**Title**

Defining *de novo* indel biology with long-read technology

**Authors**

Allison H. Seiden†, Felix Richter†, Nihir Patel, Andrew Sharp, Bruce D. Gelb\*

† Contributed equally

**Contact information**

**\***[bruce.gelb@mssm.edu](mailto:bruce.gelb@mssm.edu)

**Target journal**

*Genome Research*

**Abstract**

**Introduction**

Genomes evolve through mutation and selection. Mutations include single nucleotide variants (SNVs), short insertions and deletions (indels, ≤ 50-bp changes), and larger structural variations. Models of local indel mutation rates would serve as a powerful control for disease association studies as well as an invaluable tool for understanding evolution, especially in repetitive regions.1

Modeling *de novo* indels requires understanding the mutational mechanisms that birth them (**Table 1**).2,3 Indel mechanisms can be investigated through associations with parental age, relative genomic distribution, local sequence context, and ancestry associations. Three quarters of *de novo* indels arise through polymerase slippage, with the remaining quarter hypothesized to arise from double-strand break repair (*e.g.*, non-homologous end-joining) and/or unknown mechanisms.2,3 Polymerase slippage implies a replication-associated event consistent with spermatogenesis, yet previous studies failed to identify an association between *de novo* indels and paternal age.3,4 This could be attributable to a small sample size of parentally-assigned (*i.e.*, phased) indels or not applying previously defined indel classification techniques to the new data. Overall, these gaps in our understanding of *de novo* indel mechanisms arise from small sample sizes or not applying indel classification techniques to newer data.

Here, we describe the feasibility of using low-coverage Pacific Biosciences long-read sequencing (N=10 trios) to phase *de novo* SNVs and indels identified with short-read Illumina whole-genome sequencing (WGS). We used these data as an orthogonal technology to validate previously observed associations between parental age and *de novo* SNVs. We then developed an application programming interface (<https://pypi.org/project/sorting-hat/>) to classify indels based on sequence context. Finally, we combined the long-read *de novo* indel results with data from Illumina read-pair tracing (N=308 trios) and three-generation haplotype phasing (N=202 families)4 to characterize indel mutagenesis.

**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).5 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.6 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).7 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**).

Three-generation haplotype phasing results were downloaded from Supplemental Data.4

**URLS**

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

**References**

1. Campbell, C. D. & Eichler, E. E. Properties and rates of germline mutations in humans. *Trends Genet.* **29,** 575–584 (2013).

2. Montgomery, S. B. *et al.* The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* **23,** 749–761 (2013).

3. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res.* **25,** 792–801 (2015).

4. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549,** 519–522 (2017).

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

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

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