SEQUENCING

Short reads join hands

A transposase can link sequence fragments together for accurate haplotyping and genome assembly.

Most modern DNA sequencers report sequences in short stretches, and they do so very efficiently. But many biological questions call for long-range sequence information, which can be difficult to reconstruct from short fragments. Researchers at Illumina led by Frank Steemers have devised a simple solution that applies a transposase—an enzyme used to prepare samples for sequencing—to temporarily bind short DNA fragments together and maintain sequence order, or contiguity (Amini *et al.*, 2014).

Contiguity is critical for distinguishing the two parental patterns of sequence variation, or haplotypes, that an individual inherits, a problem known as phasing. By specifying the sequence of each gene copy, phasing can resolve cases in which variation has an ambiguous effect on gene function. It can also match transplant donors and recipients and be used to understand structural variation in cancer genomes and beyond, says coauthor Kevin Gunderson.

Computational tools can cheaply and accurately phase common variants over small regions but are largely blind to new or rare variants. Parental DNA can fill in missing information but is often too difficult to obtain or too expensive to sequence.

Traditional experimental phasing approaches generate unique sequence-indexed libraries from parental copies that are separated by physical means, which can be laborious or require specialized equipment, or by dilution and division into aliquots. The tiny quantities of DNA can introduce biases and require extensive sequencing. Pooling aliquots allows more efficient and robust sequencing but increases the chance that both copies of a given region will share an index.

The Illumina team stumbled on their contiguity-preserving transposition (CPT-seq) approach while testing related strategies. Gunderson, who helped conceive the work, recalls trouble removing a Tn5 transposase that is routinely used to fragment DNA and add adaptor sequences in preparation for sequencing. "We observed this high–molecular weight band in the gel, and we said, 'Hey, wait a minute! If we can't remove it…'"

They tested their hunch and confirmed that the transposon physically links sequence fragments until it is removed chemically. In CPT-seq, 96-aliquot dilutions of large fragments are first indexed using the transposase and then scrambled and redivided into 96 compartments. The transposase is then removed, and amplification is used to introduce a new set of indexes, effectively creating over 9,200 virtual compartments.

The large number of compartments from double indexing means the presence of "100-fold higher concentrations in any one particular dilution," says Steemers. Moreover, the transposase adaptors relieve the need for random primer amplification, which "can give a lot of background noise, chimerism and G+C bias," he says.

The protocol takes only 3 hours and uses a small amount of genomic DNA up to chromosome length rather than large numbers of cells or size-selected DNA—benefits that may make experimental haplotyping routine. "You want a clinical phased genome, which we believe will be the standard," says Gunderson.

With CPT-seq and its freely available analysis software, over 95% of new variants can be phased with a low error rate of one or two incorrect phase switches every 10 megabase pairs or so. The researchers are adapting the protocol to sequence targeted regions and aiming to improve the phased fraction of the genome.

The study was a collaboration with Jay Shendure's lab at the University of Washington. In another study led by Shendure and involving both groups (Adey *et al.*, 2014), CPT-seq was combined with a new algorithmic approach to the problem of genome assembly, successfully adding contiguity information in a size range similar to that of more laborious fosmid-based sequencing.

CPT-seq provides a simple way to 'reach far' with short-read sequencing, easing clinical and research applications that demand long-range sequence information.

Tal Nawy

Adey, A. *et al. In vitro*, long-range sequence information for *de novo* genome assembly via transposase contiguity. *Genome Res.* doi:10.1101/gr.178319.114 (19 October 2014).

Amini, S. *et al.* Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* doi:10.1038/ng.3119 (19 October 2014).