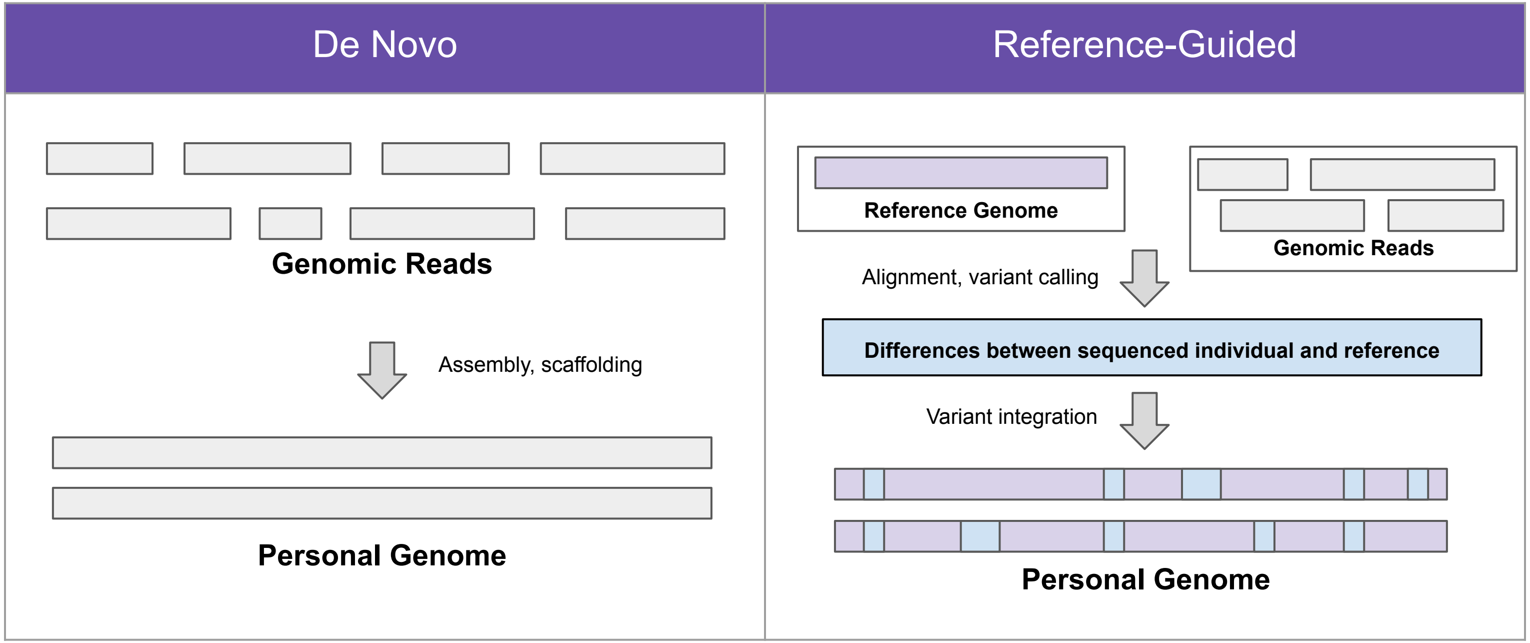
Efficient chromosome-scale haplotype-resolved assembly of human genomes

Melanie Kirsche
October 30, 2019
JHU Genomics Journal Club

Two approaches to personal genome construction



De Novo

Genomic Reads

Assembly, scaffolding

Personal Genome

Reference-Guided

Reference Genome

Genomic Reads

Alignment, variant calling

Differences between sequenced individual and reference

Variant integration

Personal Genome

This is the one we will talk about today.

Come to tomorrow’s joint lab meeting
to hear more about this one

Some Difficulties in De Novo Assembly

- Typically very high coverage is needed to good chromosome-scale contigs
- Assembly software comes with tradeoff of speed vs. assembly quality, so good assemblies take a long time
- De novo scaffolding (e.g., with Hi-C) is prone to introducing misassemblies
- Can’t use known reference variants for phasing

Existing Approaches

- Phasing long reads with short-range data (i.e., mostly linked reads)
  - E.g., Falcon-Unzip, Supernova, WhatsHap
- Trio Binning - get short reads from both of the individual’s parents and assign each long read a haplotype to assemble each set separately
  - Doesn’t work well in regions which are heterozygous in all three individuals
  - Requires sample from parents which isn’t always available

Main claim of this paper

WHdenovo produces phased human assemblies with near-chromosome-length contigs in less than 24 hours given two inputs: ~30x CCS long reads and ~30x Hi-C long-range conformation data

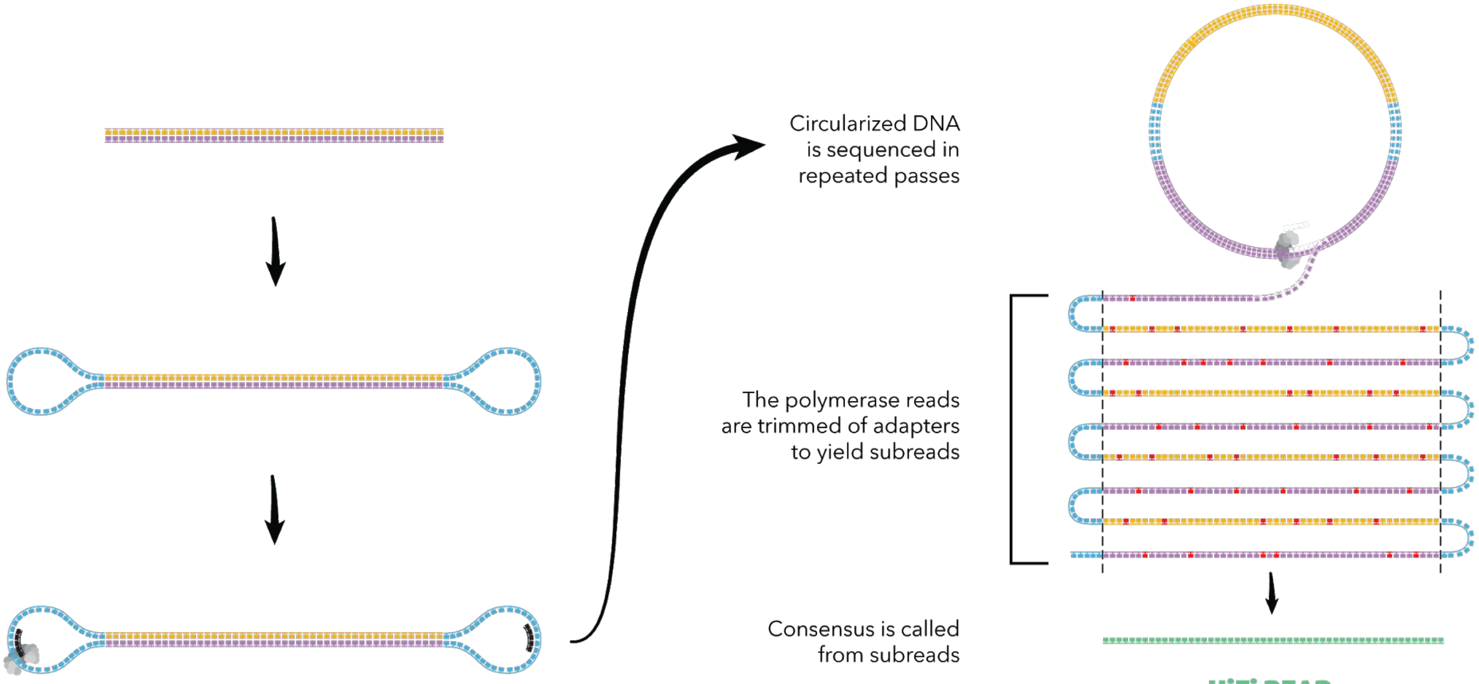
Dataset

HG002 from GIAB consortium

- Very well studied with lots of datatypes
- Lots of other assemblies of it to compare results to
- Possibly the best existing catalog of structural variant calls in an individual

30x CCS and 30x HiC

Circular Consensus Sequencing



Start with high-quality
double stranded DNA

Ligate SMRTbell
adapters and size select

Anneal primers and
bind DNA polymerase

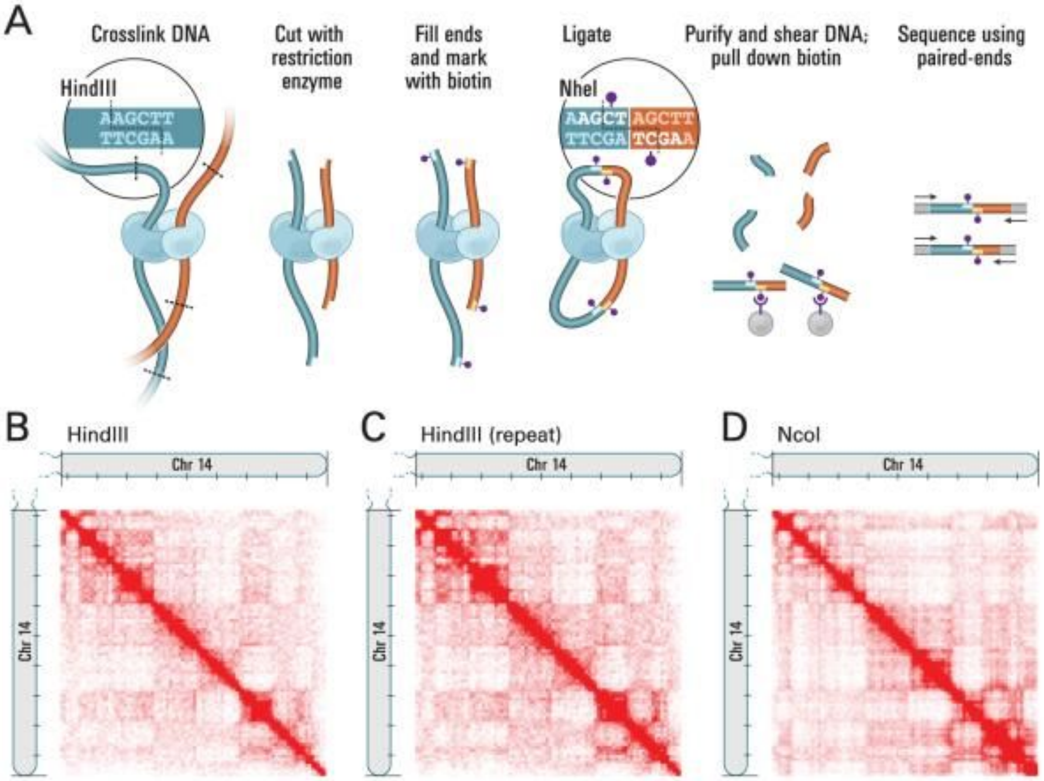
Circularized DNA
is sequenced in
repeated passes

The polymerase reads
are trimmed of adapters
to yield subreads

Consensus is called
from subreads

HiFi READ
(>99% accuracy)

Hi-C


A
Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCTAGCTT
TTCGATCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends
B
HindIII
Chr 14
Chr 14
C
HindIII (repeat)
Chr 14
Chr 14
D
NcoI
Chr 14
Chr 14

Their Pipeline

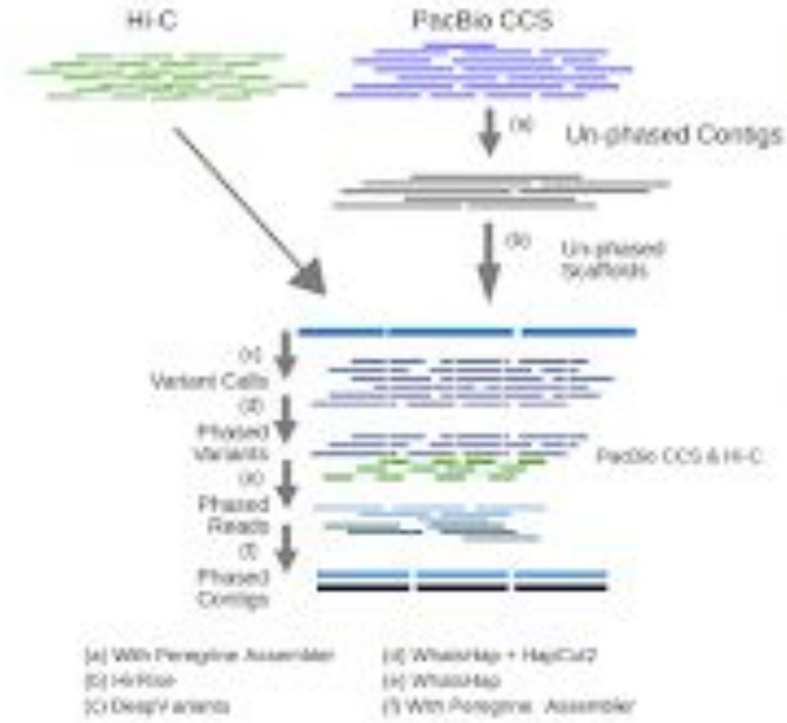


Fig 1. Outline of the assembly algorithm. (a) Assemble CCS reads into unphased contigs. (b) Group and order contigs into scaffolds with Hi-C data. (c) Map CCS reads to scaffolds and call heterozygous SNPs. (d) Phase heterozygous SNP calls with both CCS and Hi-C data. (e) Partition reads based on their phase. (f) Assemble partitioned reads into phased contigs.

Assembly algorithm

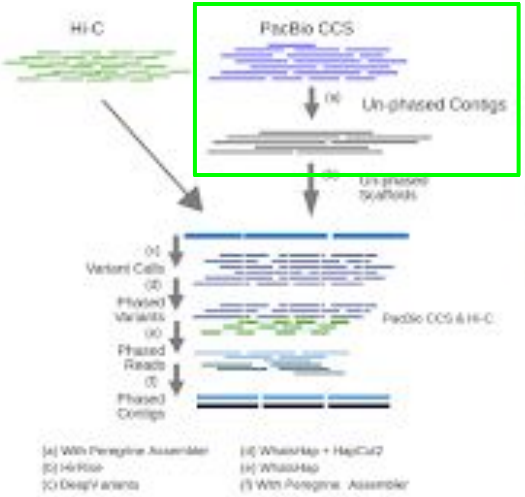
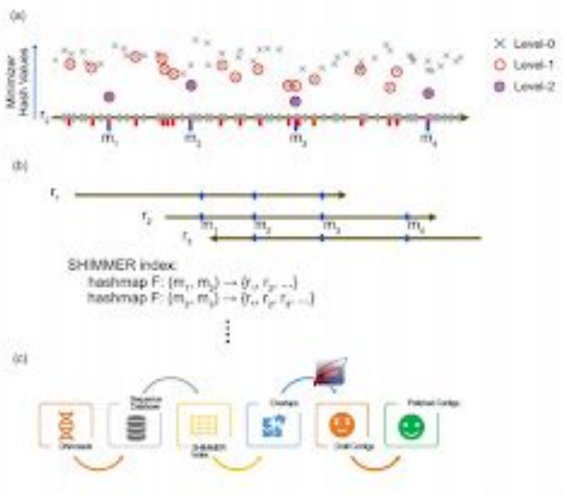


Fig 1. Outline of the assembly algorithm. (a) Assemble CCS reads into unphased contigs. (b) Group and order contigs into scaffolds with Hi-C data. (c) Map CCS reads to scaffolds and call heterozygous SNPs. (d) Phase heterozygous SNP calls with both CCS and Hi-C data. (e) Partition reads based on their phase. (f) Assemble partitioned reads into phased contigs.

Figure 1. SHIMMER index generation and the Peregrine assembler workflow.



(a) The gray tick-marks represent the locations of the level-0 minimizer along a read. The crosses represents the hash value of the minimizers. The level-1 minimizers (red tick-marks and circles) are the local minima of the windows through the neighboring minimizers. Similarly, the level-2 minimizers (blue tick-marks m_1 to m_4, and filled circles) are local minima of the level-1 minimizers over moving windows. (b) For each read, we scan the level-2 minimizers and generate a hash map that maps neighboring minimizer pair to a set of reads to speed up overlap finding. (c) The Peregrine assembler workflow. The overlapping module of Peregrine generates file that is compatible to FALCON assembler's overlap-to-contig modules. After we get the draft contigs from FALCON assembler, we apply the FALCON-sense algorithm to polish the draft contig to increase the contig accuracy.

Scaffolding algorithm/results*

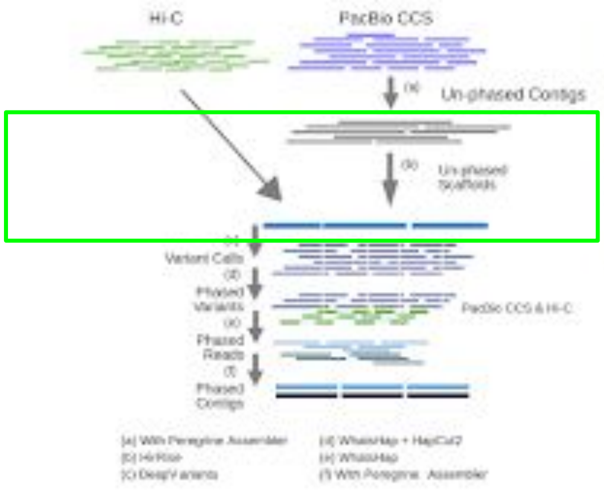


Fig 1. Outline of the assembly algorithm. (a) Assemble CCS reads into unphased contigs. (b) Group and order contigs into scaffolds with Hi-C data. (c) Map CCS reads to scaffolds and call heterozygous SNPs. (d) Phase heterozygous SNP calls with both CCS and Hi-C data. (e) Partition reads based on their phase. (f) Assemble partitioned reads into phased contigs.

*They have the option to also use Ragoo (Alonge et. al.) to scaffold using the reference, which is semi-de-novo

Hi-Rise Scaffolding

1. Mask regions of the genome with too many or too few reads mapping there (about half of the genome)
2. Give each contig pair a likelihood of being adjacent based on number of links and complicated formula
3. Construct graph with edges weighted according to these likelihoods
4. Find confident linear subgraphs with minimum spanning forest and merge those into single contig
5. Repeat steps 2 to 4 until no high-confidence joins remain

Phasing with heterozygous sites

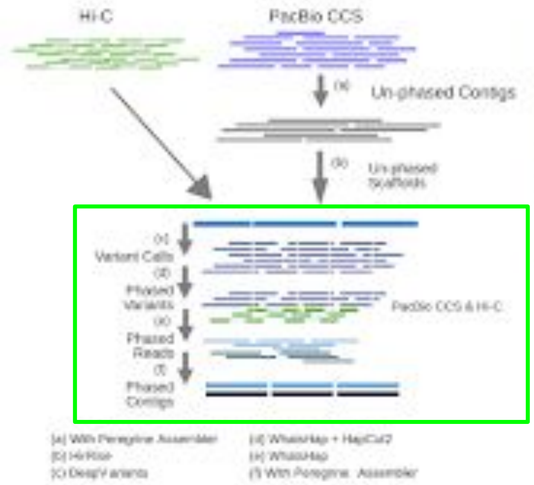


Fig 1. Outline of the assembly algorithm. (a) Assemble CCS reads into unphased contigs. (b) Group and order contigs into scaffolds with Hi-C data. (c) Map CCS reads to scaffolds and call heterozygous SNPs. (d) Phase heterozygous SNP calls with both CCS and Hi-C data. (e) Partition reads based on their phase. (f) Assemble partitioned reads into phased contigs.

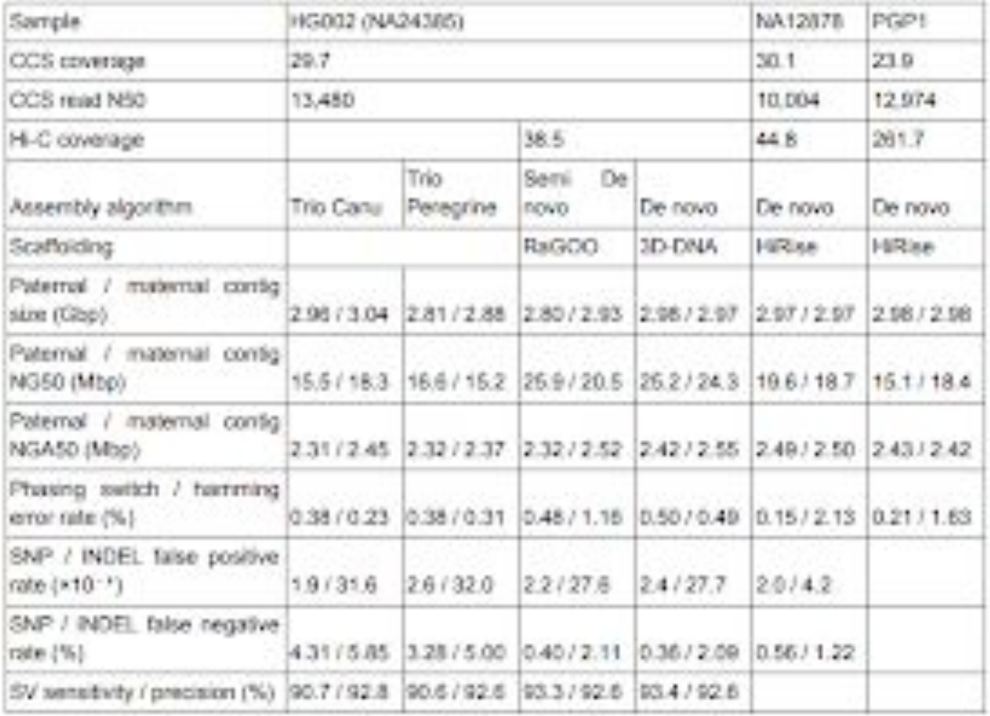
1. Map CCS reads to scaffolds with BWA
2. Run DeepVariant (neural network) to call small variants
3. Phase small variants with WhatsHap and HapCUT2 using Hi-C data (expectation-maximization algorithms)
4. Assign reads to haplotypes based on SNPs they contain
5. Re-assemble the reads on each individual haplotype

Comparison to GIAB gold standard

Used dipcall pipeline to align phased contigs to human reference and get variant calls and compare to published “high-confidence region” variant calls consisting of 2.36 Gb

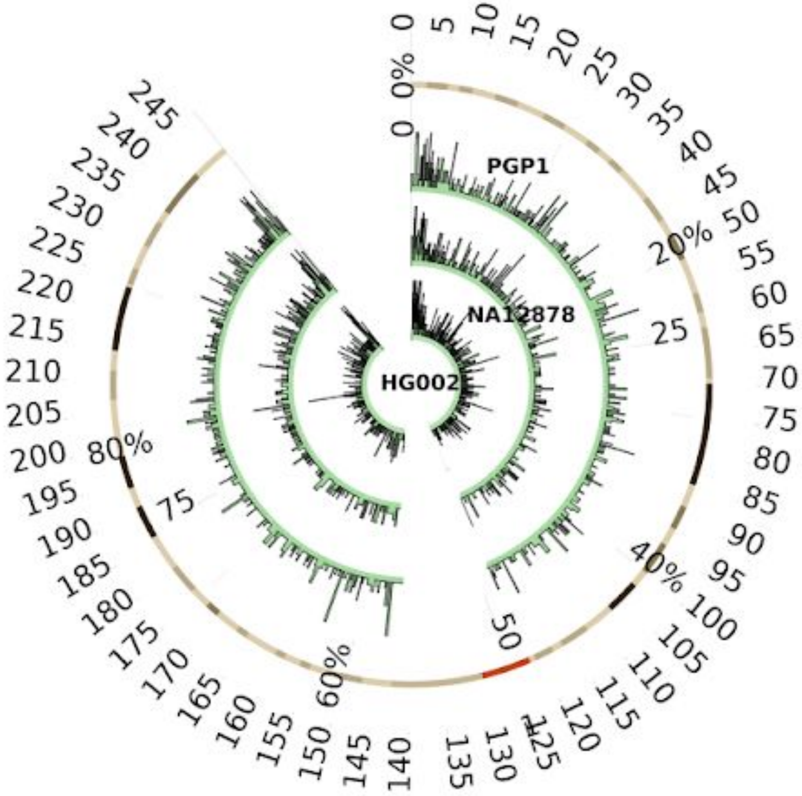
- 5,753 false SNP calls (.19% or called SNPs)
- 65,302 false indel alleles (11.68% of called indels) - 77% of these are 1 bp deletions, which is a known error mode of CCS reads
- Compared SV calls with TruVari - 93.3% sensitivity and 92.6% precision
- “Many” of them are same variants but split into multiple calls

Assembly Contiguity and Quality

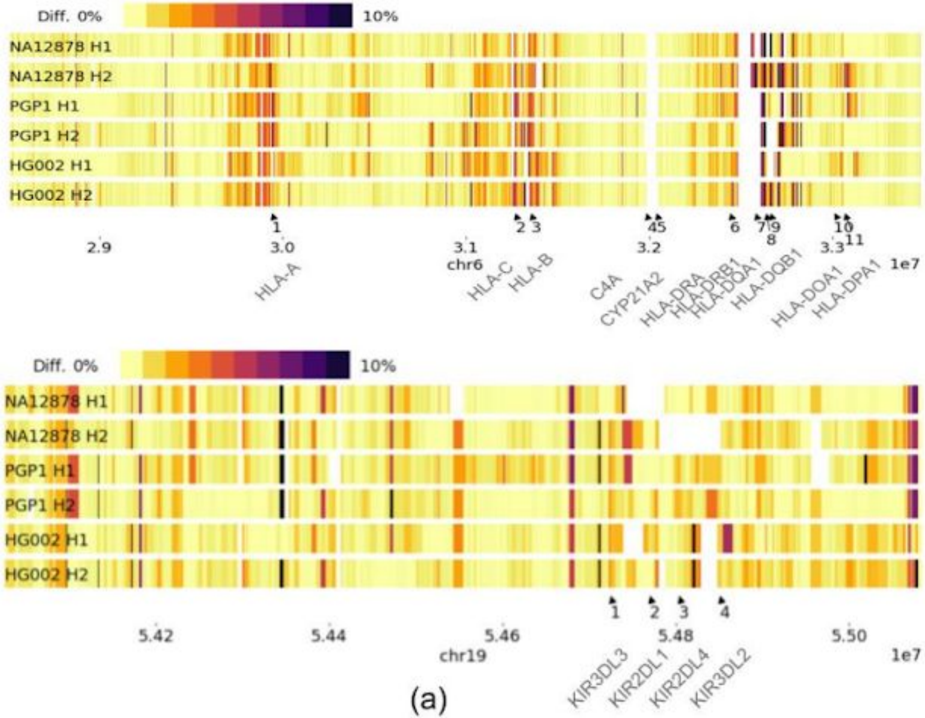


Sample	HG002 (NA24385)				NA12878	PGP1
CCS coverage	29.7				30.1	23.9
CCS read N50	13,480				10,004	12,974
Hi-C coverage			38.5		44.8	261.7
Assembly algorithm	Trio Canu	Trio Peregrine	Semi De novo	De novo	De novo	De novo
Scaffolding			RaGOO	3D-DNA	HiRise	HiRise
Paternal / maternal contig size (Gbp)	2.96 / 3.04	2.81 / 2.88	2.80 / 2.93	2.96 / 2.97	2.97 / 2.97	2.98 / 2.98
Paternal / maternal contig NG50 (Mbp)	15.5 / 18.3	16.6 / 15.2	25.9 / 20.5	25.2 / 24.3	19.6 / 18.7	15.1 / 18.4
Paternal / maternal contig NGA50 (Mbp)	2.31 / 2.45	2.32 / 2.37	2.32 / 2.52	2.42 / 2.55	2.49 / 2.50	2.43 / 2.42
Phasing switch / hamming error rate (%)	0.38 / 0.23	0.38 / 0.31	0.48 / 1.16	0.50 / 0.49	0.15 / 2.13	0.21 / 1.63
SNP / INDEL false positive rate ($\times 10^{-4}$)	1.9 / 31.6	2.6 / 32.0	2.2 / 27.6	2.4 / 27.7	2.0 / 4.2	
SNP / INDEL false negative rate (%)	4.31 / 5.85	3.28 / 5.00	0.40 / 2.11	0.36 / 2.09	0.56 / 1.22	
SV sensitivity / precision (%)	90.7 / 92.8	90.6 / 92.6	93.3 / 92.6	93.4 / 92.6		

Looking at Structural Variant Density



Long phase blocks are important for highly heterozygous regions like HLA and KIR.



(a)

Limitations

- Requires expensive CCS data
- The use of an unphased assembly may cause highly heterozygous regions to be missed

Conclusions

- Haplotype-resolved assembly is important, especially in highly heterozygous regions where keeping a single haplotype discards half the information
- CCS data is enabling simultaneous resolution of large variants (due to length) and small variants (due to accuracy)
- Phasing reads once they have an assembly to map to is more robust to repetitive regions then approaches like trio binning or pre-assembly linked-read phasing
- Aligning and variant calling relative to this individual's assembly rather than a reference makes it easier to assemble regions where the individual is very different from the human reference genome
- Code is here: https://github.com/shilpagarg/WHdenovo