohort with extant NB (epi)genomics data, we expect to catalyze our understanding of the genetic basis of NB, with insights here being applicable to the genetic basis of other childhood conditions. Moreover, we will address, for the first time, whether NB genetic risk factors are inherited or acquired de novo. Completion of this project will have a sustained and positive impact on the field by identifying clinically actionable genetic alterations in this important childhood cancer.

Genetics at the Intersection of Childhood Cancer and Birth Defects

Patrick Sleiman, Hakon Hakonarson
Children's Hospital of Philadelphia

Evidence of a connection between childhood cancers and birth defects comes from three major sources: clinical observations of syndromes, registry linkages, and case-control studies. These studies demonstrate

that children with a variety of birth defects have a significantly increased risk of developing several types of childhood cancers. However, due to the sparsity of cases, few risk factors have been consistently confirmed for specific types of birth defects and childhood cancers, and the etiology of most of these entities remains unexplained. This proposal will leverage the unique resources of The Center for Applied Genomics (CAG) at The Children's Hospital of Philadelphia (CHOP) which houses the largest genomic facility/pediatric biobank in the US. We have identified 1,205 pediatric cancer patients that were also diagnosed with a birth defect from the CAG biobank. All have banked DNA samples from peripheral blood that are ready for sequencing together with age, sex and ethnically matched controls. The patients are from diverse backgrounds and the majority of them authorize re-contact. This study will utilize two complementary analytical approaches to disease gene discovery. Patients with parental sequences will be analyzed as trios in a typical winnowing variant prioritization approach. We also propose to sequence matched controls for each of the cases allowing for powerful statistical case control approaches, namely burden tests, to be applied to the dataset. Two strengths of this study design are the large sample sizes for what are rare phenotypes and the combination of birth defects and childhood cancers in all cases which are more likely to be burdened with low frequency variants that confer risk and that more impactful variants are more likely to be discovered. Birth defects and childhood cancer share biological pathways that are important for cell growth and division. We propose that sequencing pediatric patients suffering both conditions will allow us to discover the underlying genes and in turn advance our understanding of the causes of these devastating diseases.

Genomic Analysis of Congenital Heart Defects and Acute Lymphoblastic Leukemia in Children with Down Syndrome

Philip J. Lupo, Karen R. Rabin, Jun J. Yang, Tracie C. Rosser, Stephanie L. Sherman

Down syndrome (DS), which occurs due to trisomy 21, is one of the strongest risk factors for both congenital heart disease (CHD) and acute leukemia. For instance, children with DS have a 2000-fold increased risk of atrioventricular septal defects (AVSD) and a 20-fold increased risk of acute lymphoblastic leukemia (ALL). An important and innovative aspect of the Kids First program is understanding the overlap between structural birth defects and childhood cancer. Notably, the background of DS predisposes children to both of these phenotypes, however, the genomic architecture of risk remains largely undiscovered. In this joint Kids First/TOPMed/INCLUDE project, we are currently including 2,816 samples for whole-genome sequencing, including: 1) 408 paired germline-tumor samples from children with DS-ALL; and 2) 2000 samples from families with DS-CHD. This work will advance our understanding of the developmental pathways that may lead to both structural birth defects and childhood cancer. The *objectives of this study* are to determine the genetic variants underlying AVSD and ALL risk in children with DS. Our *central hypothesis* is that risk-associated genetic variants in the background of DS lead to a higher penetrance of AVSD and ALL. Our *secondary hypothesis* is that rare variants explain a significant proportion of the increased risk of AVSD and ALL in children with DS. Our hypotheses are supported by our previous work indicating: 1) previously identified susceptibility loci in ALL genes (e.g., *CDKN2A*) have stronger effects in children with DS-ALL compared to non-DS-ALL; 2) common genetic variants and copy number variants do not explain the increased risk of AVSD among those with DS; and 3) there is an increased burden of rare variants among children with DS-AVSD compared to those with DS alone. Therefore, the aims of our study are: 1) compare whole-genome sequencing (WGS) data between children with documented DS-AVSD and

children with DS who have structurally normal hearts to identify genetic variants that perturb heart development; and 2) compare WGS data between children with documented DS-ALL and children (from Aim 1) with DS who do not have a known history of ALL. We will also examine associations between germline mutations and somatic genomic features. This study will address the fundamental question of why children with DS have elevated risks of AVSD and ALL. Insights into the genes that drive DS-AVSD and DS-ALL may have implications for improved genetic counseling, surveillance, clinical management, and treatment strategies for these children. Additionally, our findings may inform targeted therapies or interventions for children without DS who are at risk for structural birth defects and cancer.

Genomic Analysis Of Pediatric Rhabdomyosarcoma

Bailey A. Martin-Giacalone, He Li, Danielle L. Mitchell, Douglas S. Hawkins, Stephen X. Skapek, Michael E. Scheurer, Aniko Sabo, Sharon E. Plon, **Philip J. Lupo**

Rhabdomyosarcoma (RMS) is one of the deadliest childhood cancers. In addition to low five-year survival rates for intermediate- to high-risk groups (43-67%), the use of aggressive therapies often results in severe toxicities regardless of risk group. This tumor is thought to originate in skeletal muscle-like cells, and traditionally, histology has governed diagnosis and treatment. However, there is evidence to support that cytogenetics and molecular signatures of RMS tumors should dictate risk stratification and therapy protocols, as variation in the underlying genetics yield significant differences in clinical outcomes. To this end, there is a great need to understand the etiology and molecular features of this neoplasm that can be leveraged to predict clinical outcome. Through the Gabriella Miller Kids First Pediatric X01 project, we will conduct whole genome sequencing (WGS) on 1,376 germline samples including: 1) 258 RMS families from our ongoing Children's Oncology Group (COG) study titled "Genetics of Embryonal and Alveolar Rhabdomyosarcoma Study" or GEARS; and 2) 600 independent RMS probands enrolled on other COG trials or biobanking protocols. The objectives of our study address unanswered questions about RMS by determining the role of DNMs in known cancer predisposition genes and in novel susceptibility genes. We hypothesize that highly penetrant DNMs may underlie several childhood cancers, including RMS. Our efforts will contribute to understanding how genetic variation can influence RMS tumorigenesis, which may guide current risk stratification. Although there are significant studies that support the role of somatic mutations in RMS, *much less is known in relation to germline genetic susceptibility to RMS*. For instance, based on smaller clinic-based studies, about 7% of RMS cases are thought to be associated with known cancer predisposition genes. However, there have been no population-based assessments to support this estimate; much work remains in understanding the causes of the other 95% which appear to be sporadic. Further, some research efforts have led to insights into the role of de novo germline mutations (DNMs) in the etiology of seemingly sporadic diseases, but none have explored how these variations may influence pediatric RMS. Formulated on the basis of this evidence and our preliminary studies, we propose the following two specific aims: 1) identify recurrent DNMs among RMS case-parent trios; and 2) determine the prevalence of mutations in both well-established sarcoma genes and genes identified with recurrent DNMs among children with sporadic RMS. Ultimately, the findings from this study could lead to: 1) improved genetic testing and counseling strategies in RMS patients, 2) advanced surveillance and chemoprevention protocols, and 3) the identification of novel therapeutic targets for this highly fatal tumor.

Whole Genome Sequencing and Analysis of the BASIC3 Childhood Cancer Cohort

Owen Hirschi[1]; Saumya D. Sisoudiya[1]; Ryan Zabriskie[2]; Christine Eng MD[1]; Donna Muzny, MSc[3]; David Wheeler, PhD[1,3,4]; Richard Gibbs, PhD[1,3]; D Williams Parsons, MD, PhD[1,2,3,4]; Sharon E. Plon, MD, PhD[1,2,3,4]

[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas
[2]Texas Children's Cancer Center, Department of Pediatrics, Baylor College of Medicine, Houston
[3]The Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas
[4]The Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas

In the Plon laboratory, we are investigating the genetic basis of inherited susceptibility to pediatric cancer. Our research has been enhanced by the availability of whole-genome sequencing through the Gabriella Miller Kids First X01 process. Our project included submitting samples from a cohort of pediatric cancer patients and their parents who were enrolled in the NIH supported Baylor Advancing Sequencing in Childhood Cancer Care (BASIC3) study. The BASIC3 Kids First cohort (n=120) consists of unselected racially and ethnically diverse pediatric cancer patients sequentially diagnosed with solid tumor (CNS and non-CNS) cancers and their unaffected parents. Prior analysis was limited to exome sequencing of probands. Through Kids First we have now obtained high quality germline whole genome sequence data from BASIC3 subjects including proband/parent trios (n=63) parent/child duos (n=52) and singletons (n=5) for further susceptibility gene discovery. In order to discover novel links between genetics and cancer predisposition we have analyzed *de novo* single nucleotide variants, structural variants, putative pathogenic variants in known cancer genes that may have been missed in previous clinical exome studies and genes with recurrent mutations. All analysis has been done on the cloud-based computing platform Cavatica provided by Seven Bridges Genomics, which is funded by the NIH for the storage and analysis of all Kids First datasets. From our trio analysis we have seen an average of 106 *de novo* single nucleotide variants per child with 0-4 occurring in a coding region. We have identified several mutations occurring in known cancer predisposition genes directly related to the child's cancer such as *WT1* and Wilms tumor (previously reported), as well as genes with de novo variants that may play a role in the child's cancer such as *EP300* and Rhabdomyosarcoma. Data derived from this analysis informed by analysis of the larger Kids First datasets and subsequent functional assays will provide substantial new data to define the underlying genetic structure of cancer susceptibility to pediatric cancer.

Whole Genome Sequencing of Nonsyndromic Craniosynostosis

Araceli Cuellar[1], Krithi Bala[1], Cristina Justice[2], Inga Peter[3], Ethylin Wang Jabs[3], Alexander Wilson[2], **Paul Romitti[4#]**, Simeon A. Boyadjiev[1]
[1]Department of Pediatrics, University of California Davis, Sacramento, CA; [2]Genometric Section, Computational and Statistical Genomics Branch, NIH/NHGRI, [3]Department of Genetics and Genomic Sciences, Mount Sinai, New York, NY, [4]Department of Epidemiology, University of Iowa, Iowa City, IA

Craniosynostosis (CS), the premature fusion of one or more cranial sutures, is a common, major structural birth defect occurring in about 1 in 2,500 live births. About 85% of infants with CS present with nonsyndromic craniosynostosis (NCS) without associated birth defects or developmental delays. NCS is a heterogeneous condition with presumed multifactorial etiology and its causes remain largely unknown. Primary prevention strategies for NCS are limited. Our International Craniosynostosis Consortium (ICC; https://genetics.ucdmc.ucdavis.edu) has contributed to the understanding of the genetic etiology for NCS by successfully conducting the first genome-wide association studies (GWAS) for sagittal NCS (sNCS) and metopic NCS (mNCS). We identified robust associations to loci near *BMP2, BMP7,* and *BBS9*, biologic

plausible genes involved in skeletal development. An additional GWAS of over 600 coronal NCS (cNCS) case-parent trios is in progress. Additionally, Timberlake et al. (2016) performed whole exome sequencing (WES) on a cohort of sNCS, mNCS, or combined NCS cases and *SMAD6* mutations were found in 7% of the probands. Importantly, 14 out of 17 NCS cases with *SMAD6* mutations also had the risk C allele at SNP rs1884302 downstream of *BMP2*, suggesting a two-locus inheritance model. This discovery of an epistatic interaction between *BMP2* and *SMAD6* through use of GWAS and WES approaches explains only a small proportion of all NCS cases. Along with the data generated from our completed and ongoing GWAS, we believe that whole genome sequencing (WGS) is the next important step towards identifying causal variants in NCS cases because it has the power to discover both rare and common variants in coding and noncoding regulatory regions that could be missed by other high-throughput technologies. We hypothesize that WGS will identify novel genetic factors beyond those identified with GWAS and WES that contribute to the etiology of NCS.

In this study, we will sequence 300 case-parent trios (100 cases each with sNCS, cNCS, and mNCS) and 20 multiplex families (11 with sNCS and 9 with mNCS) using WGS for discovery of all types of germline variants (de novo and inherited SNPs, insertions/deletions, and structural variations). Because differences in gene expression have been reported between fused suture and unfused suture specimens from the same individual, indicating somatic mutations may contribute to the development of NCS, we will sequence 20 paired blood-derived and bone-derived DNA specimens obtained from various NCS probands for detection of somatic mutations. Our discovery specimen repository represents one of the largest collections compiled, and along with our extensive collection of independent specimens for future replication studies, represents an unparalleled resource for studying the genetic etiology of NCS. Given our past accomplishments, experienced interdisciplinary research team, and substantial resources, we are well-positioned to successfully complete the proposed research and provide critical insights into the multifactorial etiology of NCS.

Genomic Etiologies of CHARGE Syndrome, Related Conditions and Structural Anomalies

Jacob Ogle[1,2], Amanda Moccia[2], Jennifer M. Skidmore[1], Stephanie L. Bielas[2] and **Donna M. Martin**[1,2].
Departments of Pediatrics[1] and Human Genetics[2] The University of Michigan, Ann Arbor, MI 48109

Identification of the genetic basis of developmental disorders can provide a better understanding of the underlying pathogenesis, as evident from the discovery of *CHD7* variants in CHARGE syndrome. CHARGE Syndrome (Coloboma of the eye, Heart Defects, Atresia of the choanae, Retardation of growth and development, Genital abnormalities including pubertal delay and infertility, Ear abnormalities with deafness and vestibular disorders) is a multiple anomaly condition that affects a wide variety of organ systems. However, *CHD7* genetic testing does not yield a pathogenic variant for all individuals with CHARGE, suggesting additional genetic or environmental mechanisms. Our initial studies demonstrate that pathogenic variants in genes associated with other Mendelian disorders also occur in individuals with CHARGE-like phenotypes. Individuals with CHARGE Syndrome often exhibit variable expressivity and reduced penetrance of clinical features. We hypothesize that genetic evaluation of individuals with CHARGE and CHARGE-like features will help determine how non-coding regions of the genome contribute to CHARGE. To test this hypothesis, we have generated a cohort of deeply clinically phenotyped individuals with CHARGE Syndrome and related disorders and structural anomalies whose genetic testing (chromosomal microarray, single gene sequencing, next generation panel sequencing, or exome sequencing) was negative. Individuals in our cohort exhibit clinical CHARGE-like features, including structural birth defects affecting craniofacial, ocular, neurosensory, brain, heart, mediastinal, renal, genitourinary, and skeletal organs. Our cohort includes affected and unaffected family members who consented to clinical and research genetic testing and donated blood samples for DNA

and RNA isolation and sequencing. Identification of novel pathogenic genetic variants and contributing modifier alleles within the coding and non-coding portion of the genome of these individuals will improve genetic diagnosis and provide important insights toward understanding the developmental mechanisms of CHARGE and CHARGE-related structural birth defects.

Discovery of the Genetic Basis of Structural Heart and Other Birth Defects

Christine Seidman, **Bruce Gelb,** and the Pediatric Cardiovascular Genetics Consortium (PCGC)
Icahn School of Medicine at Mount Sinai
The Pediatric Cardiovascular Genetics Consortium

The Pediatric Cardiovascular Genetics Consortium (PCGC) proposes to define genetic causes for congenital heart defects (CHD) as part of the Gabriella Miller Kids First Pediatric Research Program. CHD is the most common birth defect and is often accompanied by another congenital anomaly (CA). The PCGC has recruited and clinically characterized ≥ 10,000 CHD probands and parents (CHC trios), including 30% probands with CHD + CA. From extensive exome sequence (WES) analyses in over 2800 CHD trios, genome sequence (WGS) analyses of 350 CHD trios, and other genetic studies, we identified a substantial enrichment of damaging de novo mutations in developmental genes that modulate embryonic transcription. Based on these discoveries, we hypothesize that PCGC probands with uninformative genomic analyses (WES-negative) carry mutations in critical regulatory elements that participate in developmental expression of cardiac genes. To identify these etiologies, we propose analyses of WGS in 800 prioritized WES-negative CHD trios that include probands with banked CHD tissues (n=278), one damaging variant in a recessive CHD gene (n=186), and older fathers (n=60; age>45). We will capitalize on existing RNAseq data from CHD tissues, DNA methylation studies and the extensive computational and functional data on cardiac enhancers provided by our collaborating investigators, to analyze coding and non-coding, SNVs and SVs. We will use existing resources and capabilities of the PCGC and its companion consortium in the Bench to Bassinet Program, the Cardiovascular Development Consortium, to perform confirmatory functional genomics studies using cell and animal models outside of the GMKF program. We expect that these studies will provide novel insights into the molecular basis for birth defects and fundamental knowledge about genes and pathways involved in the development of the heart and other organs. Our aims are to:
1. Define de novo and transmitted variants, both SNVs and SVs, that cause dominant, recessive, and sporadic CHD ± CA.
2. Identify pathogenic de novo and transmitted variants in coding and regulatory regions both by case-control analyses and orthogonal data sets (ENCODE, cardiac enhancers, promoters, and regulatory ncRNAs, genes with unexplained loss of expression or allelic-specific expression in CHD tissues, and genome-wide DNA methylation data).

Hear-'n-SEQ: Sequencing Kids First for Hearing

Cynthia Morton, Jun Shen, and the Hear-'n-SEQ Consortium
Brigham and Women's Hospital

Approximately 1 in 500 babies are born with hearing loss of 40 decibels or more and 1 in 100 children will lose significant hearing by school age, making it one of the single most common structural

defects affecting the pediatric population. Hearing loss can affect a child's ability to develop speech, language, cognitive and social skills. The earlier a child with hearing loss starts receiving appropriate medical and educational services, the more likely they are to reach their full potential. More than half of early hearing loss is due to genetic factors. While the majority of prelingual hearing loss is nonsyndromic, over 400 syndromes have been described that have hearing impairment as a component. It is critically important to identify the etiology of hearing loss for many reasons, as there may be important health surveillance implications, particularly with syndromic causes. Genetic testing is available for congenital hearing loss, but the current standard of care is by no means comprehensive because: 1) many types of genetic variants in known hearing loss genes are not detectable by clinical testing, and 2) it is estimated that more than 100 hearing loss genes are as yet unknown.

With this proposal called Hear-'n-SEQ, we will leverage the resources of the NIH Common Fund's Gabriella Miller Kids First Pediatric Research program to "seek-out" the genetic etiology of childhood hearing loss through comprehensive phenotypic and genomic analyses in an international cohort of hearing impaired patients. By sharing both the clinical and sequence data with the pediatric research community, we will be empowered to identify genetic pathways that underlie hearing loss as well as pathways shared with other pediatric conditions. This project will be coordinated through the Harvard Medical School Center for Hereditary Deafness (HMSCHD). The Specific Aims of the project are to: (1) build an international consortium to identify and collect well-curated patient clinical information and DNA samples from children with hearing loss and their parents (trios) or carefully selected multiple affected individuals based on the pedigree structure, (2) submit appropriate DNA samples for whole genome sequencing at an NIH-supported sequencing center, and, (3) identify the genetic etiology of hearing impairment in individuals where possible, and integrate the data collectively into a shared data resource.

Because of the tremendous genetic heterogeneity inherent in hearing loss, the proposed international collaboration will produce a maximum yield of diverse genetic causes, as it has been well established that different populations segregate distinct concentrations of hearing loss alleles. Therefore we will sample the hearing impaired pediatric populations of parts of Asia (Hong Kong), Europe (Italy and the Netherlands), the Middle East (Turkey), and the US (individuals of European, African American, Central American and Caribbean descent). In addition to identifying novel etiologies for hearing loss, ultimately this work is designed to help create a pipeline for routinely integrating genomic sequencing into clinical diagnostics, generating more refined diagnostic capabilities, and ultimately more targeted therapies or interventions for children with hearing loss.

De novo single gene variants associated with Omphalocele, Exstrophy, Imperforate anus, Spinal defects (OEIS)

Angie C. Jelin[1,2], Nara Sobreira[2,3], Joan Ko, Elizabeth Wohler[2,3], Karin Blakemore[1,2], Heather Di Carlo[4], John Gearhart[4], David Valle[2,3]

[1]*Department of Gynecology and Obstetrics, Johns Hopkins School of Medicine*
[2]*McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine*
[3]*Department of Pediatrics, Johns Hopkins School of Medicine*
[4]*Department of Pediatric Urology, Johns Hopkins School of Medicine*

Background: OEIS, a subset of Bladder Exstrophy Epispadias Complex, is a severe fetal anomaly associated with genital, urinary and spinal defects that occur during early embryogenesis. A heterogenous

etiology has been speculated, however the high rate of infertility suggests *de novo* single gene mutations could be largely causative. Candidate genes have been proposed by one prior sequencing study.

Objective(s): We sought to identify *de novo* single gene variants associated with OEIS.

Methods: Pediatric urology patients with OEIS were clinically recruited and consented. Genomic DNA was extracted from the peripheral blood or saliva of trios (probands, mothers and fathers). Whole exome sequencing was performed on exonic and flanking intron regions with the Agilent SureSelect Xt kit and paired end 100bp reads with the Illumina HiSeq2500 platform. Mutations with an internal or ExAC allele frequency of greater than 1% were removed, as were those with a depth <10. Future research will involve interrogation of intronic regions via whole genome sequencing through the Gabriella Miller Kids First Pediatric Research Program.

Results: Analysis of WES lead to identification of 7 rare *de novo* single gene mutations in 7 OEIS trios A frameshift mutation was identified in the *HSD11B1*. The remaining mutations were missense single nucleotide variants in *ZNF280B, TCP11L1, FANCE, RPL4, EIF6, WDR81*. One proband had 2 de novo mutations and 1 proband had none. Each mutation was unique to a specific proband.

Conclusions: Our etiologic investigation of OEIS by WES resulted in identification of several d*e novo* single gene mutations arising in affected probands. Additional function investigation is necessary to prove causation. This work lead to the Gabriella Miller Kids First proposal for whole genome sequencing for cases of Bladder Exstrophy Epispadias Complex (BEEC).

Genomics of Orthopaedic Disease Program

Carol A. Wise [1,2,3,4], Anas Khanshour[1], Diane Sepich[5], Chandresh Patel[6], Yared Kidane[1], Richa Singhania[1*], Nadja Makki[7], Ryan Gray[8], Nadav Ahituv[9,10], Lila Solnica-Krezel[5], Christina Gurnett[11,12], Jonathan Rios[1,2,3,4]

[1]Center for Pediatric Bone Biology and Translational Research, Texas Scottish Rite Hospital for Children, 2222 Welborn St., Dallas, TX 75219, USA

[2]McDermott Center for Human Growth and Development, [3]Departments of Pediatrics, [4]Orthopaedic Surgery, [6]Radiology, University of Texas Southwestern Medical Center, Dallas, TX 75235, USA

[5] Departments of Developmental Biology, [11]Neurology and [12]Orthopaedic Surgery, Washington University School of Medicine, St. Louis, MO 63110

[7]Department of Anatomy and Cell Biology, University of Florida Health, Gainesville, FL 32610

[8] Department of Nutritional Sciences and Pediatrics, Dell Pediatric Research Institute, Austin, TX 78723 USA

[9]Department of Bioengineering and Therapeutic Sciences, and [10]Institute for Human Genetics, University of California San Francisco, San Francisco, California, 94158, USA.

*present address Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10065

Pediatric birth defects carry significant morbidity and mortality. At Texas Scottish Rite Hospital for Children, the Genomics of Orthopaedic Disease (GOOD for Kids) program seeks to understand pediatric birth defects such as adolescent idiopathic scoliosis (AIS). The genetic contributions of AIS are complex. Through an NIH-funded Research Program Project (P01 HD084387) our goals are to identify AIS candidate susceptibility loci and genes by genome wide association studies (GWAS), which have highlighted perturbation in cartilage biogenesis pathways, and by forward genetic screens in zebrafish. In a complementary strategy, we have partnered with the Gabriella Miller Kids First (GMKF) initiative to identify dominantly inherited, high-impact AIS mutations by whole-genome sequencing (WGS) in 72 families. Screening 42 genes previously implicated in AIS by GWAS, sequence-based approaches, or animal modeling identified 10 genes with two or more segregated mutations per gene. We found that one gene, CELSR2 encoding cadherin EGF LAG Seven-Pass G-Type Receptor 2, was enriched for rare protein-altering variants in an independent cohort of 622 AIS exomes compared to gnomAD (P =5.1E-04, OR=1.48). To identify novel AIS candidate variants, we prioritized genes involved in biogenesis and homeostasis of cartilaginous tissues, or as identified from animal models for further analysis. This identified candidate missense mutations in the SSPO and HAPLN1 genes.

SSPO encodes sco-spondin, a large secreted glycoprotein, inactivation of which in zebrafish produces a curled down body axis, and partial inactivation results in scoliosis. HAPLN1 encodes cartilage link protein (CRTL1), an extracellular matrix protein that stabilizes hyaluronic acid binding to aggrecan within cartilaginous tissues Four additional HAPLN1 missense mutations were identified in the 622 AIS exomes. One mutant protein (CRTL1 p.C304S) that was predicted to misfold by molecular modeling exhibited secretion abnormalities in cell-based assays. To test the role of human HAPLN1 and SSPO mutations in vivo, we are partnering with GMKF and Knockout Mouse Project (Kids First-Komp2 Project) to characterize mouse knockins of these mutations created by CRISPR mutagenesis. Our goal is to develop a mechanistic understanding of AIS, and together with the Kids First initiative, to expand our understanding of pediatric musculoskeletal disease.

Detection of novel genetic bases for congenital cranial dysinnervation disorders (CCDDs) by whole genome sequencing.

Julie Jurgens*, Arthur S. Lee*, Eleina England, Silvio Alessandro Di Gioia, Brenda J. Barry, Wai-Man Chan, Alysia K. Lovgren, Harrison Brand, Xuefang Zhao, Katherine R. Chao, Anne H. O'Donnell-Luria, Michael E. Talkowski, Daniel G. MacArthur, Elizabeth C. Engle
*Presenting authors

Congenital cranial dysinnervation disorders (CCDDs) are defined by defective innervation of cranial nerves or nuclei and affect 0.1% of individuals worldwide. Although many CCDDs appear to be inherited, their genetic basis remains largely unknown. Through the Gabriella Miller Kids First Foundation, we conducted whole genome sequencing (WGS) of 899 individuals from 270 families with various CCDDs, with a focus on ocular and/or facial motor function. We have generated a list of 440 novel candidate causal variants in 348 genes involved in motor neuron development and/or axon growth or guidance. A few classes of candidate genes encode proteins such as semaphorins, ephrins, mediator complex subunits, neuron navigator proteins, microtubule severing proteins, nerve growth factors, netrin receptors, and chromatin remodelers. Of the 348 candidate genes, 34 are mutated in >1 family in our WGS cohort and 212 have >60% amino acid level identity in fish. Of the variants, 197 missense residues are conserved in fish at the amino acid level; 115 single nucleotide variant (SNV) sites are conserved in fish at the nucleotide level. 118 genes have putatively LOF variants including frameshifting indels, nonsense, splice site, biallelic missense, and conserved 5'UTR variants. 83 of these genes have biallelic variants predicted to result in partial or complete LOF, while 35 have monoallelic variants predicted to result in haploinsufficiency. Because this list is too extensive to conduct modeling for all candidate genes and variants, we are implementing a CRISPR/Cas9-mediated screening and modeling pipeline in zebrafish and in mouse embryonic stem cell-derived motoneurons.

In addition to identifying and prioritizing coding SNVs/indels, we have performed extensive preliminary work to develop a framework for interpreting coding and non-coding structural variants. We have performed first-pass CNV calling QC, and segregation analysis across our entire GMKF cohort, using WGS split-read and read-depth based callers, including Manta, GenomeSTRiP, and gCNV. We have also performed array-based CNV calling on 40 syndromic DRS trios, a subset of which overlap our WGS cohort. We confirm that *de novo* and inherited CNVs in DRS cases are enriched using this orthogonal technology and were able to validate specific candidate SVs called by both platforms. Based on these preliminary data, we have already solved and published one GMKF case through the identification of a truncating deletion over the novel disease gene *MACF1* and potentially solved several additional cases by

querying known or predicted disease genes. In order to further increase our sensitivity to detect pathogenic variants, we are currently performing ensemble SV calling across a battery of SV callers and joint SV genotyping against the > 20,000 WGS samples from the gnomAD database.

Finally, we are partnering with Kids First, KOMP2, Dr. Kent Lloyd at UC Davis, and Dr. Steve Murray at JAX to generate three knock-in mouse models for two high-priority candidate SNVs and a structural variant identified through WGS. Cumulatively, we expect that these approaches will enable prioritization of candidate genes in order to reveal new pathways in neuronal development and improve clinical counseling and outcomes for pediatric health.

Likely pathogenic de novo variants in congenital diaphragmatic hernia patients are associated with worse clinical outcomes

Lu Qiao[1,*], Julia Wynn[1,*], Lan Yu[1], Rebecca Hernan[1], Xueya Zhou[1], Vincent Duron[1], Gudrun Aspelund[1], Christiana Farkouh-Karoleski[1], Annette Zygumnt[1], Usha Krishnan[1], Shannon Nees[1], Julie Klevner[1], Foong Yen Lim[2], Timothy Crombleholme[2], Robert Cusick[3], Kenneth Azarow[4], Melissa Ellen Danko[5], Dai Chung[5], Brad W. Warner[6], George B. Mychaliska[7], Douglas Potoka[8], Amy J. Wagner[9], Samuel Soffer[10], David Schindel[11], David McCulley[12], **Yufeng Shen[1,#],** Wendy K. Chung[1,#]

[1]Columbia University. [2]Cincinnati Children's Hospital, [3]University of Nebraska College of Medicine, Omaha, NE, USA. [4]Oregon Health & Science University, Portland, OR, USA. [5]Vanderbilt University Medical Center, Nashville, TN, USA. [6]Washington University, St. Louis Children's Hospital, St. Louis, MO, USA. [7]University of Michigan School of Medicine, Ann Arbor, MI, USA. [8]University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [9]Medical College of Wisconsin, Milwaukee, WI, USA. [10]Northwell Health, Great Neck, NY, USA. [11]Children's Hospital Dallas, Dallas, TX, USA. [12]Meriter-Unity Point Health, Madison, WI, USA. [#]Contact: Y.S. (ys2411@cumc.columbia.edu) and W.K.C. (wkc15@cumc.columbia.edu).

Purpose: Congenital diaphragmatic hernia (CDH) is associated with significant mortality and long-term morbidity in some but not all individuals. We hypothesize monogenic factors that cause CDH are likely to have pleiotropic effects and be associated with worse clinical outcomes.
Methods: We enrolled and prospectively followed 647 newborns with CDH and performed genomic sequencing on 462 trios to identify *de novo* variants. We grouped cases into those with and without likely pathogenic (LP) variants and systematically assessed CDH clinical outcomes between the genetic groups.
Results: Complex cases with additional congenital anomalies had higher mortality than isolated cases ($P=3\times10^{-6}$). Isolated cases with LP variants had similar mortality to complex cases and much higher mortality than isolated cases without LP ($P=2\times10^{-2}$). The trend was similar with pulmonary hypertension at 1 month. Cases with LP variants had an estimated 13-16 points lower score on neurodevelopmental assessments at 2 years than cases without LP (p<1e-3), and the difference is similar in isolated and complex cases.
Conclusion: We found that the likely pathogenic genetic variants are associated with higher mortality, worse pulmonary hypertension, and worse neurodevelopment outcomes. Our results have important implications for prognosis and potential intervention and long-term follow up for children with CDH.

Metaphyseal enchondromatosis with D-2-hydroxyglutaric aciduria and variants in IDH1, IDH2 and EXT2.

Elizabeth Partan* (1,2), Lisa Kratz (5), Gary S Gottesman (4), David Valle (1,3), Nara Sobreira (1,3)

1. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD
2. Predoctoral Training Program in Human Genetics, Johns Hopkins School of Medicine, Baltimore, MD
3. Department of Pediatrics, Johns Hopkins School of Medicine, Baltimore, MD
4. Center for Metabolic Bone Disease and Molecular Research, Shriners Hospitals for Children – St. Louis, St. Louis, MO
5. Department of Neurogenetics, Kennedy Krieger Institute, Baltimore MD

Metaphyseal enchondromatosis with D-2-hydroxyglutaric aciduria (ME-HGA; OMIM 614875) is a rare disorder characterized by enchondromas and elevated levels of D-2-hydroxyglutaric acid (D-2HG). Vissers et al. (2011) reported the somatic mosaic IDH1-R132 variant in two unrelated patients with ME-HGA. Heterozygous, somatic, gain-of-function (GOF) variants of IDH1 (Arg132His, Arg132Cis, and Arg132Ser) and IDH2 (Arg172Ser) have been reported in a variety of cancers as well as Ollier disease (OD; OMIM 166000) and Maffucci syndrome (MS; OMIM 614569). OD and MS are related conditions characterized by multiple enchondromas with a ~30% risk of chondrosarcoma transformation. While wild-type IDH1 and IDH2 catalyze the oxidative carboxylation of isocitrate to α-ketoglutarate (αKG), the GOF variants cause αKG to be converted to D-2HG. The conversion of αKG to D-2HG leads to inhibition of αKG-dependent reactions promoting increased histone methylation, increased DNA methylation and impaired cell differentiation. Here, we present a 4-year-old boy with delayed tooth eruption, bilateral multiple enchondromas (shoulders, arms, elbows, wrists, hands, ribs, spine, hips, knees, ankles and feet), and D-2-hydroxyglutaric aciduria. He had a normal CGH-array. We performed WES of his blood and 2 enchondroma biopsies and identified a germline, heterozygous IDH2-T435M variant in all samples tested. We also identified a somatic mosaic, heterozygous IDH1-R132C GOF variant in both enchondromas but not in the blood sample. He also had a germline, heterozygous EXT2-T620M variant in in all samples tested. Parents' samples are now being tested to determine if the germline IDH2-T435M and EXT2-T620M variants are de novo. The IDH2-T435M is a variant of uncertain significance with a gnomAD MAF of 3.61e-3. Loss-of-function variants in EXT2 are known to cause autosomal dominant multiple exostoses (MIM133701), the EXT2-T620M is a variant of uncertain significance, predicted to be pathogenic by DANN, GERP, dbNSFP.FATHMM, LRT, MetaLR, MetaSVM, MutationAssessor, MutationTaster, PROVEAN and SIFT and a gnomAD MAF of 6.62e-4. Herein we suggest that these IDH1, IDH2, and EXT2 variants interact in this patient causing his severe multiple enchondromas phenotype with D-2-hydroxyglutaric aciduria, supporting our hypothesis that these are tumor predisposition syndromes characterized by locus heterogeneity with germline (or early post-zygotic) and additional tumor variants leading to tumor formation.

References:
1. Vissers, L.E.L.M., Fano, V., Martinelli, D., Campos-Xavier, B., Barbuti, D., Cho, T.-J., Dursun, A., Kim, O.H., Lee, S.H., Timpani, G., et al. (2011). Whole-exome sequencing detects somatic mutations of IDH1 in metaphyseal chondromatosis with D-2-hydroxyglutaric aciduria (MC-HGA). American Journal of Medical Genetics Part A 155, 2609–2616.

VHL variant in a patient with Maffucci Syndrome.

Renan Martin*, Sarah M. Robbins, David Valle, Nara Sobreira

VHL downregulates HIF1A, the main regulator of adaptation to hypoxia, by targeting HIF1A for degradation. Loss-of-function variants in VHL are known to cause von-Hippel Lindau syndrome, an autosomal dominant familial cancer syndrome, and autosomal recessive familial erythrocytosis. VHL has also been associated to other tumors such as cerebellar hemangioblastoma, pheochromocytoma and renal cell carcinoma. Maffucci syndrome (MS) is a rare disorder characterized by multiple enchondromas and vascular anomalies. Patients with MS are also at increased risk for cancer, mainly, chondrosarcomas and vascular malignancies. Here we describe a patient with MS and additional features not described before such as a region of epidermal and dermal atrophy on the left chest, a lymphoepithelioma of the nasopharynx with parapharyngeal involvement, an exostosis in the left wrist, developmental delay and intellectual disability. We performed WES on multiple samples including his blood, an enchondroma and an exostosis. A germline VHL variant (c.C505T; p.R169W) identified by WES. RNAseq confirmed that the VHL variant is present at heterozygous levels (45-70%) in RNA from all tissues tested: an enchondroma, an exostosis, a vascular anomaly biopsy and an unaffected skin biopsy. Western Blot showed that VHL in the patient's unaffected skin biopsy and vascular anomaly biopsy is within the range of the controls. WES also identified a somatic mosaic, known to be pathogenic and causative of somatic infantile myofibromatosis, PDGFRB variant (c.C1998A; p.N666K). This variant was identified in the enchondroma but not in the exostosis and in the blood. We performed digital droplet PCR and confirmed that the PDGFRB-p.N666K variant was not present in the blood, but was present in all 4 patient's enchondromas tested (21.8%, 27.5%, 50.5%, and 49.8%), in his in his exostosis (0.64%), in his unaffected skin biopsy (22%) and vascular anomaly biopsy (0.88%). Based on these findings we suggest that the VHL-p.R169W is likely pathogenic, and that, similarly to the VHL-R200W and the VHL-H191D variants that result in polycythemia, it causes overproduction of the HIF-targets by attenuating formation of the E3 ubiquitin ligase and attenuating binding of HIF1. We also suggest that this patient has a tumor predisposition syndrome characterized by the germline susceptibility VHL-p.R169W variant and additional tumor variants, such as the PDGFRB-p.N666, leading to tumor formation.

Variants in the HIF-1 pathway are associated with Ollier disease and Maffucci syndrome.

S. Robbins* 1,2; R. Martin 1; B. Cohen 3; C. Haldeman-Englert 4; M. Cernach 5; G. Gottesman 6; G. Semenza 7; D. Valle 1; N. Sobreira 1

1) McKusick-Nathans Institute in the Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 2) Predoctoral Training Program in Human Genetics, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 3) Division of Pediatric Dermatology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA; 4) Fullerton Genetics Center Mission Health, Asheville, NC, USA; 5) Escola Paulista de Medicina, Sao Paolo, Brazil; 6) Center for Metabolic Bone Disease and Molecular Research, Shriners Hospitals for Children - St. Louis, St. Louis, Missouri, USA; 7) Institute for Cell Engineering, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Ollier disease (OD, OMIM 166000) and Maffucci Syndrome (MS, OMIM 614569) are two related rare disorders characterized by multiple enchondromas, which commonly develop in the extremities around joints, resulting in deformity, leg length discrepancy, and pathological fractures. MS is distinguished from OD by the development of vascular anomalies. Both disorders are cancer predisposition syndromes; chondrosarcomas develop in ~30% of individuals. Gain-of-function variants in IDH1 and IDH2 have been

described in ~80% of the enchondromas, chondrosarcomas and vascular anomalies in patients with OD and MS. However, no predisposing germline variants have been reported to date. We hypothesize that OD and MS are tumor predisposition syndromes with germline or early post-zygotic causative variants and additional tumor variants involved in the formation of the benign and malignant tumors. Locus heterogeneity for the germline variants also seems likely. To search for causative germline variants, we performed exome sequencing of leukocyte DNA in 27 probands. We show that 9 (33%) had one or more rare variants in one of 5 genes in the HIF-1 pathway (IDH2, KDM4C, HIF1A, VHL, and EGLN1 in 3, 1, 1, 4, and 2 probands respectively). To further investigate the role of HIF-1 in the pathogenesis of OD and MS, we performed RNA-seq of fibroblast RNA from 3 individuals with variants in IDH2, KDM4C, VHL with OD and MS at normoxia and hypoxia and compared to 3 control fibroblast lines. We show that patient fibroblasts have significantly less differentially expressed HIF-1-regulated genes in response to hypoxia suggesting that the HIF-1 pathway variants in our patients lead to HIF-1 pathway dysregulation. Also, to further investigate the pathways dysregulated in patients with OD and MS, we performed RNA-seq of 4 enchondromas and 3 chondrosarcomas in comparison to a human control chondrocyte line. The genes in the KEGG pathway "proteoglycans in cancer" (KEGG:05205) were enriched in both the normoxia fibroblasts and chondrosarcomas differentially expressed gene sets; this pathway encompasses many signaling cascades, e.g. HIF-1, mTOR, and MAPK. Also, chondrocyte differentiation and cartilage development genes (HOXA11, GDF6, and PKDCC) were significantly differentially expressed in patients' fibroblasts, enchondromas and chondrosarcomas. Based on our results we suggest that dysregulation of HIF-1 and related pathways are responsible for tumorigenesis in patients with OD and MS.

Genomic Diagnostics in Cornelia de Lange Syndrome, Related Diagnoses and Structural Birth Defects

Sarah Raible[1], Sawona Biswas[1], Tiffany Hartman[1], Maninder Kaur[1], Debbie McEldrew[1], Batsal Devkota[1,2], Ian D. Krantz[1,3]

[1]Division of Human Genetics, [2]Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia, [3]Perelman School of Medicine at the University of Pennsylvania

Disorders of human morphogenesis are a major cause of human suffering for the affected individuals and their families. Congenital anomalies are identified in approximately 3% of term births, 10% of stillbirths, and in as many as 50% of first trimester spontaneous abortuses. While most, if not all, human structural birth defects have a significant genetic component, identification of genetic perturbations in isolated structural birth defects has been complicated by the complex nature of their underlying etiologies, likely involving disruption of regulatory elements that can act in a temporal and tissue specific manner, multi-gene, epigenetic and gene-environment interactions. Our approach to tease out genetic contributions to birth defects has been to identify the underlying causes of syndromic birth defects which are often Mendelian in nature and therefore lend themselves more readily to genetic causal identification. Once identified, these genetic causes of syndromic forms of birth defects can be leveraged to understand the genetic contributions to isolated birth defects seen in constellation in these syndromes. We propose to use Cornelia de Lange Syndrome (CdLS), a dominant multisystem developmental disorder consisting of a constellation of structural birth defects involving most body systems and significant growth and cognitive impairment as a prime example of this approach. We and others have shown that alterations in the cohesin and associated pathways are causative of CdLS and related diagnoses when disrupted and have more broadly been termed

"cohesinopathies" or "disorders of transcriptional regulation (DTRs)". As part of this project, genome sequence will be performed on a unique cohort of 178 probands and family members with clinically confirmed CdLS or a related diagnosis in whom molecular analysis by targeted gene sequencing, next generation sequencing (NGS) panels or exome sequencing have been negative but are strongly suspected of having an underlying genetic alteration to explain their clinical features. A total of 400 genomes will be sequenced inclusive of 96 trios, 30 duos and 52 singletons. This work will lead to the identification of genes critical in human embryonic development, provide novel insights into transcriptional regulation and help to identify genetic causes and candidate genes for isolated birth defects seen in constellation in this group of diagnoses. Most critical developmental genes are also cancer genes and the genes known to cause CdLS are no exception. CdLS is not a cancer predisposition syndrome so understanding the mutational mechanisms in these genes that lead to structural birth defects when present in the germ line and result in cancer when mutated somatically is a fundamental aspect of this research.

The Newborn Screening Translational Research Network: Advancing the Understanding of Genetic Disease in Newborns and Children Using a Shared Infrastructure of Genomic and Phenotypic Information

Amy Brower
American College of Medical Genetics

In 1961 a task force commissioned by President Kennedy recommended the establishment of a "centralized unit" charged with concentrating research on disorders of human development. In the five decades since this recommendation, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH), has supported groundbreaking efforts to help save lives, improve wellbeing, and reduce societal costs associated with illness and disability of mothers and children.

NICHD efforts are focused on maternal and child health, and their forward-looking research agenda, initiatives and subsequent discoveries in basic, translational and clinical research have been impactful across the lifespan of affected children and families. This is especially true in newborn screening (NBS) research when a 1960s NICHD led longitudinal study discovered that newborns who received a screening test for phenylketonuria (PKU), using a blood spot on filter paper taken shortly after birth, benefited from early diagnosis and treatment. This led to the adoption of nationwide screening of newborns for PKU in state based public health laboratories, and fostered years of investigations and innovations in NBS using blood based and physiological assessments of conditions with prenatal and/or neonatal onset.

Fast forward fifty-five years and the number of screened conditions has grown from 1 to 61; screening methodologies have advanced to include genomics and hospital based physiological assessments; treatment regimens now include medically complex and invasive procedures like stem cell transplant, intrathecal enzyme replacement using lumbar puncture, surgery for a cochlear implant, repair of a congenital heart defect, and genotype based therapy; and the onset of disease has expanded from the neonatal period to childhood and adult onset. Throughout the years, a better understanding of each of the 61 conditions has been possible because screening on a population basis enables unbiased ascertainment and the ability to longitudinally collect health information on each diagnosed newborn. These longitudinal natural history studies have profoundly improved knowledge of the diseases' etiology, pathophysiology, phenotypic heterogeneity, and comorbidities.

NICHD leadership and role in NBS research were formalized in the NBS Saves Lives Act of 2008 and reaffirmed in the NBS Saves Lives Reauthorization Act of 2014. This legislation authorized the Hunter Kelly NBS Research Program within NICHD to carry out, coordinate, and expand research in NBS. Through three consecutive five-year contracts, the American College of Medical Genetics and Genomics (ACMG) has led the development and implementation of a key component of the Hunter Kelly NBS Research Program, the NBS Translational Research Network (NBSTRN). The NBSTRN began as an effort to engage a variety of stakeholders across the NBS system and has matured into a dynamic and committed network comprised of researchers, public health professionals and clinicians with expertise in:

- The discovery and validation of novel technologies to screen and diagnose disease;
- The clinical care of newborns including execution of clinical trials and application of cutting edge treatments and management strategies;
- Pilots of new technologies and treatments;

- ☐ The collection, analysis and dissemination of longitudinal health and genomic data; and
- ☐ The ethical, legal and social implications of NBS research.

State-based newborn screening (NBS) programs annually screen 4 million newborns for genetic conditions that require immediate treatment to prevent morbidity and mortality. Approximately 20,000 newborns ultimately receive a diagnosis each year and enter into lifelong clinical care and management. Research that capitalizes on the NBS system offers an opportunity to improve understanding of genetic disorders in newborns and children, develop novel technologies, and study the effectiveness of new treatments. Researchers have begun to explore the use of genomics in NBS, and state-based NBS programs have begun screening for disorders that have a large amount of genetic heterogeneity, including onset of symptoms in childhood or beyond. Natural history studies that capture longitudinal information on cases resulting from this expanded NBS will establish the clinical validity and clinical utility of identified disorders, and pilots of novel NBS technologies will provide analytical validation.

Because the majority of the conditions that are part of, or candidates for, newborn screening in the United States are rare or very rare, a shared informatics system is critical. The ideal system will support the collection, analysis and sharing of genomic and phenotypic data from the basic research designed to advance understanding of the disease process, to translational research to develop technologies to screen and therapies to treat, to public health implementation of a comprehensive newborn screening program to identify newborns at risk. To support these efforts, the NBSTRN has been working to develop and implement a suite of tools, resources and expertise to translate basic research discoveries to routine clinical practice to improve health outcomes, discover novel therapies and management strategies, operationalize the use of genomic sequencing data across the lifespan and refine population-based screening and diagnosis.

We will provide an overview of key accomplishments and efforts including the use of standardized vocabularies, interactive computer systems, and robust security measures used by over 100 researchers across 33 studies. Longitudinal data from 5178 participants with an average of three clinical encounters and pilot data from 331,497 subjects have been collected. The use of the NBSTRN by researchers is creating unique cohorts of individuals with and without genetic disease and accelerating understanding of these conditions.

Accelerating research with the NCBI Sequence Read Archive on the commercial cloud

S. Sherry, A. Boshkin, J. Brister, R. Connor, D. Deriy, A. Dillman, K. Durbrow, A. Efremov, L. Fleischmann, G. Godynskiy, D. Ignatovich, A. Iskhakov, A. Johnson, M. Kimelman, A. Klymenko, A. Kochergin, C. Leubsdorf, K. McDaniel, A. Mnev, C. O'Sullivan, S. Ponomarov, W. Raetz, K. Rodarmer Jr, K. Rodarmer Sr, O. Shutov, D. Soren, A. Stine, M. Vartanian, E. Yaschenko, V. Zalunin.

The NCBI Sequence Read Archive (SRA) is NCBI's comprehensive collection of next generation sequence data from human and non-human organisms. At over 10 petabytes, SRA is NCBI's largest research archive and has exceeded the capability of most sites to replicate locally. Recognizing this, NCBI and NIH have recently replicated SRA content to both Amazon and Google commercial cloud platforms as part of the NIH STRIDES program. NCBI is providing community education, and training opportunities to build relationships with SRA users through webinars, online tutorials and a series of NCBI-hosted hackathons that bring together users to work with topical and problem-relevant subsets of SRA data.

Taken as a broad collection of sequences sampled across the tree of life, SRA data are mined for new discoveries about genomic sequence, natural variation, antimicrobial resistant genes, gene expression, methylation states, and previously undescribed genes and species, strains, or viral isolates. By size, SRA is equal parts public and controlled access data — public data includes non-human sequences and human RNAseq and genomic data where individual consent has been provided for the open and unrestricted use of their data. Controlled access data is sequence information from human research study participants supported by NIH and access is restricted and controlled through dbGaP approval protocols.

Moving SRA data to cloud platforms will benefit users by enabling meaningful and timely access to the exponentially growing corpus of SRA data. This migration of data is accompanied by improvements to the SRA Run Selector, a new data model that supports access to both submitted and normalized data formats, improvements to the SRA toolkit, and conformance to NIH's new standards for Identity and Authorization Management Researchers are invited to join the NCBI cloud user community, participate in Codeathon events, and explore our developing knowledge base of self-directed education resources to design powerful and affordable cloud-based analysis workflows. This presentation will describe new SRA resource contents, procedures for data access and use, and new cloud-related educational resources.

The INCLUDE (INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE) Project: Challenges of Building a Virtual Cohort

Melissa A. Parisi, Sujata Bardhan, Rachel Goldman, Valerie Cotton, James Coulombe
NICHD, National Institutes of Health

https://www.nih.gov/include-project
The INCLUDE (INvestigation of Co-occurring conditions across the Lifespan to Understand Down syndromE) project was launched in June 2018 in support of a Congressional directive in the fiscal year (FY) 2018 Omnibus Appropriations. The directive called for a new trans-NIH research initiative on critical health and quality-of-life needs for individuals with Down syndrome (DS). NIH dedicated almost $23 million for INCLUDE research, bolstering total funding for Down syndrome research in FY2018 to almost $60 million. NIH recently announced about $35 million in new grants in FY 2019 through the INCLUDE project. This initiative, which is led by the NIH Office of the Director and involves multiple NIH institutes and centers, aims to include people with Down syndrome in all aspects of research, such as creating opportunities for people with Down syndrome to participate in existing clinical trials and developing new opportunities through basic science studies. In addition, NIH recently sponsored a workshop to explore strategies for assembling a virtual Down syndrome cohort across the lifespan from existing and prospective studies. INCLUDE will investigate conditions that affect individuals with Down syndrome and the general population, such as Alzheimer's disease/dementia, autism, cataracts, celiac disease, congenital heart disease and diabetes. Applying the expertise and resources from multiple NIH Institutes and Centers, INCLUDE will:

Conduct targeted, high-risk, high-reward basic science studies on chromosome 21.
• Study research models of DS
• Explore the effects of multiple genes triplicated on chromosome 21 simultaneously
• Identify pathways that may be most responsive to new therapies
Assemble a large study population of individuals with Down syndrome.
• Add to or expand existing DS cohorts with 'omics data
• Develop shared databases using common data elements
• Build on the DS-Connect® registry (https://DSConnect.nih.gov) to connect families with research opportunities of interest to them

Return to top ⬆

Include individuals with Down syndrome in existing and future clinical trials.
• Bolster recruitment of people with DS in clinical trials for co-occurring conditions
• Develop new therapies for DS
• Leverage existing clinical trials infrastructure to explore differences in drug metabolism in those with DS and provide assistance in clinical trial design and training for conducting trials in people with DS

The Trans-Omics for Precision Medicine (TOPMed) Program

Matthew P Conomos and Sarah Nelson
TOPMed Data Coordinating Center, University of Washington

The Trans-Omics for Precision Medicine (TOPMed) program, sponsored by the National Institutes of Health (NIH) National Heart, Lung and Blood Institute (NHLBI), is part of a broader Precision Medicine Initiative, which aims to provide disease treatments tailored to an individual's unique genes and environment. TOPMed contributes to this Initiative through the integration of whole-genome sequencing (WGS) and other omics (e.g., metabolic profiles, protein and RNA expression patterns) data with molecular, behavioral, imaging, environmental, and clinical data.

A primary goal of the TOPMed program is to improve scientific understanding of the fundamental biological processes that underlie heart, lung, blood, and sleep (HLBS) disorders. TOPMed is providing deep WGS and other omics data to pre-existing 'parent' studies having large samples of human subjects with rich phenotypic characterization and environmental exposure data.

NHLBI's New Data Platform: Leveraging cloud technology to accelerate scientific discovery for heart, lung, blood, and sleep research

Stan Ahalt (presenting), Kira Bradford, NHLBI's Data Platform Consortium

Biomedical and genomics research represent an exponentially growing source of data that far outstrips individual researchers' ability to store and analyze. Recognizing this challenge and the need to democratize data and tool access for the scientific community, NHLBI initiated a new Data Platform in 2018. The goal of NHLBI's New Data Platform is to enable heart, lung, blood, and sleep research investigators to find, access, share, store, and compute on large data sets. NHLBI's Data Platform is a cloud-based platform providing tools, applications, and workflows to enable these capabilities in secure workspaces. NHLBI's Data Platform initially focused on opening access to TOPMed datasets, representing a high value, high return for NHLBI investigators.

NHLBI's New Data Platform is agile and iterative, defining and supporting initial scientific use cases to accelerate infrastructure development. NHLBI's Data Platform focused on two initial use cases: 1) enabling large scale genomic analyses and 2) enabling investigators to conduct Deep Learning analyses on image data. Both require addressing extensive challenges, such as handling big data, computing on a large scale, and enabling easy data access. NHLBI's Data Platform addresses these by developing solutions on a cloud-based platform utilizing technology and expertise from multiple teams.

Currently, alpha users log in to NHLBI's Data Platform with their eRA Commons ID and import datasets of interest into a secure workspace based on their dbGaP credentials. Users can access several tools to explore datasets and to gain a better understanding of cohort characteristics. Additionally, users can

conduct high powered analyses such as Genome-Wide Association Studies (GWAS) or Deep Learning on large scale datasets utilizing cloud resources. Users can publish and share their results on NHLBI's Data Platform. Users will also have access to extensive IT support for questions or suggestions for improvements.

In summary, NHLBI's New Data Platform has developed and will continue to develop many novel solutions to address biomedical and genomics research challenges. NHLBI's Data Platform utilizes cloud technology to empower big data and its computation as well as federated, centralized storage to facilitate secure data access. Additionally, NHLBI's Data Platform is working with TOPMed investigators to provide state-of-the-art, sophisticated tools for genomics and other analyses and facilitates team science, collaboration, and the ability for investigators to bring-your-own data to the platform. Thus, NHLBI's Data Platform significantly accelerates research discovery.

NCI Cancer Research Data Commons

Kim, E.[1] and the NCI CRDC Team[1]

[1]Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health, Bethesda, MD

Progress in basic and clinical cancer research requires access to well-annotated, multi-modal datasets to identify molecular underpinnings of disease. NIH and NCI have supported numerous programs to generate such datasets, including The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Atlas Consortium (CPTAC), to be leveraged by the research and clinical communities to better understand the causes of cancer and to inform the development of better treatments and prevention tools. However, researchers are still limited in their ability to draw insights and meaningful interpretations from disparate datasets by challenges with data access, aggregation, and integration. Investment in the informatics infrastructure to fully leverage diverse data types is imperative and was called out as a priority by the Cancer MoonshotSM Blue Ribbon Panel.

To this end, NCI has initiated development of a Cancer Research Data Commons (CRDC) to provide access to interoperable data repositories, analysis tools, and workspaces. The vision for the CRDC is a virtual, expandable infrastructure that provides secure access to diverse data types, allowing users to analyze, share, and store results, leveraging the storage and elastic compute of commercial cloud technology. The CRDC is built upon Data Commons Framework (DCF) which provides a core set of modular, reusable components to rapidly develop data nodes for the CRDC. Currently the CRDC hosts the Genomics Data Commons (GDC) and the Proteomics Data Commons (PDC), while additional nodes (e.g., Imaging Data Commons, Integrated Canine Data Commons) are in development. In addition, the CRDC offers three NCI Cloud Resources that provide innovative methods to query, visualize, and analyze data. These cloud-based systems allow researchers to bring bioinformatics tools and pipelines to the data, instead of the traditional downloading process of bringing data to the software. They also offer the computational capacity to enable complex analysis of very large datasets. Bioinformatics tool developers may integrate their tools through one or more of the existing Cloud Resources or may connect their tool with the CRDC using the DCF services. A critical component of the CRDC, the Cancer Data Aggregator, is under development and will facilitate the ability to search and integrate data across nodes. The NCI CRDC data nodes and Cloud Resources are currently available to the scientific community to accelerate research and promote new discoveries.

Access, Visualize and Analyze Genomic, Clinical and Phenotype Data of >10,000 Pediatric Cancer Patients and Survivors on St Jude Cloud

Alexander M. Gout*, Clay McLeod, Xin Zhou, Scott Newman, Kirby Birch, Michael Macias, Shaohua Lei, Nickhill Bhakta, Kyla Shelton, Jian Wang, Edgar Sioson, Jaimin Patel, Yutaka Yasui, Scott Foy, Antonina Silkov, Jobin Sunny, Arthur Chiao, Rahul Mudunuri, Madison Treadway, Livia Nason, Andrew Swistak, Chimene Kesserwan, Joy Nakitandwe, Sheila Shurtleff, Elizabeth M. Azzato, Lu Wang, Gregory T. Armstrong, Amar Gajjar, Carmen L. Wilson, Kim E. Nichols, Alberto Pappo, Charles G. Mullighan, David W. Ellison, Michael Rusch, Zhaoming Wang, Melissa M. Hudson, James R. Downing, Leslie L. Robison, and Jinghui Zhang.

St. Jude Children's Research Hospital

The accessibility of pediatric cancer genomic and associated clinical data is key to accelerating scientific discovery with regard to etiology, actionable targets, biomarkers and clinical associations, each of which has a direct impact the clinical management of patients and long-term outcomes. To address this challenge, we have collaborated with technology industry leaders Microsoft and DNAnexus to develop St Jude Cloud (https://stjude.cloud) – a platform comprising one of the world's largest repositories of pediatric genomics data in conjunction with a suite of unique bioinformatics analysis tools and visualizations. The platform hosts paired tumor-germline whole-genome sequencing (WGS) of 3,059 pediatric cancer patient samples, germline-only WGS of 7,745 long-term survivors, and germline-only WGS of 807 sickle cell disease patients. Our collection of cancer genomic datasets includes retrospective data of 1,688 patients from the Pediatric Cancer Genome Sequencing Project (PCGP), in addition to prospective data from 958 patients enrolled in the 'Genomes for Kids' study or our Clinical Genomics Service data – the latter representing the first pediatric real-time clinical genomics (RTCG) data sharing. Germline-only datasets on the platform consist of 4,833 participants of St Jude Lifetime Cohort Study (SJLIFE), a study that brings long-term survivors back to St. Jude Children's Research Hospital for extensive clinical assessments, and 2,912 participants of the Childhood Cancer Survivor Study, a 31-institution cohort study of long-term survivors supported by the NCI. Data access to each of these datasets is simple and fast, comprising a one-time request to access raw data. St. Jude Cloud Users may work with genomic data within a private cloud workspace, upload their own datasets for comparative analysis, utilize our unique collection of bioinformatics tools to quickly and privately gain novel insights, securely share data and results with collaborators within the platform, and access any one of a series of powerful and user-friendly interactive visualizations. In addition, we have recently developed the Survivorship Portal (https://survivorship.stjude.cloud) to streamline genomic-clinical association discovery. This portal hosts data from 3,006 SJLIFE participants with WGS and detailed and harmonized demographic, diagnostic, treatment and outcome information. Our in-house Clinical Dictionary Browser enables interactive exploration of: i) cancer-related variables (e.g. diagnosis, treatment); ii) demographic variables (e.g. age, sex, race/ethnicity); and iii) patient outcomes (e.g. subsequent cancers and severity-graded chronic health conditions). Here, users may traverse each dictionary tree of hierarchically-organized categories and view patient distributions as customizable barcharts, or cross-tabulate pairs of terms. The Survivorship Portal also incorporates GenomePaint (http://genomepaint.stjude.cloud), our new genomic data browser for visualizing and analyzing clinical and genomic data integratively. GenomePaint displays SNV, indel and copy number variants computed from WGS, and enables on-the-fly locus-specific association testing on any coding or non-coding variants to identify alleles associated with user-selected clinical terms. Access on St. Jude Cloud datasets have been granted on 150 occasions to research groups in 14 countries. We are striving to ensure St. Jude Cloud provides a dynamic platform for pediatric genomic research and as such are periodically adding new Clinical Genomics Platform pediatric genomics data to the platform and plan to integrate longitudinal clinical phenotype data in due course. We are also actively

exploring the potential for inclusion of data from non-St. Jude pediatric cancer patients from institutes within the United States and our international collaborators. In light of the recent Children's Cancer Data Initiative we are also looking forward to developing interoperability between St. Jude Cloud and other major pediatric genomics resources in order to harness the full potential of this genomic and clinical data for the advancement of treatments for pediatric cancer and other catastrophic diseases.

The Common Fund Data Ecosystem and Training in Kids First Data

Dr. Alicia Gingrich agingrich@ucdavis.edu, M.D., UC Davis
Dr. Amanda Charbonneau achar@ucdavis.edu, PhD, UC Davis
Dr. C. Titus Brown ctbrown@ucdavis.edu, PhD, UC Davis

The Common Fund Data Ecosystem (CFDE) is an NIH effort to make data use and reuse easier within and between Common Fund projects such as the Gabriella Miller Kids First program, GTEx, and HuBMAP. The CFDE's overall approach is to work with Data Resource Portals to identify challenges and opportunities for improving data (re)use.

In collaboration with the Kids First Data Resource Center (DRC), the CFDE is planning to implement a pilot training program for clinicians, molecular pathologists, and biomedical data scientists in 2020. This will consist of a combination of open online curriculum development, remote training opportunities, and in-person pilot workshops. Our overriding goal is to make this training useful to the Kids First research community. We would particularly like to work with X01 awardees to identify what kind of training is needed, when training would be useful, and how to structure training events and engagement. Please come talk to us!

The Knockout Mouse Phenotyping Program (KOMP2)

Stephen A. Murray[1], Lauryl M.J. Nutter[2,3], and K.C. Kent Lloyd[4]

[1]The Jackson Laboratory, Bar Harbor, ME, 04609, USA
[2]The Centre for Phenogenomics, 25 Orde Street, Toronto, Ontario, M5T 3H7, Canada
[3]The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario, M5G 1X8, Canada
[4]Mouse Biology Program, University of California, Davis, CA, 95618, USA

Mouse models provide a rich resource for understanding gene function, as research platforms to dissect mechanisms, and as tools to validate and study disease-relevant mutations. While the pace of human disease gene discovery is accelerating, ascribing causality to a given gene variant is often challenging and thus many cases remain unsolved. To help bridge this gap, the overall goal of the Knockout Mouse Project (KOMP2) and the International Mouse Phenotype Consortium (IMPC) is to utilize comprehensive, broad-based, high-throughput phenotyping to characterize phenotypes resulting from null alleles in all ~18,000 protein-coding genes with human orthologs in the mouse genome. To date, over 8,500 individual gene knockout mice have been produced, of which we have found that approximately 1/3 are essential for life and also highly enriched for human disease genes. Using a high-throughput embryonic phenotyping pipeline, we have uncovered numerous novel developmental phenotypes, including many that are relevant to Kids First initiatives. Our adult pipeline is designed to assess a wide range of traits associated with behavior and

physiology, and all data is released without embargo through a central portal (www.mousephenotype.org). We have recently worked with Kids First investigators to take advantage of our high-throughput production and phenotyping pipeline to engineer and characterize putative human disease mutations. We are eager to engage with the human genetics community to identify further opportunities for collaboration.

The Gene Expression Database for Mouse Development (GXD): fostering insights into the molecular mechanism of development and disease.

J.H. Finger*, C.M. Smith, T.F. Hayamizu, I.J. McCright, J. Xu, R.M. Baldarelli, J.A. Kadin, J.E. Richardson and M. Ringwald. The Jackson Laboratory, Bar Harbor, ME

Knowledge of gene expression patterns is crucial for understanding the function of genes and the molecular mechanisms that underlie human development, health, and disease. As a mammalian model organism, the mouse is heavily used in developmental research. Tissues from all developmental stages and from many different mouse strains and mutants are subject to detailed expression studies. The Gene Expression Database (GXD) systematically collects and integrates these important data through curation of the literature, by electronic submissions, and by collaboration with large-scale projects. It focuses on endogenous gene expression during development in wild-type and mutant mice and covers data from RNA in situ hybridization, in situ-reporter knock-in, immunohistochemistry, RT-PCR, northern blot and western blot experiments. Currently, GXD includes nearly 1.7 million expression results from 14,816 genes and 351,000 expression images. Data are highly curated using an extensive stage-specific mouse developmental anatomy ontology to record the time and space of gene expression; standard genetic nomenclature for genes and mutants; and many additional controlled vocabularies. As an important component of the larger Mouse Genome Informatics (MGI) resource, GXD combines its expression data with other genetic, functional, phenotypic, and disease-oriented data, thereby enabling users to search for expression data and images in many different ways, using a variety of biologically and biomedically relevant parameters. For example, one can search for the expression patterns of genes associated with specified human diseases. Further, GXD provides utilities that allow the direct correlation of mouse expression and phenotype data. The Gene Expression + Phenotype Comparison Matrix visually juxtaposes tissues where a gene is normally expressed against tissues where mutations in that gene cause abnormalities. The anatomy axis of the matrix can be expanded to view expression and phenotype at different levels of granularity. The Mouse Developmental Anatomy Browser provides access to expression data for a given anatomical structure, as well as corresponding phenotype data. All these data and utilities are freely available.

Visit the GXD Home Page at www.informatics.jax.org/expression.shtml to explore GXD.

The NIH's Rare Diseases Clinical Research Network
Working Together to Find Answers

Eileen King, PhD[1,2,*]; Maurizio Macaluso, MD, DPH[1,2,*]; Tiina Urv, PhD[4]; **Peter S. White**, PhD[1,2,3*]; and the NIH Rare Diseases Clinical Research Network

[1] Cincinnati Children's Hospital Medical Center
[2] Department of Pediatrics, College of Medicine, University of Cincinnati
[3] Department of Biomedical Informatics, College of Medicine, University of Cincinnati
[4] National Clinical and Translational Science Institute, National Institutes of Health

More than 7,000 rare diseases have been identified and characterized to date. While most rare diseases individually affect only a few hundred to several thousand people, collectively they affect more than 25 million individuals and represent a substantial fraction of disease burden and healthcare utilization in the US alone. Many rare diseases are life threatening, and about half of those affected are children. Only five percent of rare diseases have treatments available that have been approved by the U.S. Food and Drug Administration. As rare diseases individually affect only small numbers of people, they can be extremely challenging to study, and both basic and clinical research funding can be scarce.

To address those challenges, the National Institutes of Health (NIH) formed the Rare Diseases Clinical Research Network (RDCRN) in 2003. The Network's goal is to facilitate clinical research by creating rare diseases research groups to focus on related diseases; maximize efficiency by leveraging common research infrastructures across the Network; establish best practices for data collection, study design, and utilization; and make meaningful large-scale studies possible. RDCRN also directly engages with patients and their advocates and trains new investigators in rare diseases research. The RDCRN is supported by multiple NIH Institutes and Centers and led by NIH's National Center for Advancing Translational Sciences (NCATS) and the NCATS Office of Rare Diseases Research.

The Network is made up of teams of scientists, clinicians, patients, families, and patient advocates—called Consortia—that focus on particular rare diseases. The Consortia are intended to advance the diagnosis, management, and treatment of rare diseases with a focus on clinical trial readiness. Each team promotes highly collaborative, multi-site, patient-centric, translational and clinical research with the intent of addressing clinical study needs and challenges. Each Consortium must study three or more diseases, partner with rare disease patient advocacy groups, provide rare disease research training to investigators, and perform natural history studies that chart disease course and progression over time. These efforts are intended to maximize the efficiency and effectiveness of early stage clinical studies, in order to ensure that these studies lead to successful clinical trials and improved therapies for patients.

The RDCRN recently announced that its fourth funding cycle starting in 2019 will support 20 Consortia, including five new groups. It also named Cincinnati Children's Hospital Medical Center as the new Data Management and Coordinating Center (DMCC) for the Network. The DMCC manages shared resources and data from the RDCRN research studies. The DMCC emphasizes the standardization of data, increased data sharing and broad dissemination of research findings using FAIR data principles.

Advancing rare disease research by sharing high-value data for re-use is a critical goal of the program. RDCRN participants will be required to share their data with the DMCC. De-identified data collected within the Network and housed within cloud services provisioned by NCATS will become a resource for the greater rare disease research community. This data will be made available to the scientific community, stakeholders and other relevant partners in a timely way that meets all NIH human subject's protection, data safety and data sharing requirements.

In addition to NCATS, other NIH funding support for RDCRN comes from the National Institute of Allergy and Infectious Diseases, the *Eunice Kennedy Shriver* National Institute of Child Health and Human

Development, the National Institute of Neurological Disorders and Stroke, the National Heart, Lung, and Blood Institute, the National Institute of Arthritis and Musculoskeletal and Skin Diseases, the National Institute of Diabetes and Digestive and Kidney Diseases, the National Institute of Dental and Craniofacial Research, the National Institute of Mental Health and the Office of Dietary Supplements.

Established by Congress under the Rare Diseases Act in 2002, the RDCRN has included more than 350 sites in the United States and more than 50 sites in 22 additional countries. To date, the Network has encompassed 237 research protocols and included more than 56,000 participants in studies ranging from immune system disorders and rare cancers to heart and lung disorders, diseases of brain development, and more. The success of these studies has led to many new therapies and disease insights, as well as best research practices that have positively impacted many additional rare disease research stakeholders. Going forward, the RDCRN's strong focus on rare pediatric disease provides many opportunities to synergize with scientists, patients, and patient advocates within the Gabriella Miller Kid's First community.

Overview of the Undiagnosed Diseases Network

Cecilia Esteves, MPH
Department of Biomedical Informatics, Harvard Medical School, Boston, MA

In 2008, the NIH Undiagnosed Diseases Program (UDP) was established with the goal of providing care and answers for individuals with mysterious medical conditions that have long eluded diagnosis. In 2015, the UDP expanded to a national network of sites, the Undiagnosed Diseases Network (UDN). Currently, the UDN consists of twelve clinical sites, five core laboratories, and a coordinating center. The goals of the UDN are to (1) improve the level of diagnosis and care for individuals with undiagnosed diseases through the use of common protocols designed by a large community of investigators; (2) facilitate research into the etiology of undiagnosed diseases by collecting and sharing standardized, high-quality clinical and laboratory data including genotyping, phenotyping, and documentation of environmental exposures; and (3) promote an integrated and collaborative community across multiple clinical sites and among laboratory and clinical investigators prepared to investigate the pathophysiology and molecular mechanisms underpinning difficult to diagnose diseases. To date, the UDN has received 3857 applications, accepted 1458 individuals for evaluation, completed 1236 multidisciplinary evaluations, and diagnosed 351 individuals. The UDN Coordinating Center in the Department of Biomedical Informatics at Harvard Medical School built and maintains the UDN Gateway, which serves as the centralized patient and investigator portal for the UDN. Key features of the Gateway include a centralized application and case review process; workflows for submitting requests to and receiving results from the sequencing core, metabolomics core, model organisms screening centers, and central biorepository; and integrations with PhenoTips and Qualtrics. The Gateway enables sharing of identifiable phenotypic and genotypic information within the UDN. Study data is shared outside the UDN through various mechanisms including the UDN Participant Pages (udnconnect.org/participants), dbGaP (phs001232.v1.p1), ClinVar (Org. ID: 505999), and PhenomeCentral (phenomecentral.org).

The Children's Brain Tumor Tissue Consortium (CBTTC)

Presented by **Adam Resnick**
Children's Hospital of Philadelphia, Philadelphia

The [Children's Brain Tumor Tissue Consortium](#) (CBTTC) is a collaborative, multi-institutional research program dedicated to finding cures and improving treatments for children diagnosed with brain tumors. To achieve this mission, the CBTTC collects high-quality brain tumor biospecimens and associated clinical data as a part of the biorepository to facilitate the multi-omic analysis of biospecimens. Additionally, CBTTC has developed a network of research data applications that allow researchers from across the world to collaborate and discover cures.

The CBTTC consists of 17 member institutions with its operations center located at Children's Hospital of Philadelphia (CHOP), where the integration of genomic and molecular research, biorepository management and support for the informatics platforms occurs. With more than 3,000+ patients enrolled and 30,000+ biospecimen samples, this is the largest international pediatric brain tumor biorepository with a mandate for open access to researchers throughout the world.

As a research objective, the CBTTC supports the investigation of new prognostic biomarkers and therapies for some of the most difficult-to-treat tumors. The CBTTC has initiated over 100 approved scientific projects. Because the pediatric cancer genome is severely under-represented in prior genomic databases, the CBTTC has also launched the Pediatric Brain Tumor Atlas (PBTA). The Pediatric Brain Tumor Atlas (PBTA) isa multi-omic, data-rich multi-institutional initiative to accelerate discoveries for molecularly-informed therapeutic intervention in children diagnosed with brain tumors. The PBTA comprises comprehensive clinical data in addition to whole-exome sequencing, whole-genome sequencing (WGS), RNA sequencing (RNASeq), miRNA sequencing, and proteomics.

The first dataset release of PBTA occurred on September 10th, 2018. In this release there were over 30 different types of pediatric brain tumors representing over 1,000 subjects. This data is available on the Kids First DRC Portal and PedcBioPortal, leveraging the capacity of users to seamlessly move between applications. Data types include those for matched tumor/normal samples, such as WGS, RNASeq, proteomics, longitudinal clinical data, imaging data (MRIs and radiology reports), histology slide images, and pathology reports. CBTTC promotes real-time data release with no embargo period, allowing PBTA to have is up-to-date data releases with no embargo.

The combination of cloud-based analytics platforms with data from both genomics and clinical practice serves to define a new paradigm for pediatric cancer research and collaborative discovery.

Identifying reported (likely) pathogenic single nucleotide variants for rare genetic diseases in germline genomic data from pediatric brain cancer patients: a pilot study in CBTTC

Yiran Guo[1], Yuankun Zhu[1], Adam Resnick[1]

1Center for Data Driven Discovery in Biomedicine, Children's Hospital of Philadelphia, Philadelphia, PA 19146, USA.

The link between rare diseases and childhood cancer has long been suspected. As an example, the gene ACVR1 was identified to be associated with both Fibrodysplasia Ossificans Progressiva (FOP), a rare bone disorder, and Diffuse Intrinsic Pontine Glioma (DIPG), a form of deadly pediatric brain tumor. A more recently effort evaluated associations between birth defects and cancer risk among >10 million live births using a population-based registry linkage method. However due to the rarity of central nervous system (CNS) tumors and thus insufficient statistical power in analysis, associations between brain/CNS tumors and birth defects are underrepresented in the study. Children's Brain Tumor Tissue Consortium (CBTTC) presents the largest open-access pediatric brain/CNS tumor tissue and data collection and whole genome

sequencing (WGS) can provide comprehensive genetic characterization missed in previous studies. Here harnessing CBTTC, WGS, disease diagnostic information extracted from electronic medical records (EMRs), and cloud computing based bioinformatics support, we started a pilot study to identify pathogenic and likely pathogenic single nucleotide variants (SNVs) that are reported in rare genetic diseases, aiming at interrogating genetic links that intersect pediatric brain tumors with rare diseases. We used CBTTC patients' germline DNA WGS data to generate variant calls of 893 vcf files, annotated those files with snpEff and Annovar, and filtered SNVs according to their pathogenicity. We then manually checked each variant's minor allele frequency in the general population, reported inheritance, and integrated assessment in NIH ClinVar database, obtaining a list of 159 variants in 145 unique CBTTC patients. On the phenotypic/clinical information side, we extracted cancer histologies (provided by tumor boards) and non-cancer diagnostic information (in the format of ICD codes) associated with the Children's Hospital of Philadelphia (CHOP) based CBTTC patients (n=1443) from EMRs, and filtered them based on description of birth defects/rare disorders. The overlap is 75 unique CHOP-based CBTTC patients with both filtered reported SNVs (n=85) and ICD codes. This is the first of such cross-disease investigation of association between CNS tumors in children and rare genetic diseases characterized at molecular level using WGS data.

Expression Patterns of Immune Genes Reveal Heterogeneous Subtypes of High-risk Neuroblastoma

Zhenqiu Liu[1] and Christa Grant[2]
[1]Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA
[2]Department of Surgery, Penn State College of Medicine, Hershey, PA 17033, USA

Neuroblastoma (NB) is the most common extracranial solid tumor of childhood. It is a complex disease with both extreme biological and clinical heterogeneities. While the 5-year overall survival rate for patients with low- and intermediate-risk NB is over 90%, patients with high-risk neuroblastoma (HR-NB), detected in approximately 60% of the cases, have an overall survival below 50%. Standard treatments of chemotherapy, radiotherapy and surgical methods demonstrated particularly low efficacy for HR-NB patients. Although HR-NB is a heterogeneous disease, patients with HR-NB at present are treated in a similar fashion without further risk stratification.

Through mining multiple large cohorts of the gene expression data including TARGET (https://ocg.cancer.gov/programs/target), GSE49710, GSE19274, GSE45547, GSE73517, GSE85047 and GSE120572 (https://www.ncbi.nlm.nih.gov/geo/) with machine learning, we further stratify HR-NB patients into multiple intrinsic subtypes, and discover an ultra-high risk neuroblastoma (UHR-NB) subtype with the worst overall survival. A panel of immune gene signatures and pathways associated with UHR-NB are identified. The findings are validated with independent cohorts, and the biological mechanisms of UHR-NB are further explored. Particularly, we discover that the UHR-NB subtype is enriched in P53 signaling pathway (FDR < 0.01), and a gene TWIST1 identified from UHR_NB subtype is highly expressed in MYCN amplified patients (P < 0.0001) and the down-expression of TWIST1 is strongly associated with the heterozygous deletions of 1p or 11q and gain of 17q, may be a potential therapeutic target for the UHR-NB subtype of neuroblastoma. We are looking forward to collaborating with investigators from Kids First program in the areas of multiomics integration, therapeutic target identification, and network constructions.

Monarch

Peter Robinson
The Jackson Laboratory

One of the primary goals of the Kids First DRC's collaboration with Monarch is to assist in modeling and integrating data for computational utility so as to take advantage of Monarch's cross-species ontologies and algorithms that help support diagnosis and mechanistic discovery. Refining the Kids First DRC data model and curation of diseases and phenotype data using the Monarch Merged Disease Ontology (MONDO), the Human Phenotype Ontology (HPO), and the National Cancer Institute Thesaurus (NCIt) will enable improved searching and analysis. These efforts will be in accordance with standards developed within the Global Alliance for Genomics and Health (GA4GH), where Monarch is a driver project and both programs are key contributors. Additionally, the collaboration will strive to leverage and extend Monarch's patient-centered tools to help families collaborate with their clinicians on the phenotyping and patient-community development. The partnership aims to extend the translational reach of both programs by bringing together organismal data and resources, clinical phenotyping and clinical geneticists, and patients and families, supporting innovative collaboration.

Kids First X01 Mini-Posters (for PIs who could not attend)

Craniofacial Microsomia: Genetic Causes and Pathway Discovery

PI: Daniela Luquetti
University of Washington

Craniofacial microsomia (CFM), also termed hemifacial microsomia or oculo-auricular-vertebral spectrum, is the third most common congenital craniofacial condition. CFM has an estimated birth prevalence in the US of 1 in 3,500-5,600. CFM comprises a variable phenotype, and the most common features include malformations of the ear (i.e. microtia) and lower jaw (i.e. mandibular hypoplasia) on one or both sides. The etiology of CFM is largely unknown; however, the presence of multiple cases within families, mouse models with CFM malformations and the increased risk of CFM in some ethnicities suggest that genetic variants contribute to its occurrence. Although chromosomal abnormalities have been associated with CFM, only three causative genes have been identified in few cases: HOXA2, FGF3, and MYT1. Our goal in this proposal is to identify coding and non- coding variants that are genetic risk factors to CFM by performing whole-genome sequencing (WGS) of case- parent trios with CFM. We propose to perform whole genome sequencing on samples from 105 trios (individuals with CFM and their parents or affected relatives in multi-affected families) to identify candidate genes with rare de novo and inherited variants. Our hypothesis is that CFM is caused by rare new and inherited DNA variation in gene(s) related to the craniofacial development. We will analyze the data on rare de novo coding and non- coding variants. Recognizing reduced penetrance in CFM, our analysis will include analyses for variants in a dominant inheritance with incomplete penetrance model. Our approach incorporates detailed phenotype, clinical characterization, and family history for each individual. We will also integrate the WGS data with our data on gene expression from murine embryonic pharyngeal arch and external ear human embryonic tissue to ascertain tissue specific expression at the relevant time of the development of tissues in CFM. Our statistical power by sampling patients with familial and severe disease who are most likely to have a high genetic loading. This study will be conducted by an interdisciplinary team with complementary expertise in clinical aspects of CFM, clinical genetics, genomics, and bioinformatics. Successful completion of this proposal will advance knowledge in the genetic

architecture of susceptibility to CFM and will provide insight about the biological mechanisms underlying craniofacial development. The phenotypic and genomic data will be fully integrated into the Kids First Data Resource and available to all qualified investigators. The long-term goal of this project is to identify specific genetic risk factors to improve genetic counseling, enable tailored clinical care, and to provide more accurate prognosis. The results from the proposed study have potential to further research on the etiology of other craniofacial disorders, and the pathogenesis of typical and atypical craniofacial development.

Genetic Predisposition to Germ Cell Tumors

PIs: Jenny Poynter & Ching Lau
The Jackson Laboratory

Pediatric germ cell tumors (GCTs) are rare and heterogeneous tumors that most commonly occur in the gonads but also develop in other locations. Using case parent trios recruited for a Children's Oncology Group (COG), the PIs of this project found that common genetic variants associated with adult testicular GCT are also associated with both intracranial and systemic GCT, suggesting that there may be common genetic risk factors for all GCT types. Identification of additional genetic variants for Intracranial and Extracranial GCT risk will require a larger study using whole-genome sequencing (WGS) data. We recruited GCT cases and their parents through the Children's Oncology Group (COG) Childhood Cancer Research Network. In this study, we collected germline DNA for 867 GCT patients between the ages of 0-19 years at diagnosis, including 677 families with DNA samples for the complete trio. The intracranial GCT cases from the U.S., Japan, and Thailand will be included in a project selected for funding by the Gabriela Miller Kids First Sequencing program in 2018 (X01 HL145700; PIs Lau and Poynter). In the current proposal, we are proposing to use WGS and WES to evaluate 1) the contribution of rare genetic variants in GCT, 2) de novo variants, and 3) molecular signatures of tumor specimens, overall and by age group in the extracranial (testicular, ovarian, and extragonadal) GCTs recruited for this study. Whole Genome Sequencing data generated through the Gabriella Miller Kids First Pediatric Research Program will provide an opportunity to investigate the genetic origins of GCT in a diverse set of samples. Given the limited knowledge of GCT etiology and biology, the results of the proposed analyses are likely to have a big impact on the field.

Comprehensive Genomic Profiling to Improve Prediction of Clinical Outcome for Children with T-cell Acute Lymphoblastic Leukemia

PIs: David Teachey & Charles Mullighan
Children's Hospital of Philadelphia & St. Jude Children's Research Hospital

The outcome for patients with relapsed T-ALL is dismal with 3-year event free survival of <15%. Thus, the primary goal in the treatment of T-ALL is to prevent relapse, which requires accurate risk stratification. Unfortunately, no genetic alterations have been identified to date that are reproducibly prognostic independent of minimal residual disease (MRD), making it difficult at diagnosis to identify which patients are more likely to relapse. AALL0434 was a Children's Oncology Group-initiated phase 3 randomized clinical trial comparing Capizzi-style escalating methotrexate plus pegaspargase (CMTX) vs. high dose methotrexate (HDMTX), with/without six 5-day courses of nelarabine. Survival on this study was superior to any prior trial for de novo T-ALL, changing the standard of care. Yet, a substantial minority (~15%) of patients had relapsed or refractory (r/r) disease. We recently performed RNA sequencing, DNA copy number analysis, and whole-exome sequencing on 264 T-ALL patients treated on AALL0434, demonstrating recurrent alterations could be grouped into 10 different potentially targetable functional

pathways. This analysis was not powered to examine associations between genetic lesions with outcome, because too few patients with r/r disease were included. We hypothesize that comprehensive genomic profiling of the entire AALL0434 cohort will identify recurrent genetic alterations that can be segregated into biologically relevant deregulated pathways that can be combined with MRD to identify patients at risk for poor outcomes before they relapse and provide rationale for treatment with alternative therapies. In addition, a number of small recent studies demonstrated that many of the biologically relevant alterations in T-ALL occur in non-coding regions of the genome, but no large studies have performed whole genome sequencing in T-ALL. We further hypothesize that whole genome sequencing of a large cohort of patients with T-ALL will identify novel lesions in coding and non-coding regions that will be highly impactful in the understanding of T-ALL pathogenesis. We will test our hypotheses by performing comprehensive genomic profiling (whole genome sequencing, whole exome sequencing, RNA sequencing, and copy number analysis) of the entire AALL0434 cohort (n = 1430) with the following specific aims: (1) identify recurrent genetic alterations that predict poor outcome in T-ALL; (2) identify novel alterations, including non-coding alterations in T-ALL; and (3) identify germline genetic variants that predispose to T-ALL and to increased toxicity to nelarabine. The goal of the Kids First Program is to improve understanding of genetic mechanisms of disease, leading to improved diagnostic capabilities and ultimately more targeted therapies or interventions. This proposal will meet that important goal through identification of germline and somatic alterations in T-ALL that can be used to identify patients that are likely to relapse before they relapse and can be treated with new therapies.

Germline and Somatic Variants in Myeloid Malignancies in Children

PI: Soheil Meshinchi
Fred Hutchinson Cancer Research Center

Advances in genomic sequencing have allowed identification of somatic variants as potential therapeutic targets. Although myeloid disorders in children may show morphologic similarities to that seen in adults, TARGET AML initiative (Meshinchi, PI) clearly demonstrated that somatic genomic and transcriptome variants are highly distinct in children and young adults, and in fact, there are variants that are uniquely restricted to younger children. TARGET AML initiative, although modest in number, helped identify numerous somatic alterations with high therapeutic potential in younger AML patients. In addition to identification of somatic variants, analysis of the germline data provided a glimpse into the constitutional make-up of patients with AML. The identification of numerous "function altering" variants may provide an insight into possible interactions between the host and the disease, where these germline variants might alter AML risk (predisposition), response to therapy (altering target expression, drug metabolism), susceptibilities to short and long term complications (including infectious and cardiac complications) or modify risk of secondary malignancies. Armed with data from initial sequencing efforts in AML, we are poised to take full advantage of the available sequencing technology to conduct the most comprehensive genome and transcriptome interrogation of myeloid disorders in children in specimens we have amassed over the last decade. To this end, we have put in place unparalleled specimen resources from children with de novo AML, Down Syndrome AML (DS-AML), and acute promyelocytic AML (APL) treated on COG trials. In addition, thru collaboration with Dr. Resar and Kucine, we will be able to conduct the first broad sequencing study in the rare entity of myeloproliferative neoplasms of childhood (MPN-c). Identification of the somatic variants will provide valuable data on the potential genes and pathways that can be targeted for therapeutic gains. In addition, interrogation of the host's constitutional genome may yield valuable information about potential germline variants that, in combination with the somatic data, might provide a more informed approach to patient care. For those patients with predisposition mutations, chemotherapy

alone might not be adequate for a cure, and stem cell transplantation might be required. Also, those who might be at high risk of adverse secondary events (cardiac complications, secondary malignancies, etc.) can be identified early and their therapy tailored to minimize anticipated complications. Thus, we propose that the optimum outcome can only be obtained thru comprehensive interrogation of the somatic and germline genome to fully annotate the genomic makeup of the leukemia and its host.

An Integrated Clinical and Genomic Analysis of Treatment Failure in Pediatric Osteosarcoma

PI: Kenan Onel
The University of Chicago

For children with osteosarcoma, it has long been known that response to chemotherapy as measured by percent necrosis at the time of definitive surgery is a powerful prognostic biomarker. Patients with 90 percent or more tumor necrosis are likely to be cured of their disease, whereas those with less than 90 percent tumor necrosis are at high risk for treatment failure. Despite its clinical importance, however, virtually nothing is known about the genetic and molecular basis of this phenomenon. Consequently, there have been few advances in the treatment of osteosarcoma in decades. In this proposal, we will perform whole genome sequencing on serial samples obtained over time from a set of 198 patients with osteosarcoma, all treated similarly, and for whom we have complete clinical information. Of these patients, 52 have suffered a relapse of their disease. Our primary objective is to determine whether there are recurrent mutations in these relapse samples that may point towards common mechanisms of treatment failure, and may, therefore, suggest novel therapies for relapsed osteosarcoma. Our secondary objective is to determine the genetic drivers of treatment failure in osteosarcoma by analyzing within each patient the evolving spectrum of mutations selected by chemotherapy exposure over time. To our knowledge, this is the largest set of matched pre-therapy, post-therapy, and relapse samples ever assembled for any cancer. If successful, this project sets the stage for future functional studies exploiting our genetic findings to investigate the mechanisms of drug resistance in osteosarcoma. Perhaps more importantly, it also holds forth the promise of changing the paradigm for therapy in osteosarcoma, a disease that has thus far proven refractory to innovative therapies to improve the dismal survival of children with tumors that respond poorly to current chemotherapy protocols.

Phenotypic Features and Genetic Mechanisms in a PHACE Syndrome Cohort

Nara Sobriera, MD[1], Wendy Demos, MS[2], Elizabeth Partan, BS [1], Chien-Wei Lin, PhD[3], Michael Zimmermann, PhD[2], Beth Drolet, MD[4] (PI), Dawn Siegel, MD[4] (PI)

1.   McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD
2.   Genomic Sciences and Precision Medicine Center, Medical College of Wisconsin, Milwaukee, WI
3.   Division of Biostatistics, Medical College of Wisconsin, Milwaukee, WI
4.   Department of Dermatology, Medical College of Wisconsin, Milwaukee, WI

Infantile hemangiomas (IH) are the most common benign vascular tumors in infants, affecting 4-5% of children[1]. Treatment is required in 9,600 infants annually for complications, including disfigurement and impairment of vital functions. If left untreated, blindness can result from occlusion of the visual axis or significant disfigurement. Large hemangiomas can be associated with birth defects. Thirty percent of segmental hemangiomas on the face and scalp are associated with a syndromic condition with multi-organ structural congenital anomalies that include some or all of the following: posterior fossa brain

malformations, segmental facial hemangiomas, arterial anomalies, cardiac defects, eye anomalies, and sternal clefting or supraumbilical raphe (PHACE)[2]. We hypothesize that PHACE is genetically-determined based on the consistent phenotype, the classic brain and heart malformations and the lack of teratogens or environmental factors associated with the syndrome. Candidate genes will be analyzed with a functional confirmation program to create connections relevant to underlying biological networks. Gene lists will be constructed based on known developmental brain and vascular disorders and genes expressed in embryonic cranial mesenchyme and vasculature, with data mostly derived from mouse. Molecular network-based models will be constructed for normal cranial and vascular development. Developmental phenotypes in zebrafish will help to determine the functional significance of potentially disease-causing genetic variants and enhance genotype-phenotype analysis. Correlating genetic alterations with clinical outcomes will allow clinicians to predict PHACE related complications to improve tailored screening and evaluation tests. This data will also serve as an important resource in a broader context through the Kids First Pediatric Research Program.

References:
1.   Kanada KN MM, Munden A, Friedlander SF. A prospective study of cutaneous findings in newborns in the United States: correlation with race, ethnicity, and gestational status using updated classification and nomenclature. J Pediatr 2012;161:240-5.
2.   Frieden IJ, Reese V, Cohen D. PHACE syndrome. The association of posterior fossa brain malformations, hemangiomas, arterial anomalies, coarctation of the aorta and cardiac defects, and eye abnormalities. Archives of dermatology 1996;132:307-11.

Genomic Analysis of Esophageal Atresia and Tracheoesophageal Fistulas and Associated Congenital Anomalies

PI: Wendy Chung
Columbia University Health Sciences

Esophageal atresia/tracheoesophageal fistula (EA/TEF) is a rare and complex aerodigestive congenital anomaly with an estimated incidence of 1 in 2500 to 1 in 4000 live births. There is a 45% incidence of associated congenital malformations, most commonly digestive, cardiovascular, urogenital, and musculoskeletal, often part of a syndrome or complex association, with VACTERL (vertebral defects, anal atresia, cardiac defects, tracheoesophageal fistula, renal anomalies, and limb abnormalities) being most frequently recognized. Advanced surgical techniques and pre and post-operative care have improved the prognosis and survival of EA/TEF patients over the past decades. However, with improved survival, many of the long-term morbidities of EA/TEF have been exposed. It is likely that the outcome in EA/TEF patients is influenced by multiple genetic and clinical factors; however, determining which factors are critical has been limited by the lack of data, particularly genomic data. Many families and health care providers seek prognostic clinical information about other associated birth defects or genetic syndromes, but prognostic data are extremely limited unless a chromosomal anomaly is identified. Evidence is accumulating that many congenital anomalies can result from copy number variants, de novo mutations, and inherited rare mutations, often unique to the family. We propose to elucidate the underlying genomic architecture of EA/TEF and define new genes and conditions associated with EA/TEF by performing whole genome sequencing on 100 parent child trios in a clinically well characterized cohort to identify rare de novo mutations and inherited variants. We believe this information will improve genetic diagnostic methods and provide more accurate clinical prognostic information to guide clinical decisions and improve outcomes.

(Recessive) Germline Mutations in CHD

PI: Christine Seidman
Harvard Medical School

Congenital heart disease (CHD) is the most common birth defect and is often accompanied by another congenital anomaly (CHD±CA). The Pediatric Cardiovascular Genetics Consortium (PCGC) is committed to defining the molecular mechanisms for CHD±CA. We have recruited over 29,000 participants including over 6000 CHD probands and parents (CHD trios) with extensive clinical data. Whole exome sequence (WES) analyses in ~3300 CHD trios by the PCGC has defined likely genetic causes in ~40% probands. As part of the Gabriella Miller Kids First Pediatric Research Program, we propose WGS to enable the discovery of variants and mechanisms that contribute to unexplained CHD in ~60% of probands studied by the PCGC. To accomplish these goals, we will capitalize on new insights into CHD genes, identified by WES, that indicate aberrant transcriptional regulation during development is a major cause of CHD. In this application, we request WGS on 550 CHD trios so that by leveraging existing genomic datasets we will empower robust analyses of variants that alter noncoding regulatory elements of cardiac development genes in WES-negative trios. **Nested within this trio group are 100 CHD trios comprised of a proband with one damaging variant in a recessive CHD gene. In addition to genome-wide studies, focused analyses in this trio subgroup will search for noncoding variants that impact the "normal" allele.**

The Genetics of Microtia in Hispanic Populations

PI: Jonathan Seidman
Harvard Medical School

Microtia is a rare congenital deformity of the external ear, the pinna. The severity of microtia is variable and ranges from subtle deformities in the pinna to absence of the external ear. Microtia is often associated with closure of the external auditory ear canal causing significant hearing loss. Microtia can be an isolated, unilateral or bilateral malformation, or occur solely with ear canal deformities, or with additional craniofacial or syndromic manifestations. Our study of identical twins with microtia demonstrated a significant genetic contribution. The molecular pathogenesis for most microtia remains unknown. We propose to leverage our clinical acumen in diagnosis and treatment of microtia (R.E.), our relationship to the microtia community (M.T.) and our collected DNA samples from microtia patients to identify genetic variant(s) that contribute to this congenital malformation. Microtia prevalence is much higher among Native Americans and some Latin Americans (17 per 10000 Ecuadorian births) than among individuals of European-descent (0.6 -1.6 per 10,000 births). To capitalize on this epidemiologic data, we have recruited microtia cohorts from Latin America and the U.S, including clinical data and DNA samples. We propose whole genome sequence of existing samples from isolated cases, trios (proband and parents) and one large family we propose comprehensive genetic analyses to interrogate coding and non-coding sequence variants associated with microtia. We hypothesize that genetic variants that cause microtia and other less pathogenic conditions, which have relatively small impact on reproductive fitness, are likely to be tolerated and inherited, but cause malformations in only a fraction of variant carriers (i.e. reduced penetrance). We suggest that we have power to detect a variant that increases the relative risk of microtia by >2.5 (i.e. a penetrance of ~3%). We suggest that microtia likely reflects variants with low penetrance that impact genes that participate in the molecular pathways of ear development. Such variants may also contribute to other hearing and craniofacial malformations. We expect to harness the insights and reagents developed here to elucidate factors that impact the penetrance of variants. Because of the prevalence of microtia in Latin America there are microtia

support groups in Mexico, Colombia, and Ecuador. We have formed alliances, through our collaborator Melissa Tumblin (Ear Community), with these microtia support groups. We anticipate that any associations detected in the preliminary whole genome sequence (WGS) cohort will be confirmed in a second cohort of microtia patients. We request that the Gabrielle-Miller Kids First program support WGS of 821 microtia subjects and their parents as follows: a) 200 microtia probands; b) 200 trios (proband and both parents) and c) 21 members of family 3Sz.

Discovery of Genetic Basis of Fetal Alcohol Spectrum Disorders

<u>PI</u>: Christina Chambers
University Of California, San Diego

Fetal Alcohol Spectrum Disorder (FASD) is the most common birth defect worldwide, and is estimated to occur in at least 1-5% of all children in the U.S. However, not all children with prenatal exposure are similarly affected, even among those born to heavy, chronic alcohol-consuming pregnant women. Recent research has focused on the susceptibility or protective factors that seem to influence the risk for FASD. However, very little is known about the genetic risk or protective factors that may interact with prenatal alcohol exposure leading to this variable risk. In this study, we will use whole genome sequencing of well-characterized mother-child pairs, including mothers with prenatal alcohol exposure and their children with or without FASD, to test the hypothesis that genomic alterations in either the mother or her fetus or both play a role in susceptibility to the effects of alcohol. This information will be of critical value in better understanding the pathogenetic mechanisms underlying FASD. In addition, the identification of maternal or fetal genetic susceptibility factors for FASD may inform future intervention strategies for this common congenital disorder.

Genomic Analysis of Laterality Birth Defects

PI: Stephanie Ware
Indiana University and Purdue University at Indianapolis

Laterality defects occur in approximately 1:10,000 newborns and are associated with a range of structural birth defects and abnormalities of organ positioning. Gut malrotation, biliary atresia, asplenia or polysplenia, complex cardiovascular malformations, and midline defects such as neural tube defects, vertebral anomalies and rib fusions are found in various combinations in patients with laterality defects. In addition, a subset of laterality defects caused by abnormalities of cilia position or function are associated with additional medical problems such as chronic sinusitis and bronchiectasis that require specific preventive care; however these patients frequently are undiagnosed until late in disease course. The clinical picture firmly establishes laterality defects as not only diseases of significant phenotypic heterogeneity, but also ones of considerable medical and economic consequence. The goal of this project is to elucidate the genetic architecture of laterality defects in order to inform medical management and prevent complications. Laterality disorders are genetically heterogeneous and we and others have previously identified single nucleotide variants inherited in an X-linked or autosomal recessive manner as explanations for a minority of cases. In addition, we have demonstrated copy number variants (CNVs) as a mechanism of disease that requires additional investigation. We hypothesize that the majority of cases result from complex genetic inheritance. We propose to investigate this hypothesis using a multifaceted analysis approach in our extremely well phenotyped cohort of 550 probands with laterality disorders. Included within this cohort are 280 probands

who had exome sequencing which was negative for pathogenic variants in 170 clinically relevant laterality genes. However, preliminary data demonstrate increased variant burden in these cases versus controls when interrogating 809 candidate genes important for left-right patterning and cilia function. These samples are excellent candidates for gene discovery and association analyses via whole genome sequencing (WGS) which will allow broader interrogation and expansion of analyses to include non-coding regions and CNVs. We will perform burden analyses to identify genes, gene interactions, and pathways important for susceptibility to laterality disorders. Also nested within our cohort of 550 probands are 105 trios that have not had previous sequencing. All trios will be analyzed by transmission disequilibrium test (TDT) including rare variant TDT. De novo mutations will also be identified from trios for potential gene discovery. This comprehensive genetic analysis in patients with laterality disorders is necessary to identify the appropriate clinical diagnostic testing for risk stratification, to elucidate underlying genetic architecture and facilitate novel gene discovery, and to provide essential knowledge about genes and pathways impacting the development of these birth defects.

End