**Intro**

*De novo* variants

* New variants in the offspring that are not in the parents
* Fundamental to evolution and important for disease pathogenesis
* Understanding how DNVs arise has impacts on evolution, reproductive decision-making, knowledge of disease pathogenesis, and predictions mutagenic somatic events (i.e., not germline). Models of DNVs provide
* One method for understanding DNVs is assigning them to a parent of origin. This improves our understanding because the biology of sperm and oocytes is vastly different.
* Both spermatogonia and oocytes undergo approximately ~20 cell divisions as primordial germ cells before puberty
* After puberty sperm continuously undergoes mitosis, while oocytes are stable cells. Furthermore, both occupy different environments. Teasing apart the mutagenic properties of stable vs labile cells has important implications for non-germline cell types.
* Review introductions/conclusions of recent *de novo* variant papers!

*De novo* variant phasing

* In order to determine if a de novo variant arose in sperm or oocyte, they have to be phased (i.e., assigned) to the parent of origin
* Two methods for phasing DNVs
* Three-generation haplotype phasing
  + Phases x% of DNVs
  + Requires sequencing of offspring of the person to whom the DNV is assigned, which is not possible for anyone who has not had children and difficult to collect larger families.
* Read-based phasing
  + identifying heterozygous variants on the same read that are uniquely inherited from mom or dad
  + Phases x% of DNVs with Illumina
  + Limitation: read length
  + Here we use long-read technology to phase DNVs that were original identified with short-read sequencing

Patterns of *de novo* SNVs and indels

* What is known about germline *de novo* SNVs?
  + Review results
* What is known about germline *de novo* indels?

Here, we perform the first long-read phasing of *de novo* variants, replicate previous findings with *de novo* SNVs, and define the mutagenic properties of *de novo* insertions and deletions (indels).

**Methods**

**Subjects, whole genome sequencing and *de novo* variant calling**

Patients and parents (*i.e.*, trios) were enrolled in the Pediatric Cardiac Genomics Consortium (PCGC) Congenital Heart Disease Network Study (CHD GENES: ClinicalTrials.gov identifier NCT01196182).1 The protocols were approved by the Institutional Review Boards of Boston’s Children’s Hospital, Brigham and Women’s Hospital, Children’s Hospital of Los Angeles, Children’s Hospital of Philadelphia, Columbia University Medical Center, Great Ormond Street Hospital, Icahn School of Medicine at Mount Sinai, Rochester School of Medicine and Dentistry, Steven and Alexandra Cohen Children’s Medical Center of New York, and Yale School of Medicine. All subjects or their parents provided informed consent.

*Pacific Biosciences (PacBio) long-read sequencing:* DNAs of 10 PCGC patients were sequenced with PacBio long-read sequencing to a depth of 5x.

*Illumina short-read sequencing:* DNAs of PCGC samples underwent short-read sequencing at the Baylor College of Medicine Genomic and RNA Profiling Core (n=XXX), the New York Genome Center (NYGC) Genomic Research Services (n=XXX), and the Broad Institute for Genomic Services (n=XXX) following the same protocol. Genomic DNAs from venous blood or saliva were prepared for sequencing using a PCR-free library preparation. All samples were sequenced on an Illumina Hi-Seq X Ten with 150-bp paired reads to a median depth >30x per individual. Alignment, variant calling, and *de novo* identification were performed as described previously (cite WGS paper).

**Phasing**

Whatshap (0.16) was used to phase DNVs through read-back phasing.2 Inputs were either the short- and long-read alignment files, the trio VCF generated from Illumina short-read sequencing. Indels were used for short- but not long-read phasing (*i.e.*, did not use whatshap --indel flag for PacBio data). The output was a phased VCF, with the full trio and the proband’s variants phased.

Following phasing, DNVs were programmatically assigned to the parent of origin. Whasthap was used to generate a GTF from the phased VCF, where the GTF genomic coordinates represented haplotype blocks of contiguously phased variants. DNVs were assigned to a parent of origin if ≥85% of informative variants in (*i.e.*, variants in a haplotype block) were assigned to that parent.

For PacBio data, *de novo* indels were phased manually with the Integrative Genomics Viewer (IGV).3 For each *de novo* indel, we identified the 2-5 highest confidence reads (with Reference/Alternative alleles). These reads were highlighted with IGV and we identified inherited single nucleotide variants (SNVs) on the informative reads. We then assigned the *de novo* indel to the parent of origin if all SNVs on the informative reads were in agreement. We validated this heuristic approach with the *de novo* indels phased using Illumina short-read sequencing. IGV plots used for *de novo* indel phasing are provided in the code repository (see **URLs**).

Replicating Iceland Results

* Used Ref and Alt columns of dataframe to classify SNPs into mutational classes (C > A,

C > T, C > G, T > A, T > C, CpG > TpG) and included indels as their own mutational class

* In order to get CpG > TpG mutations:
  + Used bedtools getfasta with hg38 reference file to get the single nucleotides on either side of the SNP and marked every SNP that was a C adjacent to a G, or a G adjacent to a C as CpG
  + The *de novo* variants that were annotated as being in CpG regions and were C > T mutations classified as CpG > TpG, the *de novo* variants that were C > T but were not in CpG regions were classified as C > T (no overlap between the two)
* Calculated the fraction of phased *de novo* variants that were components of each mutational class and created bar chart to compare those assigned to mother versus father as parent of origin for each class
* Plotted these fractions for each proband against the age of the mother and age of the father to look for associations with parental age

Classifying Indels

* Developed Python package to automate this process
  + Named sorting\_hat
* Three classes: HR, CCC, non-CCC
  + HR: homopolymer run (mutation is in a region where there are 6 or more copies of the nucleotide being inserted or deleted)
  + CCC: change in copy count (the sequence being inserted or deleted has 1 or more repeats in the mutation region)
  + non-CCC: no change in copy count (the sequence being inserted or deleted is not repeated in the mutation region)
* Used bedtools getfasta to get bases surrounding the indel
  + If the indel was a single nucleotide, collected 6 bases on either side of the indel
  + If the indel was a sequence of 2 or more bases, collected 2\*length of sequence bases on either side of the indel
* Compared indel sequence to adjacent sequence of equal length
  + If the indel sequence and adjacent sequence were the same:
    - If the indel sequence was a single nucleotide, checked for 6 adjacent copies to see if it was HR
    - Otherwise, assigned as CCC
  + If the indel sequence and adjacent sequence were not the same, assigned as non-CCC
* Downloaded RepeatMasker from UCSC Genome Browser to obtain repeat name, repeat class, and family if the indel was in a repeat region

**URLS**

Code: <https://github.com/allisonseiden/longreadclustersequencing/>

**References**

1. Gelb, B. *et al.* The Congenital Heart Disease Genetic Network Study: rationale, design, and early results. *Circ. Res.* **112,** 698–706 (2013).

2. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *bioRxiv* 085050 (2016). doi:10.1101/085050

3. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–6 (2011).