Published in final edited form as:
Trends Genet. 2016 July ; 32(7): 392–394. doi:10.1016/j.tig.2016.05.002.

Fishing for function in the human gene pool

Iros Barozzi[1], **Axel Visel**[1,2,3,*], and **Diane E. Dickel**[1,*]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[2]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA

[3]School of Natural Sciences, University of California, Merced, Merced, California, USA

Abstract

Identification and characterization of causal non-coding variants in human genomes is challenging and requires substantial experimental resources. A new study by Tehranchi *et al.* describes a cost-effective approach for accurate mapping of molecular quantitative trait loci (QTLs) from pooled samples, a powerful way to link disease-associated changes to molecular functions.

Keywords

non-coding DNA; *cis*-regulation; QTL; GWAS

Genome-wide association studies (GWAS) have implicated thousands of genomic regions in the susceptibility to a wide spectrum of diseases and common traits [1]. However, because the majority of phenotype-linked variants fall into non-coding loci [1,2], little is known about the molecular mechanisms underlying most such disease associations. While the effects of variants that fall in protein-coding genes can be modeled computationally, albeit imperfectly, our lack of understanding of the non-coding genetic code has thus far hampered the development of equivalent prediction methods for use on the >95% of the genome that does not encode proteins. Despite these challenges, the increasing availability of individual human genomes and the technological advancements to probe gene expression and chromatin state genome-wide have begun to enable progress towards the mechanistic exploration of human non-coding sequence variants. For example, more than one third of disease-associated SNPs reside in loci with the potential to act as *cis*-regulatory elements [3], in particular enhancers. This figure is likely to grow along with the number of elements being mapped across an increasing collection of relevant cell types and primary tissues. Nevertheless, predicting the effect of a sequence variant on the activity of the respective regulatory element has remained difficult to address. Linking non-coding variants to a molecular phenotype, such as an alteration in gene expression or the binding of a

[*]To whom correspondence should be addressed: A.V., avisel@lbl.gov; D.E.D., dedickel@lbl.gov.

Disclaimer Statement

The authors declare no conflicts of interests.

transcription factor, is an increasingly feasible way to gain a foothold into the underlying mechanism of how non-coding sequence variants might increase the risk for a disease. A recent study from Tehranchi *et al.* [4] makes significant improvements to methods aimed at identifying sequence alterations affecting transcription factor binding and shines new light on the functional consequences of regulatory sequence variation.

Previously, profiling of cell lines derived from single individuals has been successfully employed to link non-coding genetic variation to molecular traits such as gene expression, transcription factor binding, or chromatin state [5,6]. For the latter two molecular traits, chromatin immuno-precipitation (ChIP) targeting transcription factors (TF) or histone modifications of interest has been performed separately on the chromatin of each individual cell line from a collection of subjects. Mapping of the ChIP-derived reads to the different alleles (Figure 1, center-left panel) can lead to the identification of quantitative trait loci potentially impinging the *cis*-regulatory activity of a genomic region (cQTLs), often referred to as bQTLs (when affecting binding of a TF) or hQTLs (when affecting histone modifications). These approaches require a large number of samples and considerable experimental and sequencing effort to identify QTLs with a high degree of statistical confidence.

Even though recent breakthroughs in genomics technologies have dramatically decreased the cost of high-throughput sequencing, the study by Tehranchi *et al.* introduces an even more cost-effective approach to cQTL mapping. This novel approach consists of pooling the chromatin from cell lines of several different individuals, then performing ChIP directly on the pool (Figure 1, center-right panel). Instead of inferring the relationship between the genotype and the binding separately for each individual, the frequency of each allele in the input material (pre-ChIP) is compared to its frequency in the enriched material (post-ChIP). This way, those variants showing an effect on the binding can be easily identified. This pooling strategy, which the authors have also recently applied to DNA methylation profiling [7], cuts down the expense substantially, resulting in costs up to 25-fold lower than the individual sample profiling of traditional studies. The pooled design also mitigates the sample-to-sample experimental variation of ChIP-seq, another considerable challenge to previous studies.

The proposed approach allowed the authors to map cQTLs for five transcription factors and the H3K4me3 histone modification in pools of human lymphoblastoid cell lines (LCLs) from 60 and 71 individuals, respectively. The resulting cQTLs were found to be in high agreement (80–99%) with previous studies using traditional approaches. Additionally, and in line with recent reports [8,9], the authors showed that bQTLs could also exert long-range effects on *cis*-regulation. Importantly, these results provide further evidence supporting TF binding as the major driver of chromatin variation [5,6]. The authors next used their results to address the major challenge of interpreting disease-associated variants identified by GWAS: the difficulty of disentangling the causative variant when many are found to be in strong linkage disequilibrium (LD). The authors reported that >3,500 of the identified bQTLs were previously associated with a phenotype (either directly or indirectly via LD), and they demonstrate how this approach could be used to pinpoint candidate causal variants within large LD blocks (Figure 1, lower panel).

Nevertheless, this strategy and all previous approaches to QTL mapping for molecular traits are still limited by three factors: 1) the number of individuals considered, where small numbers restrict the identification of low-frequency alleles; 2) the cost of sequencing at high depth, which impinges the identification of alleles with a small effect; and 3) the availability of high-quality antibodies to target relevant TFs. Larger sample sizes and the cost of sequencing are going to be increasingly negligible factors with time. The inability of generating ChIP-grade antibodies for every DNA-binding protein will be harder to mitigate. One solution will be to use the binding of a validated partner TF as readout, and to couple this information with a computational prediction about the binding of the relevant TF.

Another potential limitation is the current inability to perform these studies in models more closely relevant to human development and disease. Albeit restricted to LCLs as a model and to only five TFs, the authors highlight two SNPs with the potential to be relevant to cell types other than LCLs (myocardial infarction and cancer). Nevertheless, moving to appropriate cell types and considering the corresponding relevant TFs is of primary importance in order to investigate the effect of each variant in the proper context [10]. An attractive solution to this may be the increased use of differentiated induced pluripotent stem cells from human patients.

Despite the current limitations, the approach proposed by Tehranchi *et al.* represents a significant leap towards a higher-throughput, more systematic mapping of regulatory variation. We believe this will find broad application in human genetics as well as in regulatory genomics, and it will greatly increase our knowledge on the role of non-coding variation in development and differentiation, both in physiological and pathological conditions.

Acknowledgments

The authors are supported by National Institutes of Health grants R24HL123879, U01DE024427, R01HG003988, U54HG006997, and UM1HL098166. The authors conduct research at the E.O. Lawrence Berkeley National Laboratory under Department of Energy Contract DE-AC02-05CH11231, University of California.

References

1. Welter D, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014; 42:D1001–6. [PubMed: 24316577]
2. Abecasis GR, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]
3. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
4. Tehranchi AK, et al. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell. 2016; 165:730–741. [PubMed: 27087447]
5. McVicker G, et al. Identification of genetic variants that affect histone modifications in human cells. Science. 2013; 342:747–9. [PubMed: 24136359]
6. Kilpinen H, et al. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science. 2013; 342:744–7. [PubMed: 24136355]
7. Kaplow IM, et al. A pooling-based approach to mapping genetic variants associated with DNA methylation. Genome Res. 2015; 25:907–17. [PubMed: 25910490]
8. Grubert F, et al. Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. Cell. 2015; 162:1051–65. [PubMed: 26300125]

9. Waszak SM, et al. Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. Cell. 2015; 162:1039–50. [PubMed: 26300124]

10. Trynka G, et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet. 2013; 45:124–30. [PubMed: 23263488]

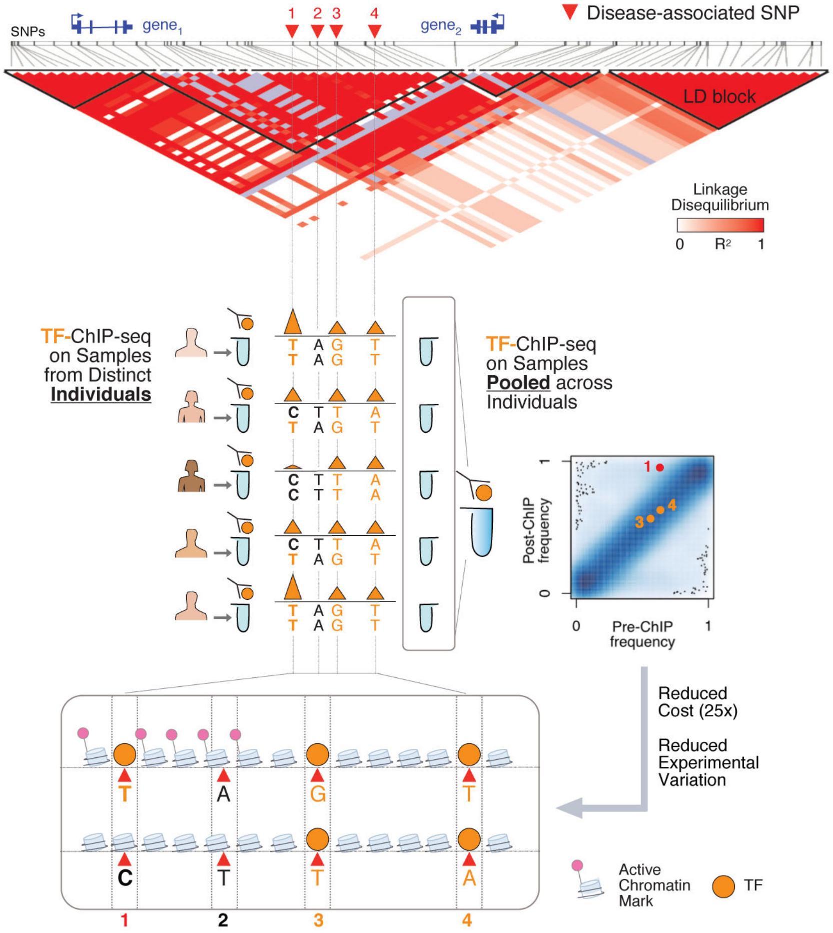

Figure 1. Mapping bQTLs: comparing a traditional approach to a pooling-based strategy

The figure shows a hypothetical scenario of using bQTL mapping to identify functional SNPs at a disease-associated locus. Traditional bQTL mapping makes use of ChIP-seq performed on different individuals to infer allele-specific binding separately for each sample (middle panel, left side). In contrast, in the pooling-based strategy introduced by Tehranchi *et al.* ChIP is performed on a pool of samples (middle panel, right side). The frequency of each allele in the input material (pre-ChIP) is then compared to its frequency in the enriched material (post-ChIP). Under this scenario, the traditional mapping shows that SNP #1 is bound by the transcription factor (TF) in individuals with genotype TT, while in individuals with genotype CC the site is barely bound; consistently, heterozygotes display intermediate levels of binding. The same result is obtained with the pooling-based strategy (see SNP #1 in the scatterplot), albeit at a much lower cost and reduced experimental variation. The lower panel shows the potential results. Out of the four linked variants previously associated with disease susceptibility, only #1 strongly affects the binding of the TF in the cell.