Exploring the use of Nanopore cDNA sequencing for haplotype phasing in F1 hybrids and polyploid plant species

Kin H. Lau and C. Robin Buell
Dept. of Plant Biology, Michigan State Univ., East Lansing, MI
laukin3@msu.edu

Background

Genome assembly is impeded by heterozygosity and polyploidy in the species of interest. As such, genome assembly strategies have traditionally targeted homozygous, diploid reference genotypes and involved collapsing heterozygous regions during assembly. However, accurate prediction of peptide sequence changes due to genic variants requires the identification of cis-linked variants (haplotypes). Recent technological advances such as linked short reads and long sequencing reads have facilitated haplotype-aware genome assembly in heterozygous, diploid individuals, but no simple solutions are available yet for heterozygous polyploids.

We are developing a pipeline to identify gene-level haplotypes using Nanopore (Oxford Nanopore Technologies) cDNA sequencing. We present our progress with data from Col-0 Arabidopsis, a maize F1 hybrid (thus with known haplotypes), and two polyploids: tetraploid potato and hexaploid sweetpotato.

Full-length reads, expression quantification and basecall accuracy of Nanopore cDNA sequencing

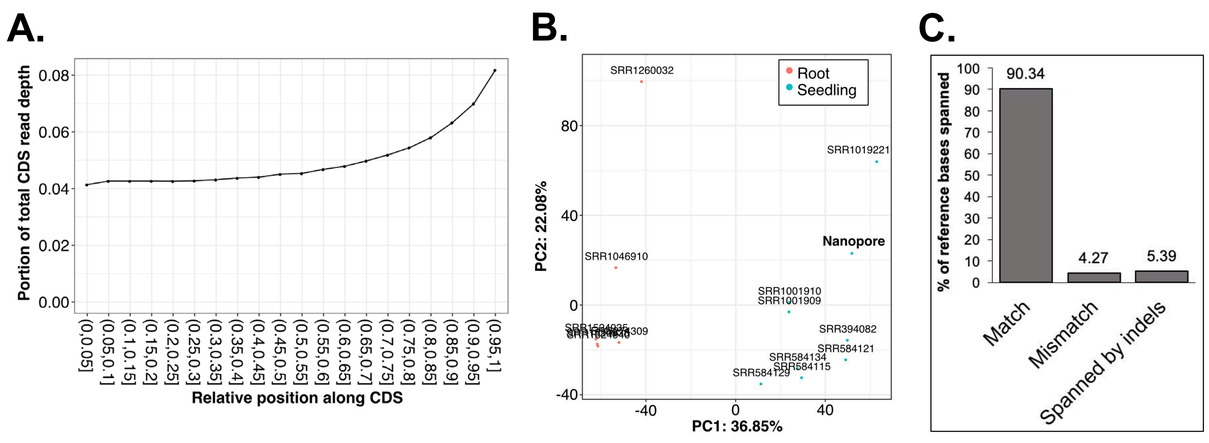


Fig 1. Pilot experiment using inbred *Arabidopsis thaliana*, Col-0. **A.** Coverage along coding sequences (CDS) of representative isoforms by Nanopore cDNA reads. **B.** PCA of log2 RPKM values for public Illumina short-read datasets and log2 reads per million values for a Nanopore library. Only genes with >10 reads aligned from the Nanopore library shown (6437 genes). **C.** Accuracy of Nanopore cDNA reads (avg. Q ≥ 7 filtered) estimated by comparing with the reference genome (TAIR10).

Summary of haplotyping pipeline

1. Call SNPs with genomes (MUMmer) or Illumina reads.
2. Parse alignments from *minimap2* (Li, 2018) for basecalls at SNPs.
3. Count haplotypes for SNPs above read depth cut-off. Low-quality bases or reads that do not overlap a SNP are indicated with a '?'.
4. Assume haplotypes with a minimum # of reads and minimum fraction of total reads are **'real'** (5 reads and 10% in e.g. below).
5. Identify the longest 'real' haplotype. Identify the 'real' haplotypes that do not concur with it; identify the longest haplotype from among these. Repeat until out of disagreeing 'real' haplotypes. These are **'representative' haplotypes** for the locus.
6. **Group** all haplotypes with a specific representative haplotype.

Table 1. Example output for *Zm00001d002501* from a maize F1 hybrid.

Gene ID	Haplotype	Reads	% total reads	'Real'?	Repr. Hap.	Hap. group
Zm00001d002501	00000	52	14.44	pass	repr	0
Zm00001d002501	11111	43	11.94	pass	repr	1
Zm00001d002501	00???	29	8.06	fail	not_repr	0
Zm00001d002501	1111?	23	6.39	fail	not_repr	1
Zm00001d002501	000?0	22	6.11	fail	not_repr	0
Zm00001d002501	11???	20	5.56	fail	not_repr	1
Zm00001d002501	111?1	15	4.17	fail	not_repr	1
Zm00001d002501	0000?	13	3.61	fail	not_repr	0
Zm00001d002501	11?11	13	3.61	fail	not_repr	1
Zm00001d002501	0?000	9	2.50	fail	not_repr	0

Assessing and tuning the haplotyping pipeline using a maize F1 hybrid

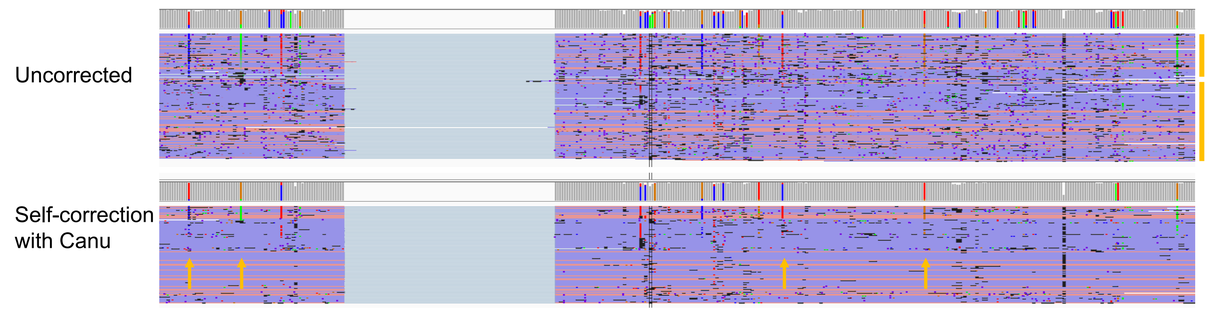
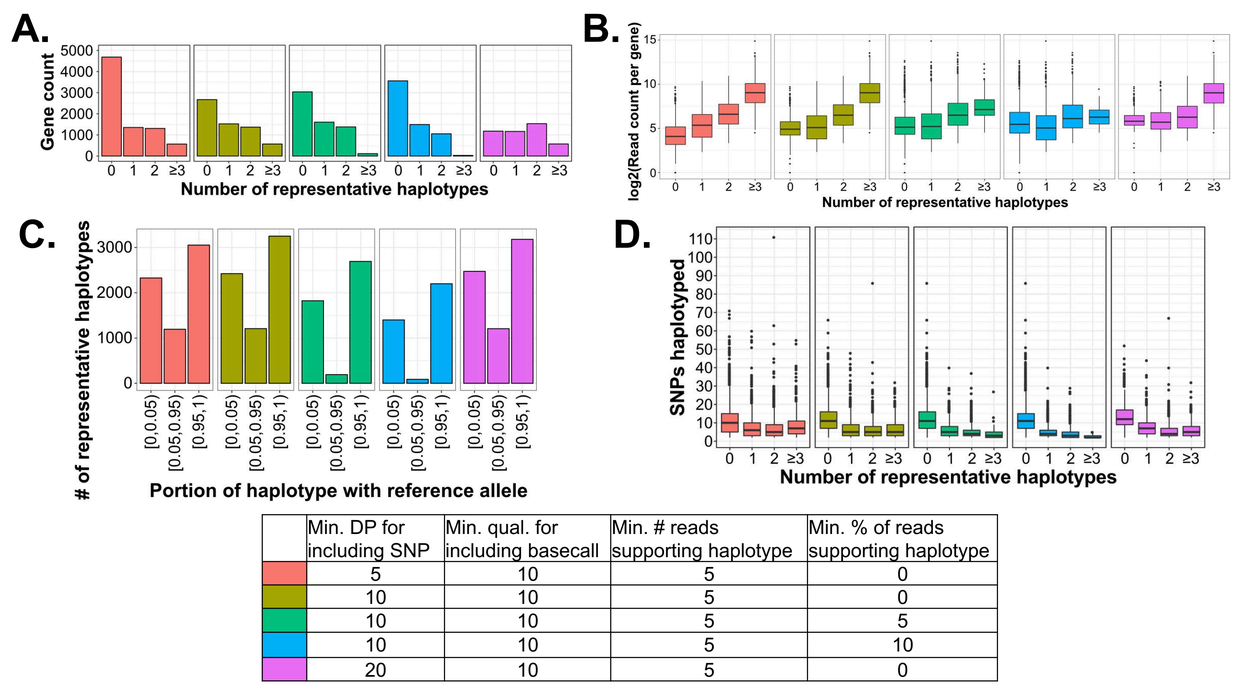


Fig 2. Haplotypes can be inferred with Nanopore cDNA reads. **(Top)** Groups of reads supporting two different haplotypes indicated (right bars). **(Bottom)** Self-correction fixes many errors but also changes real variants, causing false haplotypes; examples are indicated (arrows).



	Min. DP for including SNP	Min. qual. for including basecall	Min. # reads supporting haplotype	Min. % of reads supporting haplotype
	5	10	5	0
	10	10	5	0
	10	10	5	5
	10	10	5	10
	20	10	5	0

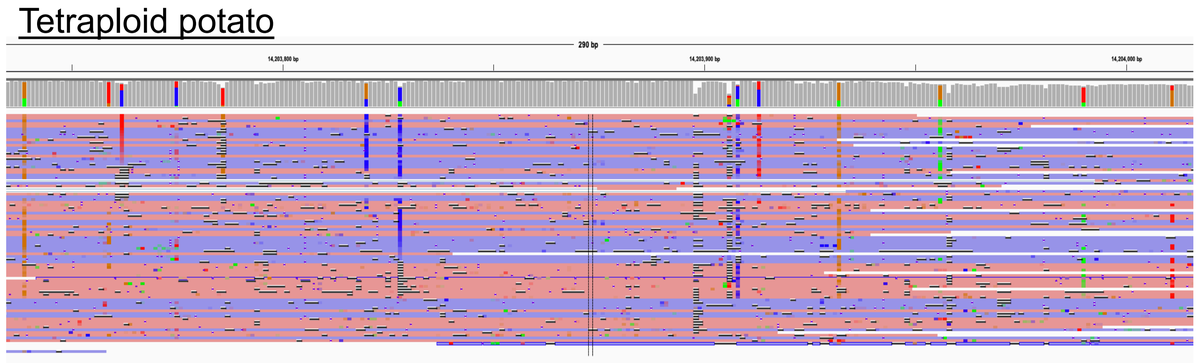
Fig 3. Haplotyping results using different script parameters. **A.** Distribution of number of representative haplotypes per gene; this should be ≤2 in a diploid. **B.** Filtering on the min.% of reads supporting haplotype curbs increase in # of representative haplotypes caused by high expression. **C.** Filtering on the min.% of reads supporting haplotype results in fewer representative haplotypes with a mixture of reference and alternate alleles, which are not expected in an F1 hybrid of inbred parents. **D.** High numbers of overlapping SNPs make identifying representative haplotypes more difficult.

Table 2. Length of longest representative haplotype for each gene under different script parameters.

Min. DP for including SNP	Min. qual. for including basecall	Min. # reads supporting haplotype	Min. % of reads supporting haplotype	Median length (bp)	Q90 length (bp)	# of genes
5	**10**	**5**	**0**	848.5	5026.3	3228
10	**10**	**5**	**0**	826.0	5026.8	3463
10	**10**	**5**	**5**	593.0	4783.2	3097
10	**10**	**5**	**10**	386.0	4501.4	2575
20	**10**	**5**	**0**	1032.0	5342.0	3271

Haplotyping potential in polyploid plants

Tetraploid potato



Hexaploid sweetpotato

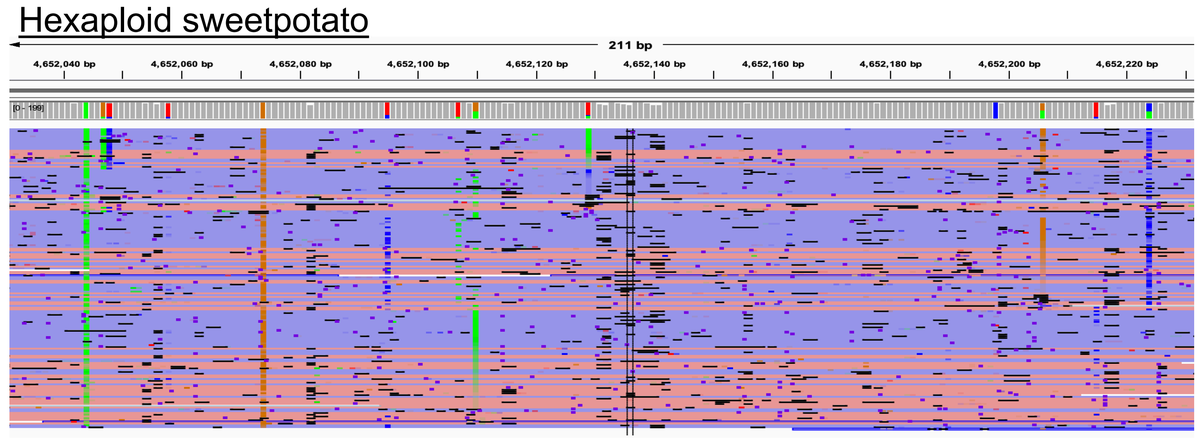


Fig 4. Haplotypes captured by Nanopore cDNA reads in polyploid plants. **(Top)** Between 3-4 haplotypes shown. **(Bottom)** Between 5-6 haplotypes shown.