Nature Reviews Genetics | AOP, published online 12 April 2012; doi:10.1038/nrg3219

FUNCTIONAL GENOMICS

Complexities of occupancy and sequence

> single-nucleotide variants at the same binding-site position in different genomic binding-site locations have differing effects on protein occupancy

Variation in non-coding regions of the genome is increasingly being implicated in inter-individual variation in complex traits, including disease susceptibility, but interpreting the functional effects of non-coding variation is particularly challenging. Two recent papers that have systematically studied the effects of SNPs on transcription factor binding show that although some trends in the relationship between sequence and binding are as expected, predicting the effects of specific SNPs will be difficult.

Maurano *et al.* mapped binding sites for the transcriptional regulator CTCF by chromatin immunoprecipitation followed by sequencing (ChIP–seq) in lymphoblastoid cell lines that were derived from 12 members of a family spanning 3 generations. They then carried out targeted resequencing of the 134 bp interval surrounding each binding site, so altogether they had high-resolution genotype and ChIP data for a total of >35,000 CTCF binding sites. Of these sites, 21% overlapped at least one SNP, allowing them to explore the relationship between SNPs and site occupancy. Overall, 5.6% of the polymorphic binding sites had a significant association of SNP genotype with CTCF occupancy and, as expected, 85% of the SNPs that affected occupancy lay within the 44 bp region where CTCF contacts the DNA at its binding sites.

However, it should be noted that most SNPs in the protein–DNA interface region do not affect occupancy, and even in the core 14 bp CTCF binding motif, only 36% of SNPs affected occupancy. Furthermore, single-nucleotide variants at the same binding-site position in different genomic binding-site locations have differing effects on protein occupancy, depending on their context. These findings indicate a buffering of the effects of SNPs. The extent of buffering seems to be dependent on binding-site strength (with stronger motifs being buffered against all but very disruptive changes) and sequence context. For example, the SNPs at position 1 in the core CTCF motif that had an effect on occupancy were all in the context of an adenine at position 5.

In the second study, Reddy *et al.* used ChIP–seq and resequencing data from a lymphoblastoid cell line from one individual (for whom the parental genome sequences were also available), and they looked at a panel of 24 transcription factors. The patterns observed are similar to those for CTCF: 13% of transcription-factor-occupied regions were polymorphic, and 5.5% of the heterozygous polymorphic sites showed allelic differences in transcription factor occupancy. However, variants in known transcription factor binding motifs only accounted for ~12% of the cases of differential allelic occupancy. These authors also analysed the occupancy data alongside allelic gene expression data and found that occupancy within 100 bp of a transcription start site is highly predictive of expression; some associations of occupancy with expression were found for more distant sites, but the long-range effects are weaker and more difficult to predict.

Together, these studies suggest that functional studies will be needed alongside informatic predictions in order to understand the functions of non-coding SNPs.

Mary Muers

ORIGINAL RESEARCH PAPERS Maurano, M. T. *et al.* Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.* **8**, e1002599 (2012) | Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* 2 Feb 2012 (doi:10.1101/gr.131201.111)
FURTHER READING Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nature Rev. Genet.* **12**, 628–640 (2011)

BRAND X