nature genetics

Little boxes

Our ability to map trait-associated regulatory variation still vastly exceeds the prospects for dissecting allele-specific effects on gene expression and activity *in vivo* in relevant tissues and organs. A small number of intensive investigations into functional variants should pave the way for scaleable strategies using high-throughput techniques and genomic data integration.

On page 714 of this issue, Jie Luo and colleagues report the investigation of 323 quantitative metabolite traits in rice and 551 SNPs at 356 loci for which genome-wide significant associations could be replicated in at least 2 populations. Five of these associations were followed up as proof of principle, as there was a readily interpretable coding variant to be investigated by transgenic overexpression. In the case of these 5 of 36 readily interpretable candidate genes, gene dose affected the levels of a relevant metabolite.

However, the main difficulty is interpreting, not finding, variants, as coding SNPs may comprise just 4.7% of the lead SNPs initially identified by genome-wide association study (GWAS). A larger proportion, some 44.8% (*Genome Res.* **22**, 1748–1759, 2012), that overlap with noncoding DNA elements identified by the Encyclopedia of DNA Elements (ENCODE) Project Consortium are candidates for regulatory variants. If SNPs with a high degree of linkage disequilibrium (LD) with lead SNPs are included, some 31% of GWAS SNPs sit in annotated transcription factor binding sites and 71% overlap with a DNase I–hypersensitive site (*Nature* **489**, 57–74, 2012). This is a considerable enrichment, as it is estimated that the recognition sites for gene-regulatory DNA-binding proteins ('little boxes') occupy a portion of the genome approximately twice as large as the exome (*Nature* **489**, 83–90, 2012).

Forging links in the chain of evidence from trait association to regulatory function is far from trivial and involves distinguishing regulatory variants from other correlated SNPs, then identifying the gene or genes on which a regulatory element operates (often at a considerable distance), establishing the tissue of gene expression relevant to the trait and, finally, determining allele-specific levels of gene expression.

When there are few SNPs in a region of LD and where regulatory elements are conserved in model genomes, causal evidence can be established in a bespoke manner, locus by locus. On page 753, François Spitz and colleagues followed up an association at 8q24 with human cleft lip and/or palate with *in vivo* reporter constructs and engineered chromosome rearrangements in mice to tie enhancer elements to *Myc* regulation in craniofacial development. On page 748, David Kingsley and colleagues identified a SNP for blond hair color in humans within a non-consensus transcription factor binding site far from the mouse *Kitl* gene (see also page 660). Elegant though these experiments are, they required extraordinary effort, and the methods cannot readily be applied to every SNP association in turn (because a deletion or inversion of a regulatory element might not have the same effect on expression as a base substitution and because a transgene in the context of a wild-type mouse genotype does not usually recapitulate the genetic architecture relevant to the human phenotype).

Judging by the high-throughput enhancer profiling of Alexander Stark and colleagues (page 685), it might seem surprising that enhancer elements are conserved at all, as apparently neutrally evolving regions of the genome give birth to thousands of enhancer elements over relatively short evolutionary timescales. This mode of enhancer evolution may have implications for the spectrum of rare and common variants we find in human populations as well as for practical aspects of modeling the regulation of gene expression in mice and other related species. As for matching enhancers to genes, systematic methods are being developed to capture hundreds of promoter regions for genes and identify their looping to distant *cis*-acting regions (for example, see *Nat. Genet.* **46**, 205–212, 2014).

Having encouraged rigorous standards in genomics and genetic epidemiology and discussion of post-GWAS strategies, we are excited to see new methods scale up to meet the challenge of dissecting regulatory genome variation. ■