nature genetics

Whole genome?

The reference human genome assembly is remarkable in its completeness and usefulness in research. However, the range of allelic variation in the human population is not well described by a haploid assembly with a profusion of alternative loci. Homozygous regions and the use of multiple sequencing technologies increasingly have roles in strategies for identifying regulatory and trait-associated variation.

It is time to stop thinking that merely more DNA sequencing will give us the variants that determine human traits: we need to explore more diverse strategies and we need to put a lot more effort into the regions of the genome that are subject to more complex modes of change over time and are therefore more difficult to relate to our biology.

The idea of a reference human genome is a good one; at a first approximation, we all have much the same genome, and we now know that one of the contributors to this fact is that most of the genome has some contribution to making us human, with very little, if any, 'junk' DNA. But, still, we differ in phenotypes, traits and disease predispositions, and our genomes come in many flavors of combinations of rare and common variation on many different spatial scales.

The ways in which nucleotide and structural variation affect protein-coding genes are relatively straightforward to describe. Documenting copy number variation is somewhere in the middle of the difficulty spectrum, with most of the difficulty inherent in developing reproducible methods for converting sequence or genotype data into counts of gene number, but once this is achieved trait association can be feasible, as demonstrated in the example of association of amylase gene copy number with obesity (*Nat. Genet.* **46**, 492–497, 2014). In contrast, variants acting at a distance to regulate transcriptional units and chromatin domains with coordinate expression patterns have a much larger range of possible mechanisms by which they can alter gene function. Thus far, the search for regulatory variants has focused largely on well-defined, simple duplications and deletions and (mostly) diallelic single-nucleotide polymorphisms (SNPs), in the hope that these will have measurable effects on well-defined targets of DNA-binding regulatory proteins.

Much of what remains to be understood in the genome is intrinsically difficult to sequence because of base-pair composition or simple sequence repeats, difficult to assemble because of repeat lengths exceeding read lengths in current short-read sequencing technology and difficult to interpret because the sequences reflect their history, with multiple iterations of replication and division having given rise to complex repeats. Some of the most complex compound repeats are reminiscent of the disrupted runs of shuffled playing cards from multiple decks or the compound folded layers of butter and dough in Danish pastry.

One very promising strategy has been to generate effectively haploid reference genome sequences by long-read and single-molecule sequencing technology (*Genome Res.* **24**, 2066–2076, 2014, and *Nature* **517**, 608–611, 2015). This technology, currently relatively slow and expensive, is not restricted to whole-genome sequencing but can also be useful for sequencing selected genomic regions. We encourage the use of a range of sequencing technologies to explore highly variable and complex genomic regions in a large number of human samples. SNP genotyping can be used to identify informative individuals homozygous for haplotypes of interest that can then be examined in detail via long-read sequencing.

The most highly variable regions of the human genome may be better described by multiple, chromosome-sized haplotypes from real genomes than by a single, idealized haploid reference genome with alternative loci of different sizes. Of course, working with multiple reference sequences requires a different analytical approach and new software to be useful, but this strategy will let us unlock more of the genomic basis of heritable human traits. ■