The next phase in human genetics

Vikas Bansal, Ryan Tewhey, Eric J Topol & Nicholas J Schork

Experimental haplotyping of whole genomes is now feasible, enabling new studies aimed at linking sequence variation to human phenotypes and disease susceptibility.

The maternal and paternal copies of each chromosome in the human genome have distinct combinations of nucleotides that are functionally important, but knowledge of this 'haplotype' information (**Fig. 1a**) has been absent from all but a handful of studies of genomes[1–3]. The reason for this is largely technical: determining haplotypes, or 'phasing,' is not trivial[4,5]. Two papers in this issue report experimental methods for phasing at the genome scale. Fan *et al.*[6] physically separate the chromosomes in single cells using a microfluidic device, essentially phasing before genome analysis. Kitzman *et al.*[7] prepare standard mixtures of maternal and paternal chromosomes for whole-genome sequencing using a new protocol that enables phasing through bioinformatics analysis. Haplotyping strategies such as these should transform human genome sequencing and the study of the phenotypic effects of combinations of genome sequence variants.

Haplotype information is essential for human genetic research because of the fundamental importance of diploidy in human biology, as seen in phenomena such as haploinsufficiency, recessive acting variants, dosage compensation and parent-of-origin imprinting effects. A simple example that shows the need for phasing is compound heterozygosity, which occurs when an individual is a heterozygote carrier of two different mutations at different loci in the same gene. In such cases, a phenotype may arise only when the two mutations are present on different chromosomes, disrupting both copies of the encoded protein[3]. In a more complex example, the two mutations may give rise to a mutated protein from one chromosome and aberrant allele-specific expression or methylation from the other. Assessing the combined effect of mutations implicated in compound heterozygosity requires phase information; simply knowing that an individual is heterozygous at the two loci is not enough. Resolving phase is also important for addressing a range of other problems in human genetics, including characterizing the genomes of under-studied populations[7], comparing chromosomal segments shared between *homo sapiens* and distant ancestors, and detecting complex genomic structural variation.

Existing methods for phasing an individual's whole genome have involved either analysis of related individuals[3], which is often not possible, or labor-intensive one-by-one cloning and sequencing of many large fragments of the genome[1], which is not scalable. Other approaches, such as pedigree- and population-based statistical phasing algorithms, have been used in traditional linkage and linkage disequilibrium mapping. But they are not comprehensive, typically resolving haplotypes only for specific genomic regions and particular variations co-segregating with a phenotype. Moreover, because these other strategies use probabilistic haplotype information, they do not directly observe all of the nucleotide content of a haplotype and are not appropriate for rare variants, which may not have been observed in enough people to be statistically informative. Imputation methods based on phasing algorithms, which infer missing genotype information from other available data, also suffer from the same limitations. Finally, previous haplotyping methods based solely on DNA sequencing reads can be incomplete without additional information to anchor those reads to a chromosome[4,5].

The two papers in this issue are among the first potentially scalable and accurate methods for experimentally phasing entire human genomes. Kitzman *et al.*[7] describe a cost-effective strategy for assembling long haplotypes by sequencing many haploid subsets of an individual genome using next-generation sequencing platforms (**Fig. 1b**). The first step in this approach is to generate a single whole-genome fosmid library with long inserts, in this case ~37 kb. This library is then randomly partitioned into pools such that each pool is essentially a haploid mixture of clones derived from either the maternal or paternal DNA at each genomic location. High-throughput sequencing of each pool provides haplotype information for each clone in that pool. Overlaps between haplotypes derived from different pools are then pieced together to assemble even longer haplotypes. Notably, the method of Kitzman *et al.*[7] is similar to the approach taken to sequence the genome of Craig Venter[1], which used fosmid libraries and standard Sanger sequencing to obtain long-range haplotype information as well as computational algorithms to assemble long haplotypes. The approach of Kitzman *et al.*[7], though, is more scalable and amenable to short-read sequencing.

Kitzman *et al.*[7] apply their method to study the genome of a female with ancestry from western India. They assemble haplotype contigs, or blocks, with a length of ≥386 kb for about half of the genome. Using the resulting phase information, they identify 10 genes (from a candidate set of 44 genes) that harbor two or more rare heterozygous functional mutations that are on different homologous copies of the gene and might therefore cause compound heterozygous phenotypic effects. The authors also use the phase information to identify haplotypic segments that are enriched for novel variants and differ substantially from previously sequenced HapMap populations, suggesting that haplotypes, in contrast to the genotype information captured in the HapMap initiative, contain much more information about ancestry. Lastly, the authors are able to use the fosmid data to detect complex structural variants—a difficult task when both homologous chromosomes are sequenced together.

Vikas Bansal, Eric J. Topol & Nicholas J. Schork are at Scripps Health, La Jolla, California, USA; Vikas Bansal, Ryan Tewhey, Eric J. Topol & Nicholas J. Schork are at The Scripps Translational Science Institute, La Jolla, California, USA; Eric J. Topol & Nicholas J. Schork are in The Department of Experimental Medicine, The Scripps Research Institute, La Jolla, California, USA.
e-mail: nschork@scripps.edu

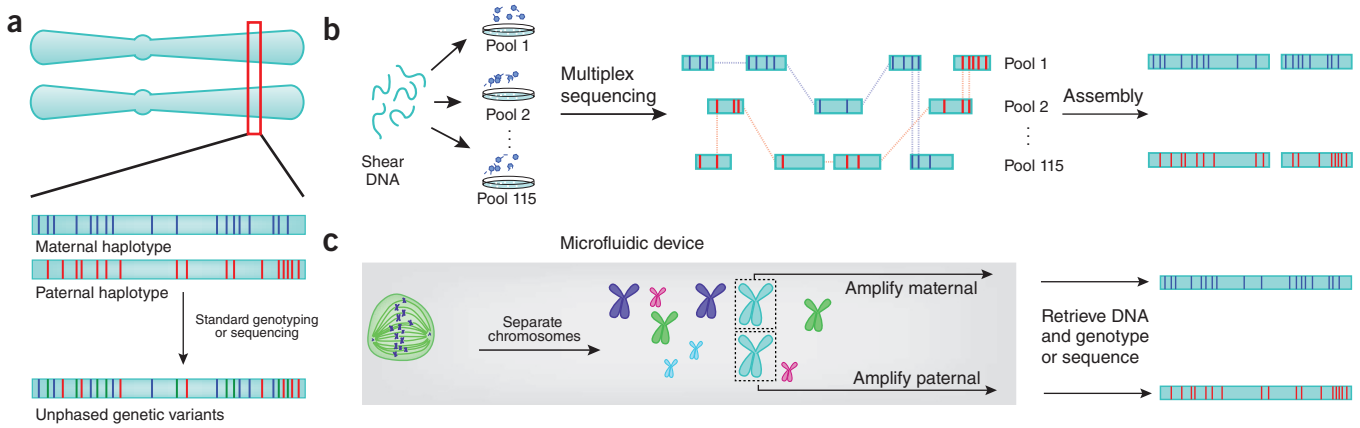

Figure 1 Experimental approaches to haplotyping genomes. (**a**) In standard analysis of a mixed pool of maternally and paternally inherited chromosomal DNA, haplotype information is lost. Blue vertical lines represent sequence variants on the maternal chromosome homolog; red vertical lines represent variants on the paternal homolog; green lines represent homozygous variants. (**b**) Kitzman *et al.*[7] exploit large-insert fosmid clone libraries that allow sequencing reads derived from a single chromosomal homolog to be associated with each other. Cloning and sequencing is multiplexed, enabling efficient construction of contigs that span large genomic regions. Although this approach may be easier to implement in most sequencing laboratories, it is less robust than the approach of Fan *et al.*[6] owing to the need for assembly. (**c**) Fan *et al.*[6] use a microfluidic device to separate and amplify homologous chromosomes during metaphase in single cells, enabling the individual chromosomes to be sequenced and nearly perfectly phased.

One drawback of the approach of Kitzman *et al.*[7] is that it requires stitching together phased contigs, albeit rather large ones. This may result in switching errors, where chromosomal segments are accurately haplotyped but misrepresent complete chromosomes. Such errors can occur once or many times over different chromosomes. These concerns are eliminated in the approach developed by Fan *et al.*[6], which resolves phase directly by isolating individual copies of each chromosome in a single cell with a microfluidic device. Thus, there is no need to assemble contigs. After isolation of single chromosomes, a haploid mixture of clones derived from either the maternal or paternal DNA can be genotyped using standard single-nucleotide polymorphism arrays or shotgun sequenced to generate haplotypes spanning entire chromosomes (**Fig. 1c**).

Fan *et al.*[6] validate their approach by haplotyping the genomes of a mother-father-child trio from the HapMap project. The experimentally determined haplotypes are highly concordant (99.8%) with previously calculated haplotypes inferred using family and population information, demonstrating the accuracy of the approach. The authors also demonstrate the potential of their method for clinical diagnostics by haplotyping a fourth individual, P0, whose genome has already been sequenced, at the highly polymorphic HLA locus. Phase information for this genomic region is very important for matching transplanted donor organs with a potential host. Although the approach taken by Fan *et al.*[6] is direct and the most optimal for phasing, the need for sophisticated microfluidic devices and sequencing technologies able to handle individual chromosomes may delay its routine use.

The work of Fan *et al.*[6] and Kitzman *et al.*[7] highlight the obvious, yet often overlooked, diploid nature of the human genome and expose the incompleteness of available individual genome sequences that do not phase genetic variants. Going forward, discussions of individual human genomes should refer either to a single maternally or paternally derived haplotypic complement of DNA or to the two genomes that each person possesses. This simple change in language should help emphasize the importance of studies that account for the phase information that is the hallmark of the human genome.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Levy, S. *et al. PLoS Biol.* **5**, e254 (2007).
2. Wang, J. *et al. Nature* **456**, 60–65 (2008).
3. Roach, J.C. *et al. Science* **328**, 636–639 (2010).
4. Bansal, V., Halpern, A.L., Axelrod, N. & Bafna, V. *Genome Res.* **18**, 1336–1346 (2008).
5. He, D., Choi, A., Pipatsrisawat, K., Darwiche, A. & Eskin, E. *Bioinformatics* **26**, i183–i190 (2010).
6. Fan, H.C., Wang, H., Potanina, A. & Quake, S.R. *Nat. Biotechnol.* **29**, 51–57 (2011).
7. Kitzman, J.O. *et al. Nat. Biotechnol.* **29**, 59–63 (2011).

Crafting rat genomes with zinc fingers

Meng Amy Li & Allan Bradley

Expressing zinc-finger nucleases in zygotes enables targeted transgene integration in the mouse and rat genomes.

Mammalian oocytes and zygotes are extraordinarily resilient receptacles that provide a conduit from designs etched in laboratory notebooks to living animals. The first transgenic mammals were generated three decades ago by injection of naked DNA into the pronuclei of mouse zygotes. Until now, however, pronuclear injection has allowed insertion of exogenous DNA only at random sites in the genome, and site-specific engineering has proved extremely difficult. In this issue, Cui and colleagues[1] have finally overcome this barrier, making use of zinc-finger nucleases (ZFNs) to stimulate targeted integration of transgenes by homologous recombination in mouse and rat zygotes. This technology will dramatically alter the speed

Meng Amy Li & Allan Bradley are at Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom
e-mail: abradley@sanger.ac.uk