**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 implemented two methodological improvements and leverage these to dissect indel mutagenesis mechanisms. First, we assigned *de novo* variants to parents-of-origin (*i.e.*, phasing) with low coverage long-read whole genome sequencing, achieving better phasing compared to short-read sequencing (medians of 84% and 23%, respectively). Second, we wrote an application programming interface to classify indels into three subtypes by sequence context. Using these methods and three cohorts with different phasing methods (*Ntrios*=540), we observed that one *de novo* indel subtype (change in copy count) was significantly correlated with father’s (*P*=7.1x10-3) but not mother’s (*P*=0.45) age at conception. We replicated this effect in three cohorts without *de novo* phasing (*Ppaternal*=1.9x10-9, *Pmaternal*=0.61, *Ntrios*=3,391). Although this is consistent with polymerase slippage during spermatogenesis, we did not find an association with replication timing and further note that paternal age explained <10% of variance in *de novo* CCC indels. These results suggest that spermatogenesis-specific events have a minor role in CCC indel mutagenesis, one not observed for other indel subtypes or for maternal age in general. 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

**Association with replication timing**

Six embryonic stem cell (BG02 line) Repli-seq bigwig tracks were downloaded from ENCODE.11,12 *De novo* indels were overlapped with each of the six tracks and read depth was standardized within each experiment. To avoid ambiguous replication timing associations, indels that could be clearly assigned to one of six tracks (*i.e.*, read depth ≥1 standard deviation higher than the next best track) were kept for downstream associations.

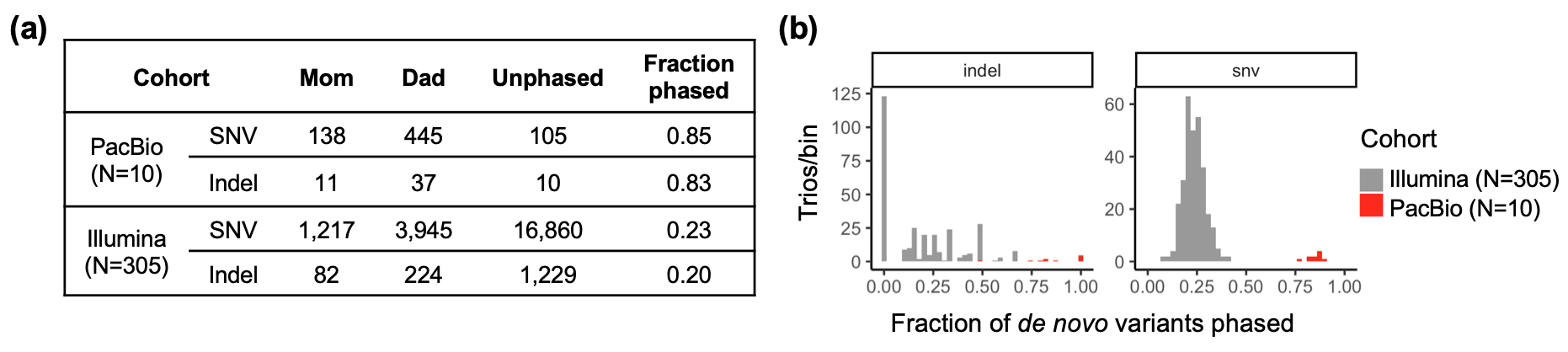
**Statistics**

Correlations were quantified with the Spearman’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 final meta-analysis p-value.

A mixed effects Poisson regression was also performed as secondary evidence and to provide models for scatterplot fitting using the glmer function from lme4.13 Parental age slope and intercept were considered cohort-specific random effects, while cohort was a fixed effect (model formula: Indels/trio ~ parental age + (parental age | cohort) ). For Poisson models, all variance/mean ratios were <2.

**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).

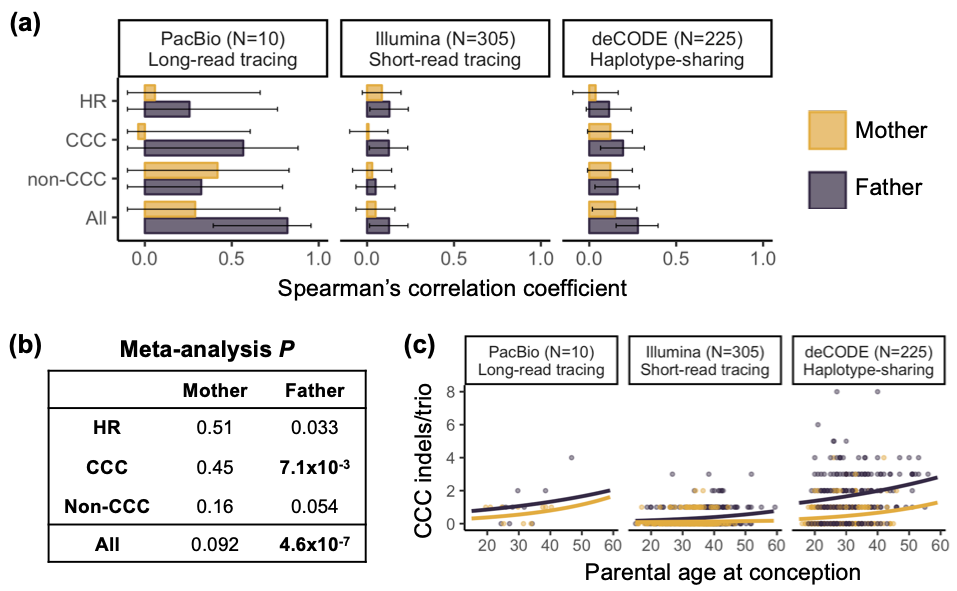


**Figure 1. The fraction of de novo variants (DNVs) phased with Illumina short-read or PacBio long-read sequencing through read-pair tracing.** **(a)** The total number of SNVs and Indels phased in each cohort. **(b)** Histograms of the fraction of de novo indels

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 We observed the same directionality in all three variant classes (**Supplemental Figure 1**).

Others previously observed a correlation between unphased *de novo* indels/trio and paternal but not maternal age at conception.5 We corroborated and dissected this observation in cohorts representing three types of phasing, PacBio long-read and Illumina short-read read-pair tracing (described above) as well as indels phased with three-generation haplotype phasing.5 We first classified *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 percent of *de novo* HRs (5-13%), CCCs (45-48%), and non-CCCs (39-48%) across all three methods, consistent with previous results (**Supplemental Table 1**).3 We found the highest deletion/insertion ratio in non-CCCs (range 9:1 to 24:1), consistent with previous results and mutagenesis during non-homologous end-joining repair.3 We then compared the number of indels per class with parental age (**Figure 2** and **Supplemental Figure 2**). In a meta-analysis of all approaches, the only statistically significant correlation occurred between CCCs/trio and father’s age (two-sided meta-analysis *P*=7.1x10-3). This result was confirmed with Poisson mixed effects regression across all three cohorts (*P*=1.5x10-3).

In order to derive further insight into the *de novo* indel genesis mechanisms, we sub-divided the CCCs. Despite the lower number of *de novo* CCC indels identified in repeat-masked10 genomic regions (SINEs, LINEs, LCRs, *etc*.), *de novo* CCC indels in repetitive regions had significant correlations with paternal age (two-sided meta-analysis *P*=0.02, 233 paternal indels) while CCCs outside of repeats did not (*P* =0.14, 278 paternal indels). We also investigated if the number of *de novo* CCC indels in fathers was associated with replication timing in embryonic stem cells, building on previous correlations with SNV data. However, we observed no association between father’s age and the fraction of paternally-phased *de novo* CCC indels (**Supplemental Figure 3**).



**Figure 2. Correlation between parental age and indel classes across three phasing methods.** **(a)** Spearman’s correlation coefficient and 95% confidence intervals between maternal or paternal age at conception and three indel classes for three methods. **(b)** Fisher’s method meta-analysis of correlation p-values shows a significant correlation between paternal age and CCC indels. **(c)** Scatter plots illustrating the correlation between age at conception and the number of CCC indels, fitted with mixed effects Poisson regression (**Supplemental Figure 2** for HR, non-CCC, and all indel scatter plots). HR=homopolymer run, CCC=change in copy count (excluding HRs), non-CCC=indels not classified as HR or CCC

To further corroborate these findings, we tested the correlation between parental age and *de novo* indels in three trio cohorts with unphased *de novo* variants, all ascertained with Illumina short-read sequencing at a depth of 30x. Consistent with findings in the phased data, we observed the strongest evidence for an association between all *de novo* indels and paternal (two-sided meta-analysis *P*=1.4x10-14) but not maternal (*P*=0.04) age at conception (**Supplemental Figure 4**). We observed statistically significant associations between father’s age and both CCC and non-CCC indels, with CCCs having larger effect sizes. There was no association with maternal age for any *de novo* indel class and no association between number of HRs and either parent.

**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). *De novo* indel identification is notoriously challenging, typically having 85% PCR true positive rates. However, our results, which represent the largest aggregation of phased and unphased *de novo* indels to-date, are robust across thousands of WGS trios ascertained through multiple centers, pipelines, sequencing methods, and phasing methods.

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).

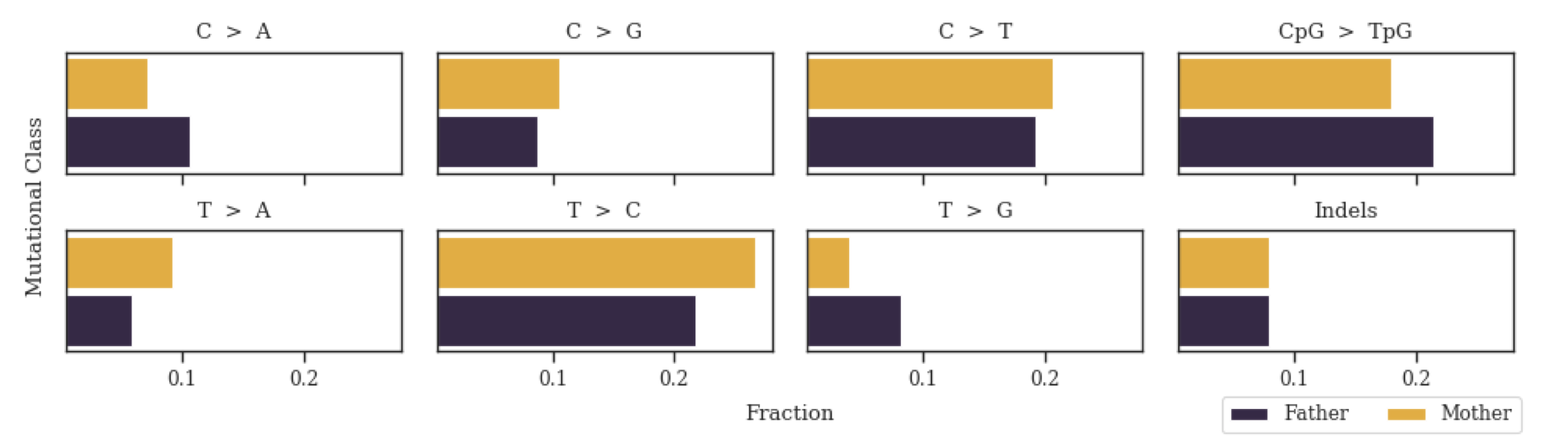
10. Smit, A., Glusman, G. & Hubley, R. RepeatMasker 4.0. Available at: http://www.repeatmasker.org/.

11. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

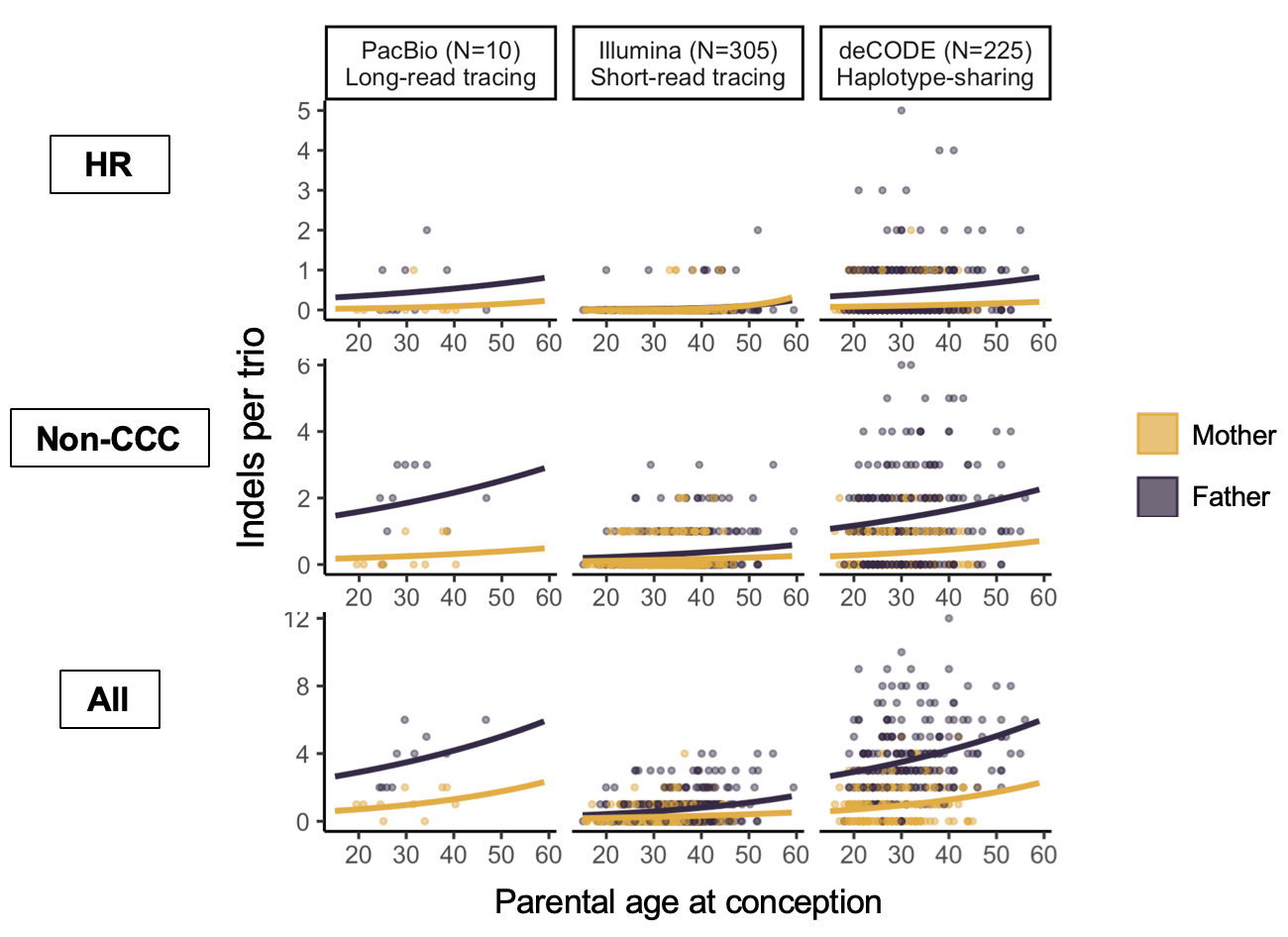
12. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46,** D794–D801 (2018).

13. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67,** 1–48 (2015).

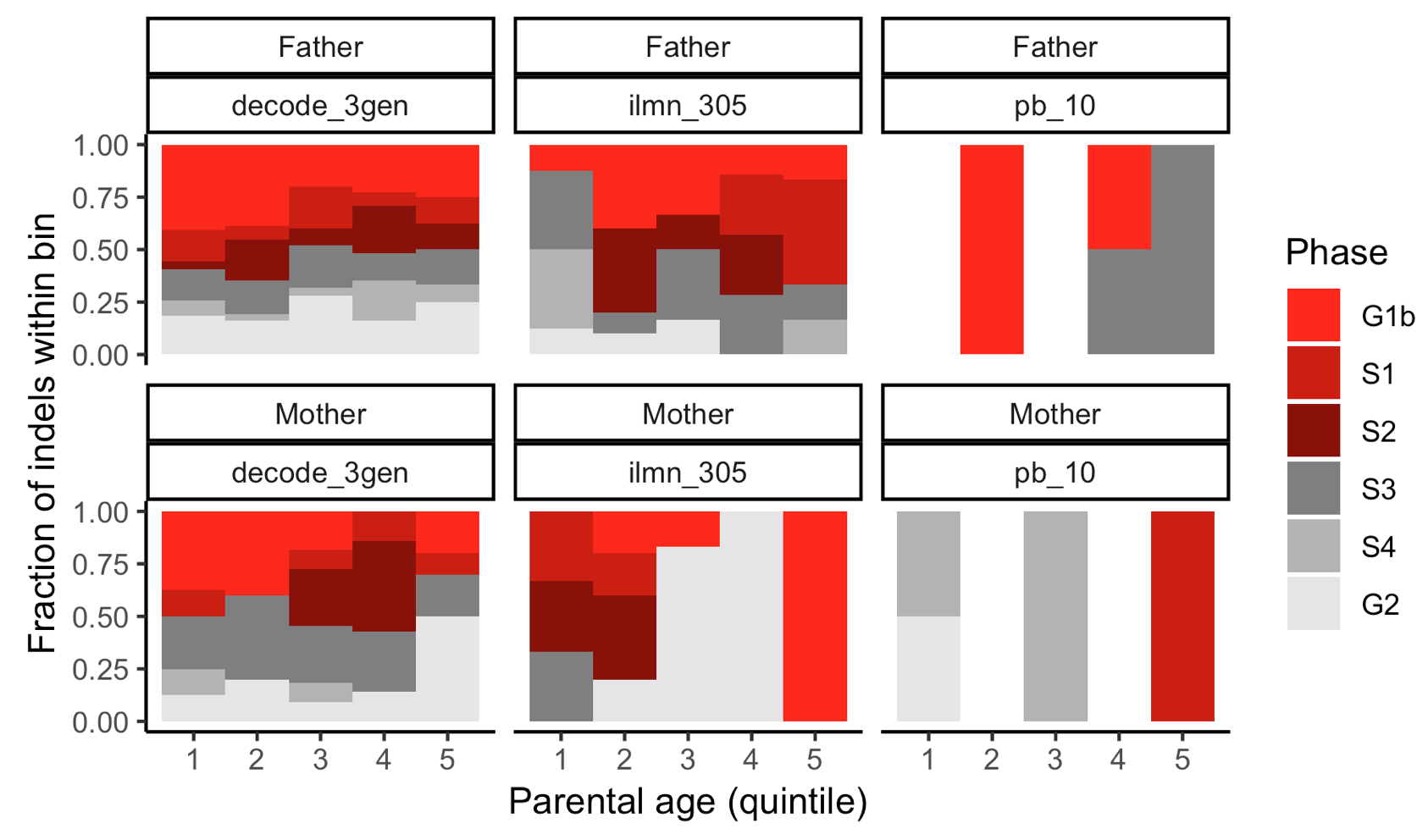
**Supplemental information**



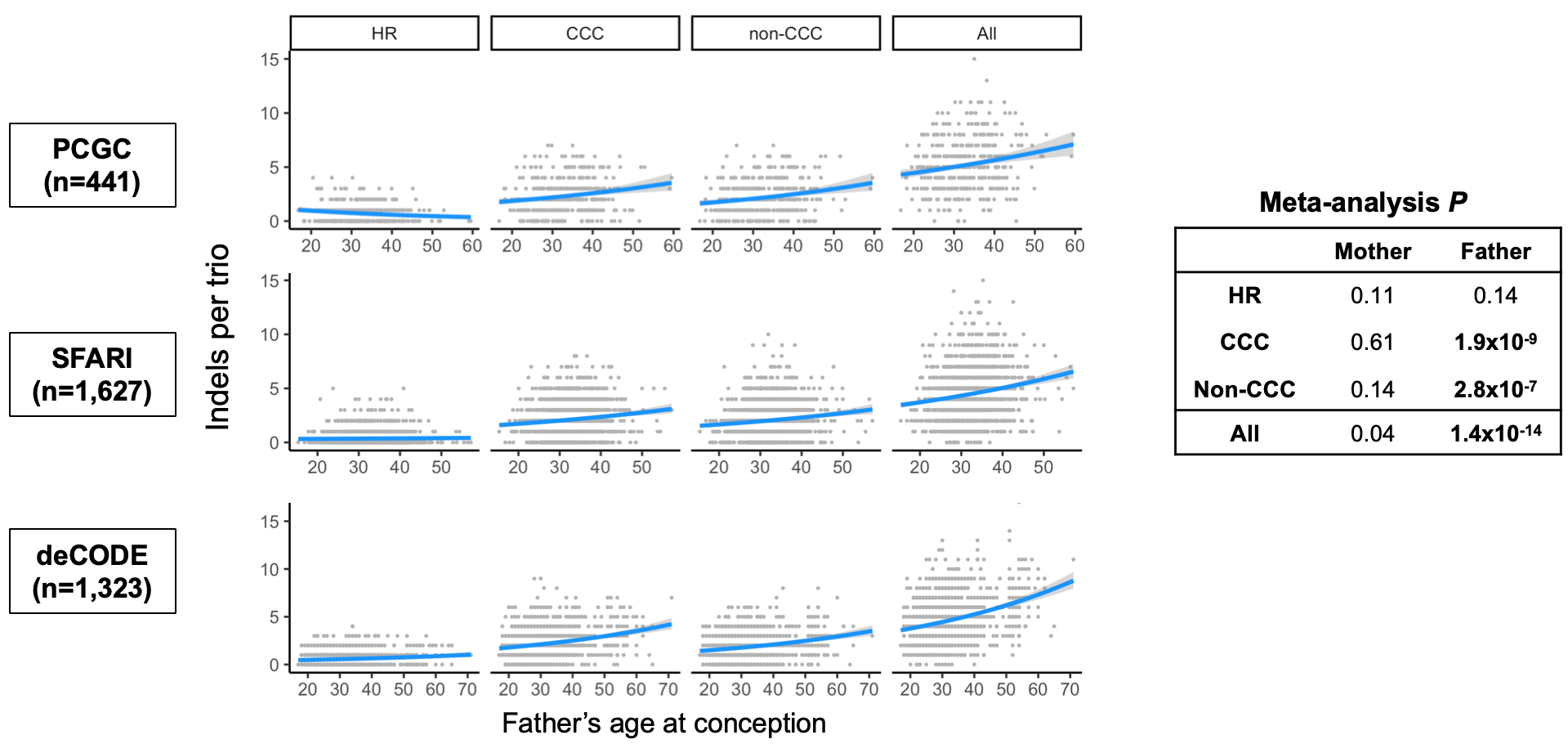
**Supplemental Figure 1. Replication of previous *de novo* SNV patterns with long-read phasing by read-pair tracing.** Fraction of *de novo* variants (DNVs) in major classes phased to each parent.



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



**Supplemental Figure 3**. **No association between number of *de novo* CCC indels and replication timing in embryonic stem cells.** Parental ages were binned into quintiles and (1=youngest, 5=oldest). Within each bin, the fraction of *de novo* CCC indels falling in each of six replication time points (colors) was plotted. There was no increase in the fraction of *de novo* CCC indels in early-replicating regions (*i.e.*, G1b, S1) with increasing paternal age.



**Supplemental Figure 4**. **Correlation between father’s age at conception and number of unphased indels per trio.** Trios without phasing information reveal correlations consistent with a pronounced paternal age effect in both CCCs and non-CCCs, in contrast to phased *de novo* indels where only CCCs where significantly correlated. P-values were obtained from Poisson multiple regression coefficients and meta-analyzed with Fisher’s method.