Red indicates a to-do

**Title**

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

**Authors**

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

† Contributed equally

**Contact information**

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

**Target journal**

*Genome Research*

**Abstract**

*De novo* insertion/deletion (indel) genesis mechanisms, such as polymerase slippage, have been hypothesized but not well-characterized in human parent-child trios. We demonstrated the utility of assigning *de novo* variants to parents-of-origin (*i.e.*, phasing) with low coverage long-read whole genome sequencing, wrote an application programming interface to classify indels by sequence context, and found that change in copy count (CCC) *de novo* indels were significantly correlated with father’s age at conception (*P*=9.1x10-5). However, paternal age explained <10% of variance in *de novo* CCCs, suggesting that mechanisms outside of spermatogenesis-associated DNA replication have a significant role in forming indels. These results have implications for indel modeling in genome evolution and disease.

**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 could 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 experimentally (*e.g.*, with biochemical assays or structural biology)4 or at the population level through associations with parental age, relative genomic distribution, local sequence properties, 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.3,5 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 frameworks 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 results from three phasing methods, the long-read sequencing results (N=10) as well as data from Illumina read-pair tracing (N=305 trios) and three-generation haplotype phasing (N=225 families),5 to characterize insertion/deletion mutagenesis.

**Table 1. Description and examples of indel classes.**

|  |  |  |  |
| --- | --- | --- | --- |
| *Position* | *Indel* | *Context* | *Type* |
| chr8:117967436 | +T | TTAAATATTTTTTT | HR |
| chr5:52931910 | -A | CCAATTAAAAAAA | HR |
| chr12:71038252 | -GTG | TAACTTGTGGTGTTT | CCC |
| chr2:158890470 | +GAG | CAGAACTGAGGAGCAT | CCC |
| chr20:39588834 | -AGG | GAGAGAAGGAGATGT | non-CCC |
| chr4:14567444 | +AACCC | ACATAATATATAACCCAACCTACCTT | non-CCC |

HR=homopolymer run, CCC=change in copy count

**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).6 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 target depth of 10x.

*Illumina short-read sequencing:* DNAs of PCGC samples underwent short-read sequencing at the Broad Institute for Genomic Services. 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 (also known as read-pair tracing).7 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).8 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.5

**Replicating *de novo* SNV results**

Variants were classified into one of eight mutational groups (C>A, C>G, C>T, T>A, T>C, T>G, CpG>TpG, and indel), with reverse complements grouped together (*e.g.*, C>T and G>A are both C>T). Bedtools getfasta was used to identify CpG>TpG mutations by marking reference C alleles with an adjacent G, and reference G alleles with an adjacent C.9

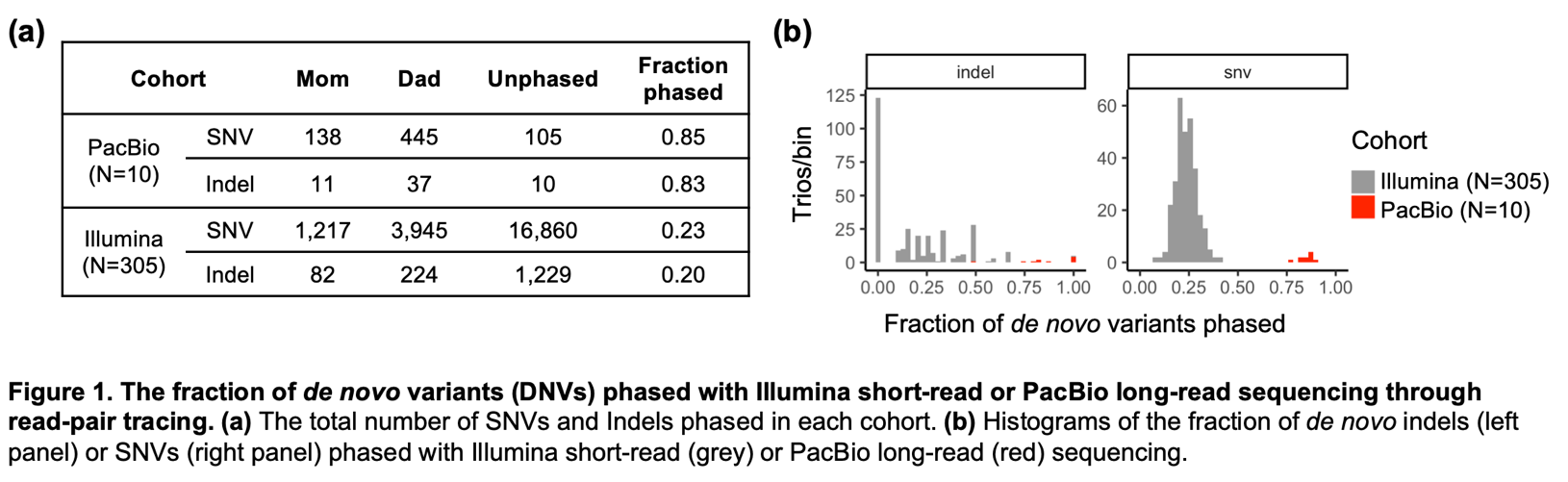
**Indel classification application programming interface (API)**

The three classes of indels were defined as follows. Homopolymer run (HR) mutations occur in regions with 6 or more copies of the inserted/deleted single nucleotide. Changes in copy counts (CCCs) occur if the sequence being inserted or deleted has one or more repeats in the directly flanking bases. Finally, any indel not falling in the above categories was considered a non-CCC. The sorting-hat API was made to automate this classification process. Sorting-hat calls the bedtools getfasta command and collects flanking base pairs depending on indel length.9 Sorting-hat collects the six flanking bases if the indel was a single nucleotide, or 2\*L flanking bases if the indel length (L) was greater than one. Sorting-hat also optionally annotates indels with the encompassing repeat if the RepeatMasker track from the UCSC Genome Browser is provided.10

**Statistics**

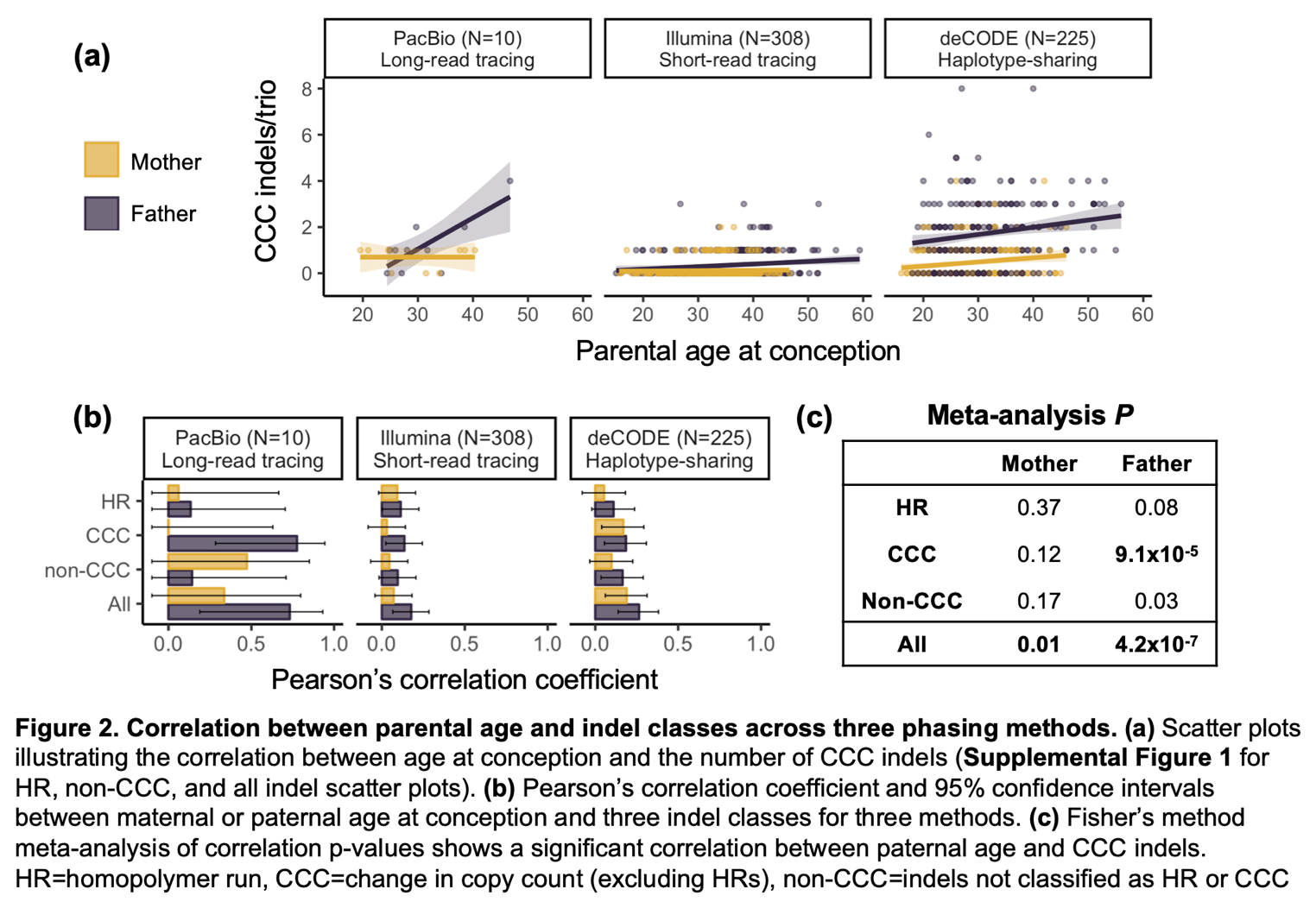
Correlations were quantified with the Pearson’s correlation coefficient, with statistical significance assessed with two-tailed p-values. The meta-analysis p-values were calculated with Fisher’s method, in which constituent p-values were log-transformed, multiplied by -2, and summed, generating a chi-squared test statistic (with 2 \* Np-values degrees of freedom, N=number of constituent p-values) that was converted to a p-value.

**Results**

The most common method for phasing *de novo* variants (DNVs) depends on read length. Previous read-pair tracing with short-read sequencing assigned ~20% of *de novo* SNVs to a parent-of-origin.3,5 We used read-pair tracing to phase DNVs identified with short-read whole genome sequencing (**Figure 1**), achieving similar results for short-read phasing (23%, N=305 trios) but a considerably higher fraction phased with low-coverage (<10x) long-read technology (84%, N=10 trios).

Having phased DNVs in two cohorts, we first sought to replicate previous *de novo* SNV associations with parental age. Previously, C>A and T>G DNVs were observed to have significant enrichment for paternal origin, while C>T DNVs were found to have significant enrichment for maternal origin.5 Although not significant (check significance), we observed the same directionality in all three variant classes (**Supplemental Figure 1**). We also observed the same correlations with parental age (re-do the fraction within each ID instead of overall fraction), with the paternal age effect being more consistent than maternal age effect.

* *Option:* When considering trinucleotide context, the most significant enrichment (and/or correlation with parental age) we observed was for XXX, consistent with the Iceland data.
* *Option:* Consistent with previous results, we observed that the paternal mutation signature is most similar to cancer signature XXX, while the maternal mutation signature is most similar to cancer signature YYY (cite Yufeng study if applicable)
* *Option*: correlation of specific mutations (or 3mers) with ancestry principal components? *e.g.*, was the TCC→TTC mutation rate higher in Europeans (Harris and Pritchard, 2017)?

We previously observed a correlation between unphased *de novo* indels/trio and paternal but not maternal age at conception (cite WGS). We dissected this observation by classifying *de novo* indels into homopolymer runs (HRs), non-HR changes in copy count (CCCs), and non-CCCs (**Table 1**) using the sorting-hat API (see **Methods**). We observed a similar fraction of *de novo* HRs, CCCs, and non-CCCs (**Supplemental Table 1**) across all three methods, consistent with previous results.3 We also found deletion-insertion ratios ranging from X-Y among non-CCCs. We then compared the number of indels per class with parental age (**Figure 2** and **Supplemental Figure 2**). The strongest correlation occurred between CCCs/trio and father’s age, significant for all three phasing methods in addition to the meta-analysis. We also observed a modest correlation between maternal age and *de novo* indels in any class.

* *Option:* Characterize distribution of indels across genome. Are there hotspots similar to Montgomery 2013?
* *Option:* For supplemental table 1, break down by fraction in repetitive elements
* *Option:* Did we find the same deletions near C>A (or whatever the complete genomics/Iceland/Molly finding was?)
* *Option:* Calculate meta-analysis correlation coefficient. However, parental age only explained a minority of the variance in CCC count per trio, suggesting that most arise through mechanisms independent of parental age
* *Option:* predict the region-based indel background mutation rate, possibly use in case-model and control-model comparison
* *Option:* Repeat analysis separating repetitive vs non-repetitive elements
* *Option:* Test NHEJ mechanism directly through nearby palindrome enrichment

**Conclusion**

We phased *de novo* variants to gain a deeper understanding of germline indel mutagenesis. Phasing with low-coverage long-read sequencing was successful, illustrated by our ability to phased four times as many variants as short-read sequencing. We replicated previously defined *de novo* SNV associations, as well as the predisposition of deletions among *de novo* non-CCC indels. The latter observation is consistent with the hypothesis that most non-CCCs arise with double-strand break repair through non-homologous end-joining.

We observed the novel but expected correlation between *de novo* CCC indels and father’s age at conception. These results were highly significant across three phasing methods. Because CCCs are hypothesized to arise during polymerase slippage, the paternal age association is consistent with spermatogenesis-associated DNA replication errors. However, paternal age explained <10% of variance in *de novo* CCC counts per trio, suggesting there are other likely polymerase-mediated events that are contributing to CCCs. We also observed trends supporting a possible role for both maternal and paternal age in the development of other types of *de novo* indels, including a trend for *de novo* HRs (which is also hypothesized to be associated with polymerase slippage). Finally, we observed a weak correlation between maternal age and any *de novo* indel, suggesting that indels arise through events other than replication-associated errors.

Taken together, we have demonstrated that grouping *de novo* indels into HRs, CCCs, and non-CCCs provides a valuable framework for interpreting indel mutagenesis, and have developed an API to make this easier. The CCC-paternal-age correlation is consistent with hypothesized mechanisms but explains a low percentage of indel mutagenesis, suggesting there are other, as of yet unknown, mechanisms at play.

**URLS**

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

Sorting-hat: <https://pypi.org/project/sorting-hat/>

**Acknowledgements**

The results analyzed and published here are based in part on data generated by Gabriella Miller Kids First Pediatric Research Program projects phs001138.v2.p2/phs001194.v2.p2, and were accessed from dbGaP (www.ncbi.nlm.nih.gov/gap). This work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

**Funding**

This work was supported by the National Institute of Dental and Craniofacial Research Interdisciplinary Training in Systems and Developmental Biology and Birth Defects [T32HD075735 to F.R.], Mount Sinai Medical Scientist Training Program [5T32GM007280 to F.R.], and National Heart, Lung, and Blood Institute (NHLBI) grants for the Pediatric Cardiac Genomics Consortium [U01-HL098188, U01-HL131003, UM1-HL098147, U01-HL098153, U01-HL098163, UM1-HL098123, UM1-HL098162, UM1-HL128761, UM1-HL128711].

**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. Garcia-Diaz, M. & Kunkel, T. A. Mechanism of a genetic glissando\*: structural biology of indel mutations. *Trends Biochem. Sci.* **31,** 206–214 (2006).

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

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

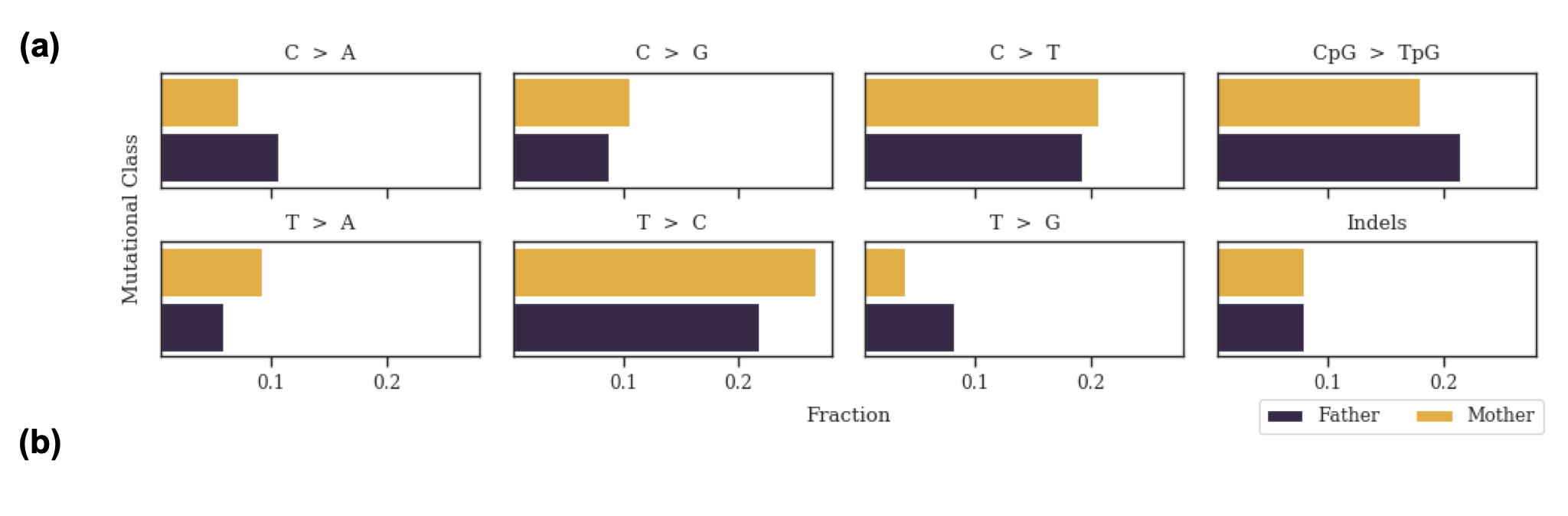
7. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *bioRxiv* 085050 (2016). doi:10.1101/085050

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

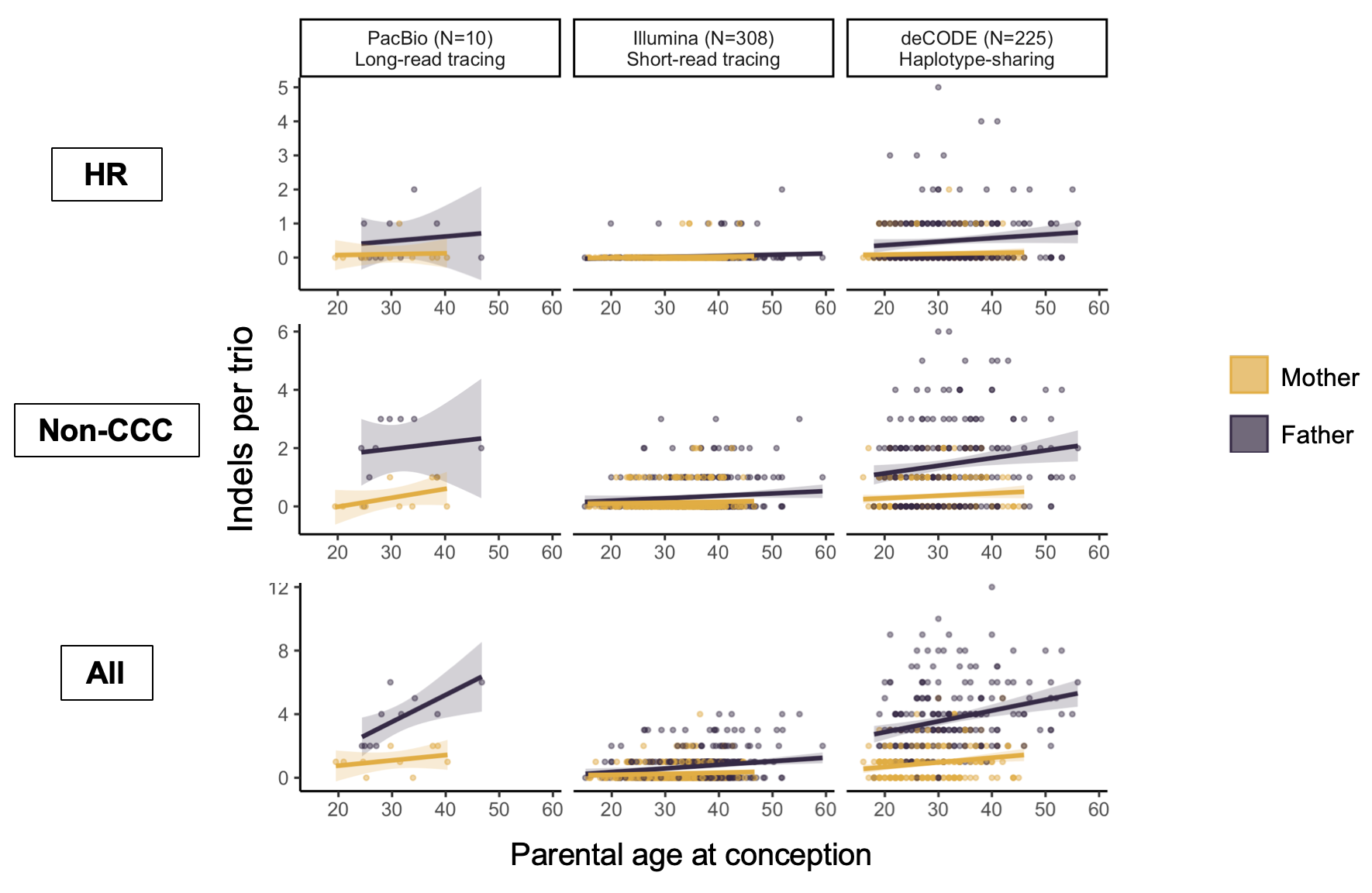
9. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

10. Smit, A., Glusman, G. & Hubley, R. RepeatMasker 4.0. Available at: http://www.repeatmasker.org/.

**Supplemental information**



**Supplemental Figure 1. Replication of previous *de novo* SNV patterns with long-read phasing by read-pair tracing.** **(a)** Fraction of *de novo* variants (DNVs) in major classes phased to each parent. **(b)** Correlations between DNVs in each class and parental age.



**Supplemental Figure 2.** Number of HR, non-CCC, or all indels per trio (y-axis) as a function of parental age (x-axis).