Nature Reviews Genetics | Published online 18 Apr 2016; doi:10.1038/nrg.2016.53

GENE REGULATION

Finding genetic target sites

EyeWire/Getty Images

There is increasing realization that genetic variation in non-coding regulatory elements, such as enhancers, has a major role in evolution, complex traits and diseases. However, a key challenge is to identify the genes regulated by these elements in order to dissect the gene regulatory networks through which they act. A new study reports a machine-learning approach that leverages multilayered epigenomics profiles to predict which promoters are targeted by particular enhancers.

Various approaches have been devised to predict the target genes of regulatory elements. These strategies have largely relied on simple but intuitive criteria, such as prioritizing genes based on genomic proximity to the regulatory element and/or based on quantitative trait locus analyses (that is, genes for which expression levels correlate with genetic variation in the element). Whalen *et al.* sought to determine whether target gene predictions could be improved by taking advantage of the rich information within existing epigenomic profiling data sets; such profiles include DNA methylation status, numerous histone modifications, chromatin accessibility, binding locations of diverse transcription factors and chromosome architecture proteins, and gene expression data.

The team focused on six human cell lines of different tissue types for which extensive and diverse molecular data sets are available. They identified putative active enhancers and promoters based on characteristic histone modifications and open chromatin, and then classified enhancer–promoter pairs that were less than 2 Mb apart into target versus non-target pairs using evidence of enhancer–promoter physical contact from chromatin conformation capture data. They then built their 'TargetFinder' machine-learning algorithm to dissect which of the molecular profiling features could predict the enhancer–promoter interactions.

Overall, the researchers found that no single feature provided strong predictive power, but that the algorithm could identify an optimal combination of features that was highly informative and achieved a false discovery rate (FDR) of only 8–15%. By comparison, a standard alternative method based on the nearest actively transcribed gene had an FDR of 53–77%.

Within the optimal combination of features, those that contributed most to the predictive power across cell lines included DNA methylation, histone marks associated with transcription elongation, and binding sites of repressive or architectural proteins. Consistent with chromosomal looping being required to bring enhancers into proximity with their target promoters, regulatory contacts were associated with the binding of looping factors (such as CTCF, the cohesin complex and some zinc-finger proteins) near the interacting elements. Furthermore, features in the intervening regions between an enhancer and a promoter were predictive of contacts and provided insight into the molecular mechanisms of looping, such as putative combinatorial interactions between proteins, and the underappreciated or cell-type-specific roles of various transcription factors.

The study also pointed to less commonly profiled features that might have future value for predicting target sites. CTCF is known to undergo post-translational sumoylation, and supplementing the input data with sumoylation profiles enhanced the predictive power of the algorithm.

It will be interesting to determine the contribution that TargetFinder will make to our ability to uncover the molecular consequences of disease-associated regulatory variation, and to see whether profiling data from additional cell types will facilitate insights into diverse disease-affected tissues.

Darren J. Burgess

ORIGINAL ARTICLE Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* http://dx.doi.org/10.1038/ng.3539 (2016)