Personalized Phased Diploid Genomes of the EN-TEx Samples

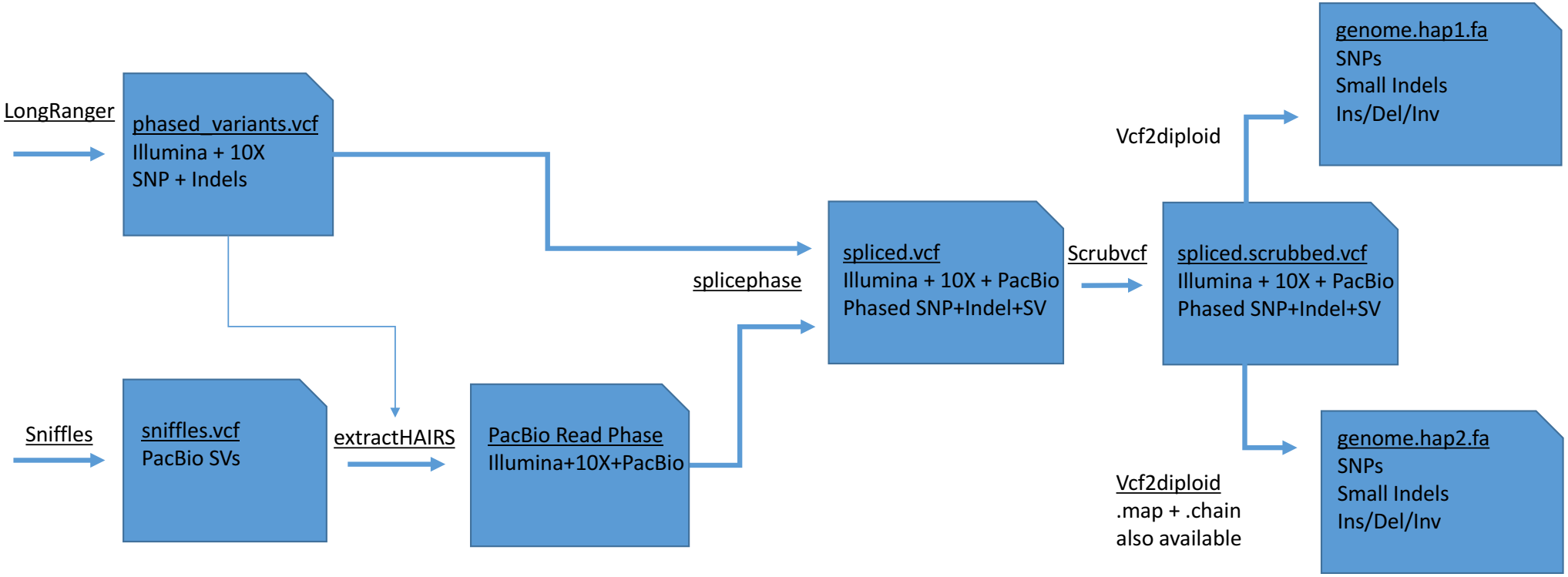
Michael Schatz, Fritz Sedlazeck, Han Fang, Maria Nattestad, Ruibang Luo, Srividya Ramakrishnan, Charlotte Darby, Philipp Rescheneder, Alex Dobin, Carrie Davis, Ashwin Prakash, Anna Vlasova, Alessandra Breschi, Roderic Guigo, Tom Gingeras

Nov 6, 2017

ENTex Update

ENTex Phased Diploid Genomes Version 1.0

http://labshare.cshl.edu/shares/schatzlab/www-data/encode/diploid/2017.10.26/

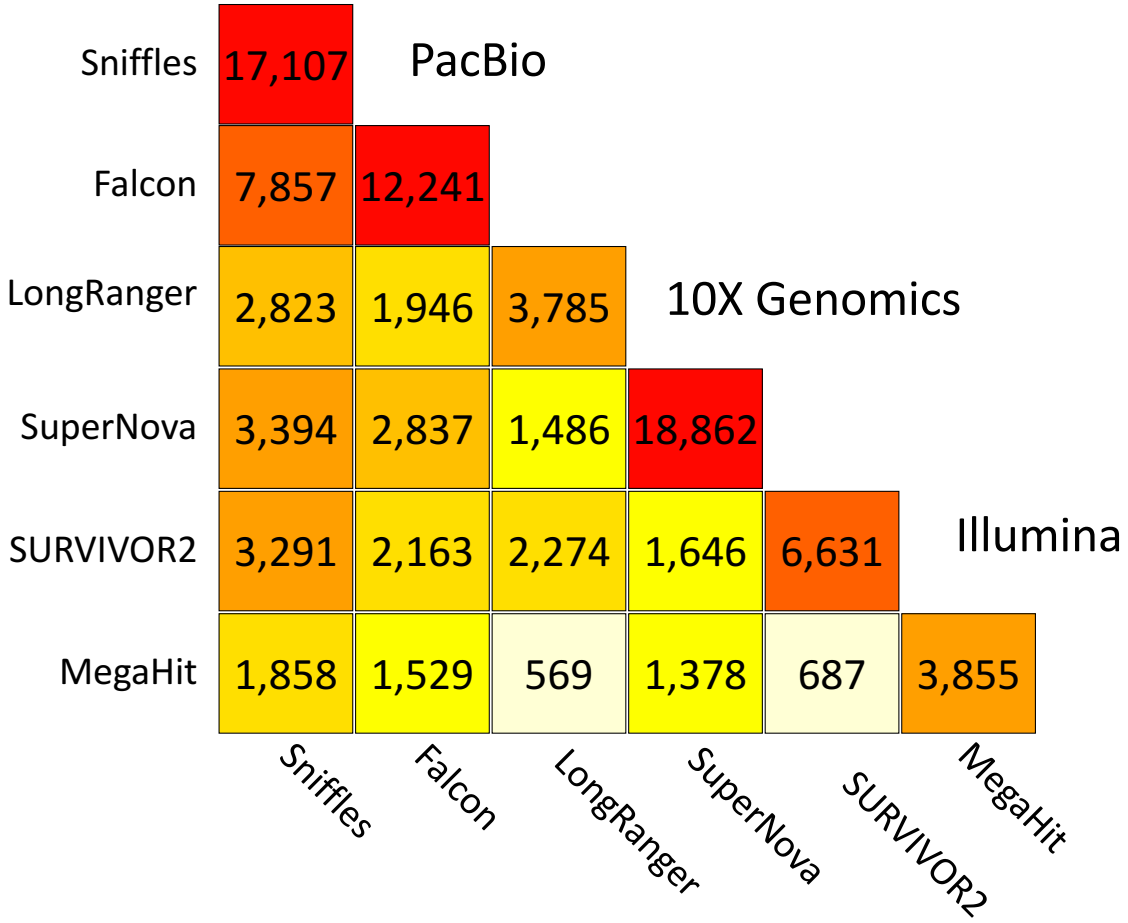


SV phasing software available at: https://github.com/schatzlab/phase_sniffles (name will change soon)

Why this approach?

- ***Major alternative is phased diploid de novo assembly followed by whole genome alignment***
  - Goal: Phase and find all variants directly from assembler, including very long insertions or other complex SVs!
  - aka FALCON + Assemblytics (Note I was corresponding author on both of these papers)
- ***Practical and Engineering Issues for assembly:***
  - Assembly takes days to weeks, mapping is done "overnight"
  - Sourcecode is specific to PacBio only, but we will use multiple technologies including ONT :)
  - After assembly, we still have to align to reference to find variants, annotate genes, etc
    - Repeats that are too long to reliably map reads, are also too long to reliably assemble contigs
- ***Potential sensitivity issues for assembly:***
  - Deeper coverage is needed over a region to successfully error correct and assemble, worse for het SVs
  - Hard to correctly align and find SVs near the end of contigs, while mapping has consistent performance
  - Variants flanked by repeats will probably fail to assemble, but could be captured by mapping
- ***Potential specificity issues for assembly:***
  - Ends of contigs are unreliable as they have lowest coverage and enriched for repeats
  - Hard to correctly characterize homo. versus het. since both alleles have to assemble and align well
  - Mis-assemblies hard to detect and will cause false variants; can have 3 or more contigs spanning some regions
- ***Empirical result:***
  - We benchmarked 6 SV calling approaches (3 mapping-based, 3 assembly-based), and found mapping (Sniffles) has the highest sensitivity and specificity

Structural Variations Concordance (ENC-002)



Main Diagonal

- Calls per tool

Outer triplets

- Concordance by Technology

Inner triplets

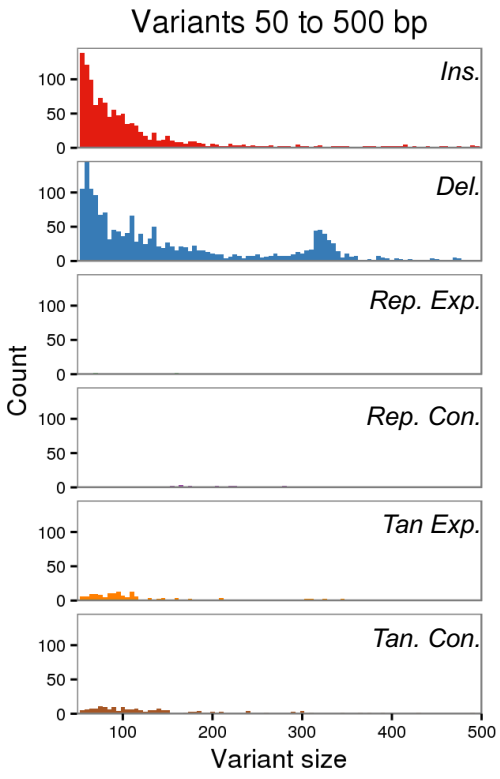
- Concordance by Assembly
- Concordance by Mappers

Overall:

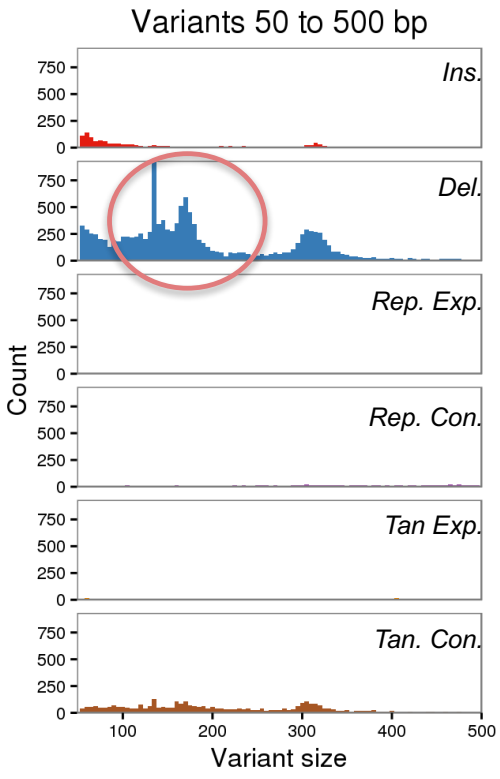
- We need multiple technologies and approaches

Missing Insertions from Short and Linked Read?

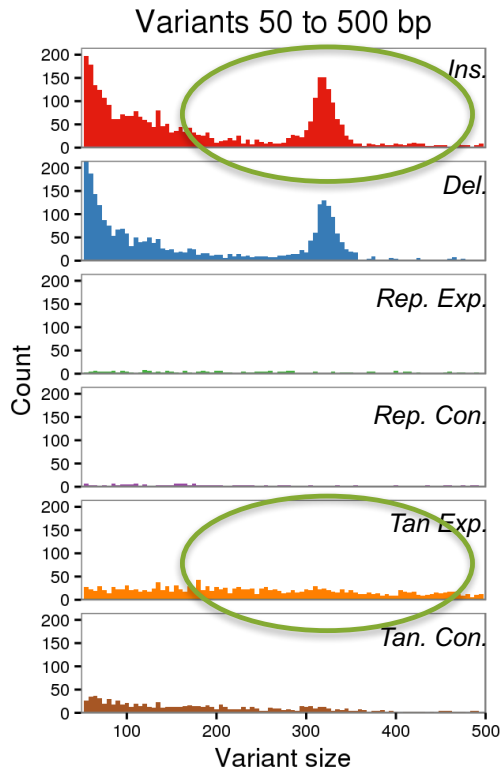
Illumina



10X Genomics



PacBio



Discussion and Future Work

- ***Phasing was performed using 10X genomics only, and the phase block N50 size is around 5Mbp.***
  - Only variants over the main chromosomes as 10X/LongRanger does not work well with alternative chromosomes
  - Not all variants (SNVs/Indels/SVs) are phased if there are not enough flanking SNVs to make a reliable call and will have a 0/1 or 1/0 genotype
  - The boundaries of the phase blocks can be determined by the PS field in vcf file
  - **TODO: Integrate in Hi-C once data quality is confirmed**

- ***For SVs, the vcf file only contains insertions, deletions and inversions that are at least 50bp***
  - Standard variants supported by at least 10 PacBio reads, or at least 5 reads for the sensitive version.
  - The sensitive version identifies a few thousand more SVs, although may have more false positives
  - **TODO: Bench validation to establish sensitivity limits on SVs and indels**

- ***The sequence fidelity of the SV insertions will only be around 90%***
  - The reported sequence is derived from a single raw PacBio read although multiple reads support the SV call
  - **TODO: Developing local assembler to improve the sequence accuracy, augment with additional variant types**

- ***Variant calls have been post-hoc filtered***
  - Remove any overlapping calls using simple left to right scan, preferring SVs to small variants
  - Only includes gender-appropriate chromosomes (male: 1X + 1Y; female: 2X)
  - **TODO: Better resolution of nested SVs and very long insertions**