Multiplexed phasing of clinically relevant long human PCR amplicons with short reads

Jessica M. Smith, Danny Yun, Li Wang, Joe Mellor, and Gavin Rush

seqWell, Inc. Beverly, MA USA

AGBT 2023

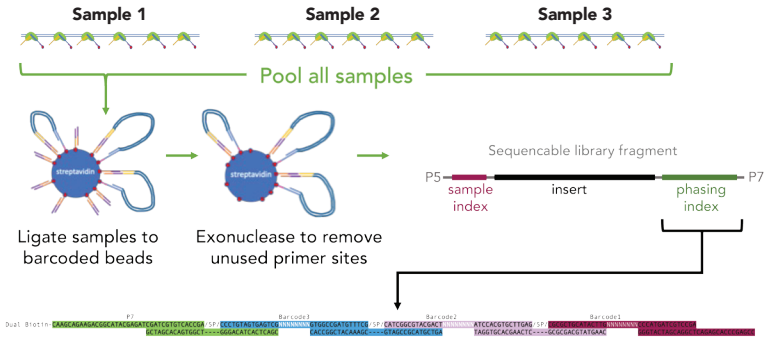
Introduction

While short-read sequencing technology can call variants with high fidelity, phasing variants over gene-length regions remains a challenge. Long-read sequencing methods are better suited for this application but have inadequate accuracy and are difficult to scale to many samples. We have developed a novel library preparation technology that links short reads across long DNA fragments and multiplexes many samples for fast and affordable parallel phasing of dozens of gene-length amplicons.

We used six kilobase amplicons of the gene Cyp21A2 as a demonstration of our technology because for best diagnostic accuracy, genotyping of Cyp21A2 must fully resolve the variants on both alleles. Variants in this gene are associated with congenital adrenal hyperplasia caused by 21-hydroxylase deficiency. Pathogenic variants in Cyp21A2 correspond to disease states ranging from mild to severe and affected individuals often have two different pathogenic alleles.

plexWell Long Read workflow

Figure 1: Samples undergo i5 indexed tagmentation in separate reactions, then are pooled and applied to barcoded beads to generate library fragments with phasing information



Phasing index is synthesized on the surface of streptavidin-coated beads with three eight base barcodes. Each barcode has 96 different indexes, yielding 884,736 different barcode combinations to identify long DNA fragments.

Cyp21A2 amplicons

- We generated ~6,000 base pair Cyp21A2 amplicons from two individuals (NA12217 and NA14733) with a gene fusion disrupting one allele of the gene, resulting in effectively haploid at this loci
- An equimolar combination of these two amplicons yields a sample with known phasing
- Positions are given relative to the length of the amplicon

Position	Ref	Alt	Genotype
337	C	T	1|0
357	C	A	1|0
617	C	T	1|0
641	A	C	1|0
675	C	T	1|0
878	C	A	0|1
906	G	A	1|0
1223	T	A	1|0
1810	C	G	1|0
2471	C	T	1|0
2916	G	A	1|0
2924	G	A	1|0
3371	C	T	1|0
3476	A	G	1|0
4292	T	A	1|0
4752	T	G	1|0
5224	G	C	1|0
5255	C	T	1|0
5268	C	T	1|0
5378	A	G	1|0

Table 1: Truth genotype phasing of mock diploid Cyp21A2 amplicons

Sequencing and phasing results

- We multiplexed 24 samples generated with four different input mass levels across an Illumina MiSeq flowcell, resulting in roughly 737,000 reads per sample resulting in a 98% barcode-aware duplication rate (99% normal duplication rate)
- We achieve high coverage across the amplicons while identifying only 5,000-25,000 unique phasing barcodes (5%-2.5% of the884,736 total barcodes)

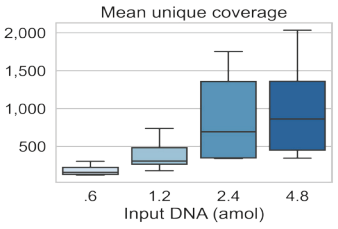


Fig 2: Amplicons are sequenced to high depth to capture phasing information

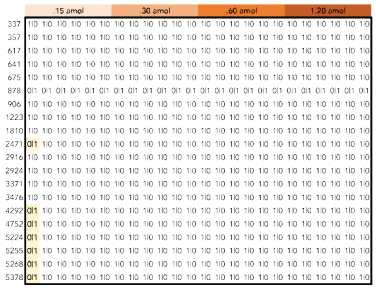


Table 2: Results showing 23/24 samples perfectly phased over a range of input DNA amounts. Phased genotypes highlighted in yellow are inaccurate calls.

- Phased data from these samples showed an average of 2.9 linked reads per phasing barcode with average fragment insert size of ~220 base pairs
- Heterozygous variants at 20 positions were phased correctly at all positions for > 95% of samples tested (23/24 samples perfectly phased)
- Phasing performance was unchanged when fastqs were downsampled to 100k reads/sample, suggesting that comparable performance is potentially attainable at higher multiplexing levels (e.g. 96 plex on a MiSeq).

Phasing analysis

After calling variants on aligned sequencing reads using open-source bioinformatics tools (Fig 3).

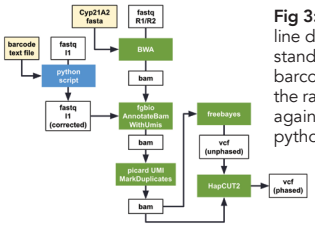


Fig 3: Bioinformatics pipeline diagram including mostly standard tools. The phasing barcodes are extracted from the raw i7 read and corrected against a whitelist via a python script.

Heterozygous variants were linked via the phasing barcode and haplotypes assembled using maximum likelihood estimation. As demonstrated in Fig 4, accurate phasing is robust to some phasing mismatch.

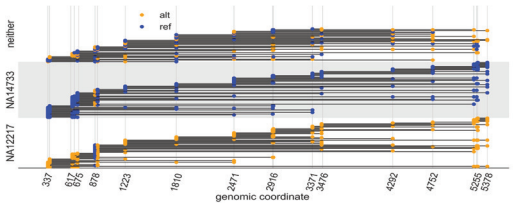


Fig 4: Identified linkages for heterozygous variants in one sample

Summary and Conclusions

- The pooled format of the phase barcoding process presents significant workflow advantages for processing large numbers of samples with an efficient and scalable workflow. We anticipate this technology will have significant potential to be extended to other applications including single cell sequencing, bacterial and viral metagenomics, and full-length transcript assembly.
- This product is still in development. Contact info@seqwell.com for more information.