Genomics 104 (2014) 105–112

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Strategies to fine-map genetic associations with lipid levels by combining epigenomic annotations and liver-specific transcription profiles

Ken Sin Lo [a], Swarooparani Vadlamudi [b], Marie P. Fogarty [b], Karen L. Mohlke [b], Guillaume Lettre [a,c,*]

[a] Montreal Heart Institute, Montreal, Quebec, Canada
[b] Department of Genetics, University of North Carolina, Chapel Hill, NC, USA
[c] Université de Montréal, Montreal, Quebec, Canada

ARTICLE INFO

Article history:
Received 7 November 2013
Accepted 25 April 2014
Available online 2 July 2014

Keywords:
FOXA1
FOXA2
HNF4A
BIRC5
ENCODE
GWAS

ABSTRACT

Characterization of the epigenome promises to yield the functional elements buried in the human genome sequence, thus helping to annotate non-coding DNA polymorphisms with regulatory functions. Here, we develop two novel strategies to combine epigenomic data with transcriptomic profiles in humans or mice to prioritize potential candidate SNPs associated with lipid levels by genome-wide association study (GWAS). First, after confirming that lipid-associated loci that are also expression quantitative trait loci (eQTL) in human livers are enriched for ENCODE regulatory marks in the human hepatocellular HepG2 cell line, we prioritize candidate SNPs based on the number of these marks that overlap the variant position. This method recognized the known *SORT1* rs12740374 regulatory SNP associated with LDL-cholesterol, and highlighted candidate functional SNPs at 15 additional lipid loci. In the second strategy, we combine ENCODE chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) data and liver expression datasets from knockout mice lacking specific transcription factors. This approach identified SNPs in specific transcription factor binding sites that are located near target genes of these transcription factors. We show that FOXA2 transcription factor binding sites are enriched at lipid-associated loci and experimentally validate that alleles of one such proxy SNP located near the *FOXA2* target gene *BIRC5* show allelic differences in FOXA2-DNA binding and enhancer activity. These methods can be used to generate testable hypotheses for many non-coding SNPs associated with complex diseases or traits.

1. Introduction

Genome-wide association studies (GWAS) have identified thousands of robust associations between single nucleotide polymorphisms (SNPs) and complex human diseases and traits [13]. These SNPs are often in linkage disequilibrium (LD) with many other known and unknown DNA sequence variants and are located within non-coding regions of the human genome. For these two reasons, at most GWAS loci it has been difficult to identify the genes and variants that are responsible for phenotypic variation. The 1000 Genomes Project has generated an extensive catalogue of genetic variation across several human populations, partly addressing the first challenge in GWAS fine-mapping projects [1,2]. As for the second challenge, investigators from the Encyclopedia of DNA Elements (ENCODE) Project recently summarized results from comprehensive whole-genome analyses of transcription, transcription factor association, chromatin structure, and histone modification, allowing for a functional annotation of non-coding DNA variants [9]. Furthermore, the ENCODE data might be useful to pinpoint functional regulatory variants from strongly correlated, but not functional, LD proxies. Many groups have already utilized their own epigenomic datasets or ENCODE data to show enrichment of chromatin marks at GWAS loci, to identify relevant tissues for experimental design or to prioritize candidate functional genes and DNA sequence variants [8–10,16–18,23,26,29,32].

Additional work is needed to refine these existing methods. We also need to develop new tools when there is no evidence in human tissues that the associated non-coding SNPs control gene expression, that is when the SNPs are not expression quantitative trait loci (eQTLs). In an effort to broaden the application of this approach by the community, we further extended the use of epigenomic data to prioritize functional candidate SNPs by developing two novel approaches, and we applied these approaches to 95 loci associated with lipid levels in humans [28]. We were particularly interested in testing if gene expression datasets from relevant knockout mouse models could help prioritize candidate functional genes and variants at GWAS loci. Such a strategy could have broad implications as it may offer an alternative when there is no eQTL evidence or the human tissues are not readily accessible for transcriptomic studies. Our results demonstrate that combining human genetic, epigenomic and mouse expression data can provide additional fine-mapping resolution at GWAS loci. As a proof-of-principle,

[*] Corresponding author at: Montreal Heart Institute, 5000 Rue Bélanger, Montréal, Québec H1T 1C8, Canada. Fax: +1 514 593 2539.
E-mail address: guillaume.lettre@umontreal.ca (G. Lettre).

http://dx.doi.org/10.1016/j.ygeno.2014.04.006
0888-7543/© 2014 Elsevier Inc. All rights reserved.

we functionally tested and validated a variant in LD with a lipid sentinel SNP that interferes with the binding of the FOXA transcription factors and is located near a FOXA2 transcriptional target gene as determined by the transcriptomic characterization of *Foxa2*$^{-/-}$ mouse livers. Our two methods, applied individually or together, should be broadly applicable to other human complex traits and diseases.

2. Results

2.1. Enrichment analysis

For this study, we obtained from the ENCODE Project all DNAseI hypersensitive sites (DHS) and ChIP-seq peaks from HepG2, which are hepatoblastoma cells that have been extensively used to study lipid metabolism. For comparison, we also analyzed the same data in the three tier 1 ENCODE cell lines: B-lymphoblastoid cells GM12878, erythroleukemia cells K562 and human embryonic stem cells H1-hESC. In this article, we use the term "epigenomic annotation" to refer to any DHS or ChIP-seq peak reported by the ENCODE Project in these four cell lines. To quantify the overlap between ENCODE epigenomic annotations that mark regulatory DNA sequences and individual SNPs at GWAS loci, we counted epigenomic annotations in each cell line that overlap the SNP and assessed significance using a simple enrichment analysis framework. We considered variants in LD ($r^2 \geq 0.8$, European-ancestry individuals from the 1000 Genomes Project) with the GWAS sentinel SNPs and then used 5000 matched sets of markers to assess the statistical significance of the enrichment (see Methods section and Supplementary Fig. 1).

Applying this approach to 95 lipid loci, we found enrichment of DHS and most histone marks associated with transcription regulation. The enrichment was stronger in HepG2 cells than in the three other cell lines analyzed: 70% of marks (7 of 10) had enrichment $P < 0.0002$ for HepG2, whereas the corresponding proportions for GM12878, K562 and H1-hESC were 20%, 50% and 20%, respectively (Supplementary Table 1). This result is consistent with previous reports that used similar or complementary strategies, and emphasizes that most functional lipid variants identified by GWAS may exert their effect on phenotypic variation through the regulation of gene expression [8–10,16–18,23,26,29,32].

2.2. Integrating human eQTL data

A large meta-analysis of genome-wide association results for lipid levels highlighted variants at 24 of 95 lipid loci that are eQTL in human liver at $P < 5 \times 10^{-8}$ [25,28]. Given our enrichment results, we reasoned that the specific causal variant(s) at each of these eQTL should be either the sentinel SNP itself or a marker in strong LD with it, and marked by epigenomic annotations in HepG2 cells. Because the presence or absence of epigenomic annotations at markers within the same locus should be independent of LD between them, ENCODE data could help prioritize functional variants even if they are perfectly correlated (a limitation of the genetic approach in fine-mapping GWAS loci).

The simplest strategy to combine epigenomic annotations and DNA polymorphisms is to count the number of DHS and ChIP-seq peaks that physically map in the human genome at the same position as DNA polymorphisms. Our hypothesis is that the best functional candidate variant at an eQTL lipid locus should have the highest number of overlaps with epigenomic annotations in HepG2, thus allowing discrimination between variants in strong LD. Obviously, this one causal variant-one locus hypothesis would not be valid if there is evidence of indepencent association signals or in the presence of several causal variants in strong LD, as recently proposed in the genomic context of super-enhancers [7,14,22]. However, under the several causal variants-one locus model, our framework might still identify at least one of the potential functional variants. For this analysis, we used all DHS and histone mark peaks; we also included ChIP-seq data for all available transcription factors since most of them were examined specifically in hepatocytes or are general activators or repressors of transcription without a clear cell- or biological pathway-specificity. Importantly, epigenomic annotations are biologically correlated as many mark the same chromatin state (e.g. promoters, enhancers) [12]. However, they also each provide experimental evidence that a genomic region is transcriptionally important. In addition, the accumulation of DHS and ChIP-seq peaks from different experiments (and for ENCODE, different laboratories) at a given position in the genome decreases the likelihood of false positives. For these reasons, we treated all DHS, histone marks and transcription factors ChIP-seq data from ENCODE HepG2 independently (including technical replicates when available) and used them to annotate SNPs. Merging technical replicates to only analyze intersecting peaks had no significant impact on the results.

Results from this analysis are summarized in Table 1. At 19 of the 24 eQTL, the variant with the highest number of overlaps with ENCODE epigenomic annotations in HepG2 was different than the reported sentinel lipid SNP. The candidate SNPs prioritized by the ENCODE data were also on average closer, although not significantly, to the transcription start site(s) of the eQTL gene(s) than the sentinel lipid SNPs (78 ± 82 vs. 88 ± 93 kilobases (kb)), but still sufficiently far to suggest an influence on enhancer as opposed to promoter activities. We performed a receiver operating characteristic (ROC) curve analysis to determine the number of overlapping epigenomic annotations that maximize both sensitivity and specificity of finding candidate SNPs at eQTL. We compared the number of epigenomic annotations for each SNP within the 24 eQTL with the number for each SNP in the 71 non-eQTL, focusing on the SNP with the highest number of epigenomic annotations in each locus. At a threshold of 16 ovelapping epigenomic annotations, the area under the curve (AUC) is 0.618, the sensitivity 67% and the specificity 61%. If a SNP has ≥16 epigenomic annotations in HepG2, it is more likely to be located at an eQTL in liver (Fisher's exact $P = 0.03$, odds ratio and 95% confidence interval = 3.1 [1.1–9.6]). Using a threshold of 16 epigenomic annotations, we found a functional candidate SNP for 16 of the 24 lipid and gene expression levels loci (bold in Table 1). For each of the 16 loci, we list all SNPs in strong LD ($r^2 \geq 0.8$) that overlap with ≥16 epigenomic annotations in Supplementary Table 2.

As a positive control, we evaluated the priority of rs12740374, a SNP near *SORT1* previously proposed to be a causal lipid variant at this locus by interfering with binding of C/EBP transcription factors [20]. At the *SORT1* locus, we identified rs12740374 as the most likely functional regulatory variant based on 44 epigenomic annotation overlaps in comparison with 23 overlaps for the second most likely SNP (empirical $P = 0.048$, calculated using the two variants with the highest number of annotations in each of the 5000 matched sets of 95 SNPs) and 13 overlaps for rs629301, the sentinel lipid SNP (Fig. 1A). Another promising example is at the *NFATC3* locus. The sentinel lipid SNP rs16942887 that is associated with *NFATC3* expression levels in human livers is located 191 kb upstream of its transcription start site. The highest priority candidate SNP at the locus in our analysis, rs7188085, has 81 epigenomic annotation overlaps in HepG2 (vs. 20 for rs16942887) and is located only 5.3 kb upstream of *NFATC3* (Fig. 1B). This variant and many others presented in Table 1 are strong functional candidates.

2.3. Combining ENCODE and mouse transcriptomic data

Despite a very strong enrichment of epigenomic annotations correlated with transcriptional regulation (Supplementary Table 2), only 36% of the 95 loci associated with lipid levels in humans were reported to harbor eQTL variants [28]. Many factors could explain this observation: transcriptomic profiling was performed in the wrong tissues, the genotypic effect on gene expression was too weak to be detected, the transcripts of interest were not measured or were undetectable, etc.

One alternative to gene profiling in human samples is to use the mouse, where the relevant tissues are readily accessible, and assume that transcription factor homologs will target a large set of overlapping