Nature Reviews Genetics | AOP, published online 24 July 2012; doi:10.1038/nrg3308

TECHNOLOGY

When two halves are better than one whole

Advances in DNA-sequencing technologies have supplemented SNP microarrays to facilitate the genotyping of individuals and thus to provide a wealth of information on human variation. However, phasing the haplotypes (alleles that are co-located on the same chromosome) from these data usually requires laborious follow-up sequencing or indirect statistical methods. A new method can now efficiently and accurately achieve human genome sequencing and haplotyping from as few as 10–20 cells.

Peters *et al.* developed long-fragment read (LFR) sequencing. Briefly, this method involves diluting large fragments (10–1,000 kb) of picogram-quantity genomic DNA into 384-well plates, such that each well contains 10–20% of a haploid genome. Crucially, this means that for the genomic loci in each well, only one parental allele is typically present. Next, libraries for high-throughput sequencing are prepared by in-well enzymatic steps, including the ligation of well-identifying barcode sequences. These libraries are then pooled and sequenced. During data analysis, the barcode sequences allow bioinformatic deconvolution of the data from each well and thus the phasing of haplotypes based on the single allele present. Because each allele is represented in overlapping fragments in ~40 different wells, haplotypes can be phased genome-wide.

To test the accuracy of the method, the authors used LFR sequencing on the genomes of seven humans for which haplotype information is known from previous studies, such as the HapMap project. They achieved a ~92% phasing rate, which was comparable to other methods. This was increased to 97% by increasing the numbers of replicates or the initial quantity of DNA.

As only a small quantity of genomic DNA is required (that is, DNA from ~10–20 cells), LFR sequencing is likely to be applicable to a wide range of sample types. Although the authors noticed a high rate of errors in the raw sequence reads owing to the multiple rounds of DNA amplification that are required to generate sufficient quantities for sequencing, these were largely eliminated at the data-analysis stage because maternal and paternal alleles had been sequenced separately and multiple times, resulting in one false positive per 10 Mb.

As examples of the biological applications of this haplotype information, the authors assigned the parental origins of various *de novo* mutations, which cannot be inferred by statistical methods based on population history. Additionally, the impact of multiple variants in the coding region and/or *cis*-regulatory elements of a gene will depend on whether they are co-located. The authors found that each individual in the analysis had ~40 genes with predicted inactivating mutations in both alleles, adding to our growing appreciation of the pervasiveness of homozygously inactivated genes in human populations.

The accuracy, cost-effectiveness and biological insights make this an attractive approach for genomic research and potential clinical applications.

Darren J. Burgess

ORIGINAL RESEARCH PAPER Peters, B. A. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012)
FURTHER READING Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nature Rev. Genet.* **12**, 703–714 (2011).