Results

**Phasing *de novo* variants**

* Phasing with read-pair tracing depends on read length
* Previous read-pair tracing with short-read sequencing phased 20% of *de novo* SNVs (cite 231 illumina trio study, Iceland)
* At the other extreme, 3-generation haplotype phasing was successful for almost all de novo variants, phasing a mean of x% de novo indels and y% de novo SNVs (cite iceland)
* We achieved similar phasing results with Illumina short-read phasing (**Figure 1**, mean 23% phased)
* In contrast, we phased 84% of DNVs with low-coverage (depth <10x) long-read phasing
* *Option:* we observed (no) bias in the types of variants phased (compare fraction phased per ID for different variant classes)

**Replication of previous *de novo* SNV associations**

* First sought to replicate previous *de novo* SNV associations with parental age using an orthogonal phasing technology
* Previously, C>A and T>G DNVs were observed to have a significant paternal enrichment, while C>T DNVs were observed to have a significant maternal enrichment (cite Iceland).
* 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 (need to re-do the fraction within each ID), with the paternal age effect 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)?

**Exploration of *de novo* indel mutagenesis**

* Builds on previous observation of indel correlation with paternal but not maternal age
* First developed an API that classifies indels based on sequence context (sorting-hat, see **Methods**)
* Total number and fraction of HRs, CCCs, and non-CCCs that were insertions and deletions, inside and outside of repeats, for the three methods
  + Also, maternal vs paternal
  + The predisposition of deletions among non-CCCs was consistent with non-homologous end-joining repair (*option:* test this mechanism directly with nearby palindrome enrichment)
* *Option:* Distribution across genome
* *Option:* Did we find the same deletions near C>A (or whatever the complete genomics/Iceland/Molly finding was?)
* We then compared the number of indels in the three classes (as well as all indels) with parental age, finding the strongest correlation between CCCs per trio and father’s age (**Figure 1** and **Supplemental Figure 2**). The correlation p-value were highly significant for CCCs (*P*=9.1x10-5)
  + *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

We previously observed a correlation between indels/trio and paternal, but not maternal age (cite WGS). In order to build on this observation

**Conclusion**

Interpretation of results

* Phased *de novo* variants to gain a deeper understanding of indel germline mutagenesis
* Phasing with low-coverage long-read sequencing was highly successful
  + Phased four times as many variants as short-read sequencing
  + Achieving closer to 100% phasing could be possible with longer read lengths (in this study, median PacBio read was 7,000 bp)
* We replicated previously defined de novo SNV associations
  + Exception was the paternal X>Y de novo, which was only nominally significant in the Iceland study and for which we observed an opposite directionality
* The predisposition of deletions among non-CCCs was consistent with non-homologous end-joining repair (*option:* test this mechanism directly with nearby palindrome enrichment)
* We observed the novel but expected correlation between CCC indels and father’s age at conceptus
  + Highly significant across three different phasing methods
  + CCCs are hypothesized to be caused by polymerase slippage
  + The paternal age association is consistent with DNA replication during spermatogenesis (?spermiogenesis) and the implied polymerase slippage during replication
    - However, paternal age explained <10% of variance in *de novo* CCC counts per trio
  + We also find trends supporting a possible role for both paternal and maternal age across other variant classes
  + HRs are also associated with polymerase slippage
  + We also observed a weak correlation between maternal age and any de novo indel

Implications of results

* We have demonstrated that grouping *de novo* indels into HR, CCC, and non-CCC classes, possibly through the sorting-hat API, provides a valuable framework for interpreting indel mutagenesis at the scale of a single generation
* The fact that paternal age explained <10% of variance in *de novo* CCCs suggests that polymerase slippage during DNA replication accounts for only a minority of indel appearances
  + This cannot be explained by replications before puberty, since the paternal genome experiences roughly the same number of mitosis replications post-puberty every 1.5 years
  + This suggests there could be additional mechanisms involved in *de novo* indel formation, such as the type or relative concentration of polymerase reprised for DNA replication or repair.
* *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
* 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)?

Story: (1) first to phase *de novo* variants with PacBio sequencing, (2) replicated previous *de novo* SNV results, (3) demonstrated that the class of *de novo* indels most associated with DNA replication (CCCs) are mostly explained by paternal age, consistent with spermatogenesis

When classifying SNPs, results were consistent with Iceland paper

* correlation between paternal age and number of de novos
* C > T more maternal than paternal
* C > A more paternal than maternal
* Paternal age effect more consistent across mutational classes than maternal age effect

CCC indels were positively correlated with paternal age and not correlated with maternal age – confirms the idea that CCC mutations are related to polymerase slippage during replication

|  |  |  |  |
| --- | --- | --- | --- |
| *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 |

Note: only one of the non-CCC indels was an insertion, rest were deletions. See if this is a pattern with Illumina data as well as Netherlands/Oscar’s data

In a multiple variable regression with both parental age and fraction of SNVs phased to any parent, the fraction of SNVs phased was not a significant predictor of number of indels per proband in the Illumina data (all P>0.12) but was significantly negatively correlated with the number of maternally phased indels/proband in PacBio data (for All indels *P*=0.0078, for CCC indels *P*=0.014).

All indel class correlations with parental age were positive

Correlation p-values and plots with parental age run with regressions\_age.R

P-value meta-analysis done with metanalysis\_indels.py