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/

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

Table 1
Overlaps of epigenomic annotations from ENCODE HepG2 and sentinel lipid SNPs associated with gene expression levels in human livers. For each sentinel lipid SNP, we identified SNPs in linkage disequilibrium ($r^2 \geq 0.8$, European populations from the 1000 Genomes Project) and counted the number of overlaps with ENCODE peaks for all DNAse I hypersensitive sites and ChIP-seq data available. ENCODE top candidate SNPs with ≥16 epigenomic annotation overlaps are in bold (see text for details). TSS, transcription start site; bp, base pairs. Human liver eQTL data from [25,28].

Sentinel lipid SNP	Chr:Position (hg19)	Transcript(s) associated with genotypes at the sentinel lipid SNP in human livers	ENCODE top candidate SNP (in LD with sentinel lipid SNP; highest number of epigenomic annotation overlaps)	Chr:Position (hg19)	Number of overlapping ENCODE epigenomic annotations	Distance between ENCODE top candidate SNP and gene TSS (bp)	Distance between sentinel lipid SNP and gene TSS (bp)	Distance between sentinel lipid SNP and ENCODE top candidate SNP (bp)
rs12027135	Chr1:25,775,733	*RHCE*	**rs9438904**	Chr1:25,756,860	46	9497	28,370	−18,873
		RHD				−157,880	−176,753	
		TMEM50A				−92,072	−110,945	
		TMEM57				527	−18,346	
rs2131925	Chr1:63,025,942	*ANGPTL3*	**rs631106**	Chr1:62,901,807	47	161,379	37,244	−124,135
		DOCK7				−252,232	−128,097	
rs629301	Chr1:109,818,306	*CELSR2*	**rs12740374**	Chr1:109,817,590	44	−24,950	−25,666	−716
		PSMA5				−151,480	−150,764	
		PSRC1				−8200	−7484	
		SORT1				−122,973	−122,257	
		SYPL2				191,509	190,793	
rs1260326	Chr2:27,730,940	*IFT172*	**rs780094**	Chr2:27,741,237	23	28,666	18,369	10,297
rs13107325	Chr4:103,188,709	*SLC39A8*	rs13107325	Chr4:103,188,709	0	−77,946	−77,946	0
rs9488822	Chr6:116,312,893	*FRK*	rs9488822	Chr6:116,312,893	4	−69,028	−69,028	0
rs10128711	Chr11:18,632,984	*SPTY2D1*	**rs7943121**	Chr11:18,656,062	49	42	−23,036	23,078
rs174546	Chr11:61,569,830	*FADS1*	**rs174538**	Chr11:61,560,081	49	−24,448	−14,699	−9749
rs11220462	Chr11:126,243,952	*ST3GAL4*	rs2066985	Chr11:126,251,286	5	22,024	29,358	7334
rs7134594	Chr12:110,000,193	*MMAB*	**rs10161126**	Chr12:110,042,348	17	30,990	−11,165	42,155
rs8017377	Chr14:24,883,887	*NYNRIN*	rs72694393	Chr14:24,874,193	9	−6202	−15,896	−9694
rs2929282	Chr15:44,245,931	*CKMT1A*	rs4270152	Chr15:44,224,668	5	−239,585	−260,848	−21,263
rs1532085	Chr15:58,683,366	*ALDH1A2*	rs2043085	Chr15:58,680,954	4	322,833	325,245	−2412
		LIPC				43,220	40,808	
rs11649653	Chr16:30,918,487	*VKORC1*	rs11640961	Chr16:30,979,818	4	−126,458	−187,789	61,331
rs16942887	Chr16:67,928,042	*NFATC3*	**rs7188085**	Chr16:68,113,873	81	5395	191,226	185,831
rs11869286	Chr17:37,813,856	*PERLD1*	**rs881844**	Chr17:37,810,218	33	−34,092	−30,454	−3638
rs7206971	Chr17:45,425,115	*TBKBP1*	**rs4793978**	Chr17:45,698,175	18	74,454	347,514	273,060
rs7241918	Chr18:47,160,953	*LIPG*	**rs7239867**	Chr18:47,164,717	32	−76,291	−72,527	3764
rs7255436	Chr19:8,433,196	*ANGPTL4*	**rs10413136**	Chr19:8,452,879	16	−23,869	−4186	19,683
rs439401	Chr19:45,414,451	*APOC4*	**rs584007**	Chr19:45,416,478	31	29,016	31,043	2027
rs386000	Chr19:54,792,761	*LILRA3*	rs386000	Chr19:54,792,761	2	−11,460	−11,460	0
rs2277862	Chr20:34,152,782	*CEP250*	**rs2104417**	Chr20:34,127,871	25	−84,649	−109,560	−24,911
		CPNE1				−124,988	−100,077	
rs6065906	Chr20:44,554,015	*PLTP*	**rs1057208**	Chr20:44,563,007	49	22,004	13,012	8992
rs181362	Chr22:21,932,068	*UBE2L3*	**rs2266959**	Chr22:21,922,904	18	−886	−10,050	−9164

genes in both species. In particular, we tested the hypothesis that the disruption of specific transcription factors in mouse livers could help identify functional lipid genes and variants. First, we performed an enrichment analysis of all the ENCODE HepG2 ChIP-seq transcription factor data over the sentinel and correlated SNPs at the 95 lipid loci and identified ten transcription factors that preferentially bind to these regions: *CEBPB*, *ELF1*, *FOXA1*, *FOXA2*, *HEY1*, *HNF4A*, *HNF4G*, *MBD4*, *MYBL2*, and *NFIC* (Supplementary Table 3). This enrichment was reproducible across technical replicates. These transcription factors may define regulatory networks that are important to control lipid metabolism in humans. Of particular interest, we saw an enrichment for three families of transcription factors expressed in the liver and previously implicated in lipid metabolism: CEBPB, FOXA1 and FOXA2, and HNF4A. Second, we identified publicly available transcriptomic profiles in livers of control mice and liver-specific knockout animals for *Foxa1* and *Foxa2* [3], and *Hnf4a* [5]; unfortunately, such data was not available for *Cebpb*. For each of these conditional gene knockout strains, we retrieved the list of mouse genes whose expression in liver was significantly changed when compared to control animals: 385, 1009 and 1179 genes for *Foxa1*, *Foxa2* and *Hnf4a*, respectively (see Methods section). Third, we searched if any of the human homologs of these target genes were located within an arbitrary window defined as 250 kb on each side of the 95 sentinel lipid SNPs. For *FOXA2*, we found ten target genes located within nine of the 95 lipid loci; all but one of these loci contain at least one FOXA2 ChIP-seq peak in HepG2 (Table 2 and Supplementary Table 3). Results were similarly encouraging for *HNF4A*: there are 20 transcriptional target genes located at 17 of the 95 lipid loci, and for 14 of these 17 loci, there is at least one annotated HNF4A ChIP-seq peak in HepG2 (Table 2 and Supplementary Table 3). Because we demonstrated a strong statistical enrichment of FOXA2 and HNF4A ChIP-seq peaks at the human lipid loci, and because we focus our query on genes modulated by the disruption of these transcription factors in mouse livers, we argue that the genes listed in Table 2 are strong biological candidates for influencing lipid levels in humans. Our screen re-identified genes previously implicated in lipid metabolism, such as *SORT1* and *GALNT2*, but also other genes with unanticipated functions in regulating lipid levels (Supplementary Table 3) [15,20]. There were no *FOXA1* target genes among these genomic regions, perhaps consistent with the previous finding that *FOXA1* is preeminently involved in cell cycle regulation [3].

2.4. Finding and characterizing potential functional variants

Our analysis presented in Table 2 also allowed us to try to predict functional variants. Indeed, if a sentinel lipid SNP (or an LD proxy) overlaps a FOXA2 or HNF4A ChIP-seq peak in HepG2 and disrupts a predicted binding site for these transcription factors, it is likely to be biologically relevant. We queried the HaploReg database and found that four SNPs disrupted binding motifs for FOXA2 and HNF4A (Table 2: rs3776702 and rs4969182 for *FOXA2*; rs838882 and rs12185764 for *HNF4A*) [30]. Many of the loci listed in Table 2 do not contain SNPs that disrupt predicted FOXA2 or HNF4A binding sites. This is consistent with results from the ENCODE Project that showed that ChIP-seq can identify

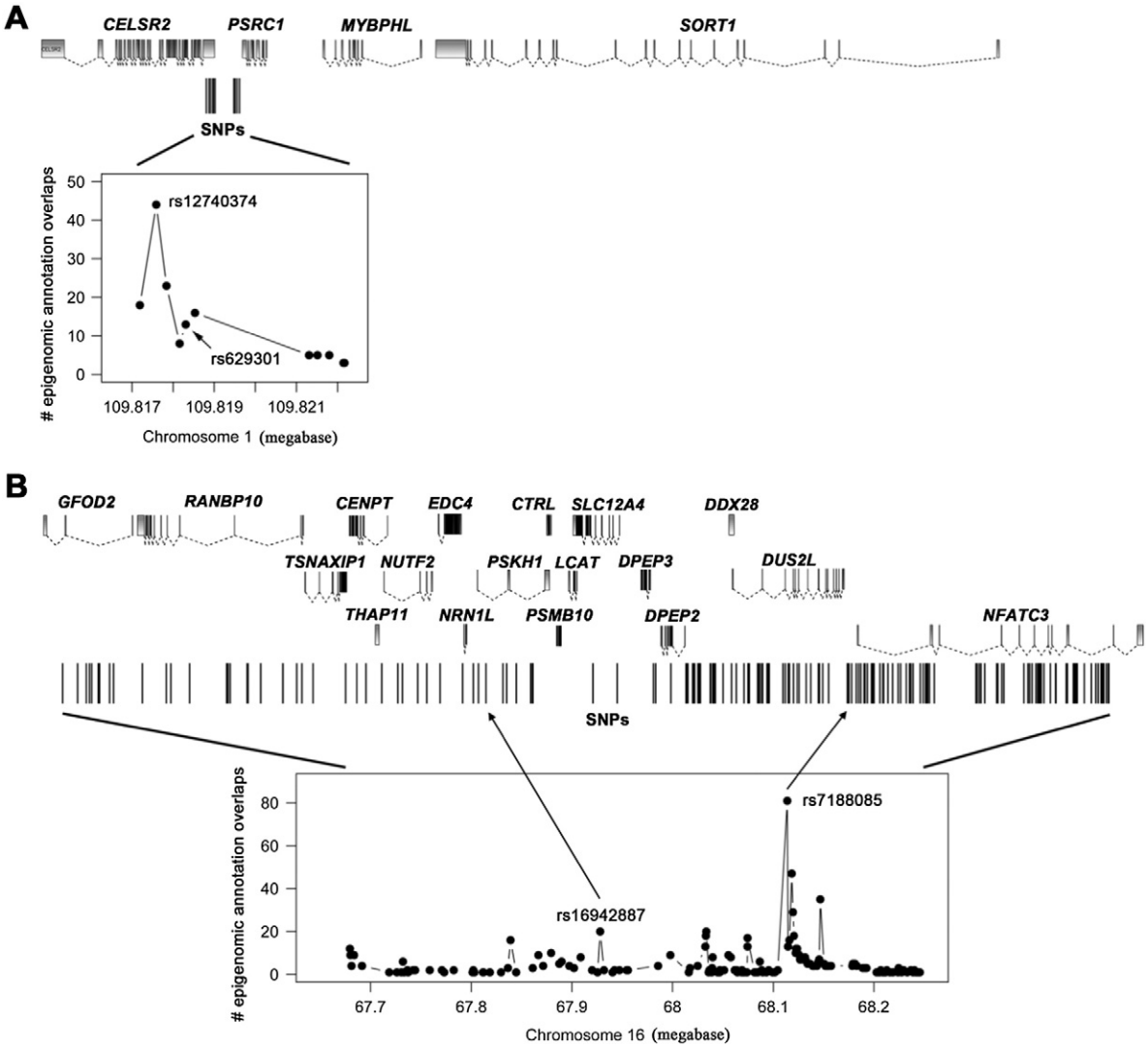

Fig. 1. ENCODE HepG2 epigenomic annotations at the (A) *SORT1* and (B) *NFATC3* lipid loci. For the sentinel lipid and eQTL SNPs (*SORT1*: rs629301; *NFATC3*: rs16942887) and their linkage disequilibrium proxies ($r^2 \geq 0.8$, European populations from 1000 Genomes Project), we counted the number of overlaps with peaks from HepG2 DNase I hypersensitive sites, histone marks or transcription factor binding ChIP-seq data. For both loci, the SNP with the highest number of epigenomic annotations is different than the published sentinel SNP.

numerous and robust transcription factor peaks with no consensus binding motif in the underlying DNA sequence [21,31]. In the absence of canonical binding sites, it is impossible to predict the effect of SNPs on transcription factor binding; this requires functional validation. Therefore, many of the variants listed in Table 2 might be functional even if they reside in FOXA2 or HNF4A ChiP-seq peaks that do not contain canonical binding motifs.

Finally, we sought to functionally validate one of our predictions. We selected rs4969182, which is in LD with the sentinel lipid SNP rs4129767 ($r^2 = 0.96$), overlaps with a FOXA2 peak in HepG2 and is located 171 kb away from the apoptosis-related gene *BIRC5*, a transcriptional target of *Foxa2* in mouse livers (Table 2). rs4969182 is a C/T bi-allelic variant, and the C-allele disrupts the motif recognized by FOXA transcription factors. Using reporter assays in HepG2 cells, we showed that the DNA sequence surrounding rs4969182 has enhancer activity, and that the T-allele recognized by FOXA2 shows significantly increased transcriptional activity compared to the C-allele (Fig. 2A, $P = 2.6 \times 10^{-5}$ and $P = 5.0 \times 10^{-6}$ in the forward and reverse orientation, respectively). Next, using electrophoretic mobility shift assays (EMSA), we tested if alleles of rs4969182 differentially affected DNA binding to nuclear proteins. Our results showed that proteins from HepG2 nuclear extracts bind probes containing either the C- or the T-allele, but that binding is stronger for the T-allele-containing probe (Fig. 2B). Competition of T-allele-containing labeled probe with excess unlabeled probe with the T-allele more efficiently competed away allele-specific bands than excess unlabeled probe with the C-allele, providing support for allelic differences in protein-DNA binding (Fig. 2B). Antibodies against FOXA1 and FOXA2 appear to weaken the probe–FOXA interaction but did not supershift the protein–probe complexes (Fig. 2B). Other examples exist of EMSA experiments in which antibodies appear to impair binding without causing a clear supershift of the complex [20].

3. Discussion

Characterization of the epigenome by the ENCODE Project provides a framework to functionally annotate non-coding SNPs identified by GWAS. Based on the observation that GWAS SNPs are enriched for chromatin marks linked to transcriptional regulation, we designed two novel strategies that integrate gene expression profiling with epigenomic characterization. We used SNPs associated with lipid levels as a test set because the large number allows us to derive meaningful statistics and also because relevant cells and tissues are characterized. First, we showed that at eQTL loci, simply counting the number of epigenomic annotations that overlap with associated SNPs can improve fine-mapping resolution. This is particularly useful to distinguish markers in strong LD, such as SNPs at the *SORT1* locus (Fig. 1A). Second,

Table 2
Identification of *HNF4A* and *FOXA2* potential regulatory variants at lipid loci. For each of these two transcription factors, we identified target genes in mouse livers, and then searched if the human homologs of these target genes were located within a 500 kilobase window around the 95 sentinel lipid SNPs. For the lipid loci that contain at least one target gene, we then query if the SNPs (or linkage disequilibrium proxies) overlapped with corresponding ENCODE ChIP-seq peaks and disrupted predicted binding sites. Two SNPs for each *HNF4A* and *FOXA2* (in bold) met all these criteria.

A-HNF4A

Sentinel lipid SNP	Chr:Position (hg19)	*HNF4A* target gene(s)	Is there a *HNF4A* ChIP-seq peak at the locus?	SNPs in LD with sentinel lipid SNP that overlap with ENCODE *HNF4A* peaks (Do they disrupt predicted *HNF4A* binding sites?)
rs4846914	Chr1:230,295,691	*GALNT2*	Yes	rs4846913 (No) rs2144300 (No)
rs1260326	Chr2:27,730,940	*FNDC4*, *SLC30A3*	No	
rs1800562	Chr6:26,093,141	*SLC17A3*, *HIST1H4D*, *HIST1H4F*	No	
rs17145738	Chr7:72,982,874	*MLXIPL*	Yes	rs34060476 (No)
rs2081687	Chr8:59,388,565	*UBXN2B*	No	
rs11136341	Chr8:145,043,543	*NRBP2*	No	
rs7134594	Chr12:110,000,193	*MMAB*, *MYO1H*	Yes	rs10744826 (No) rs10161126 (No)
rs1169288	Chr12:121,416,650	*HNF1A*	No	
rs4759375	Chr12:123,796,238	*SETD8*	Yes	rs10846506 (No)
rs838880	Chr12:125,261,593	*BRI3BP*	Yes	rs838881 (No) **rs838882** (Yes) rs838884 (No)
rs16942887	Chr16:67,928,042	*DUS2L*, *PSMB10*	Yes	rs7188085 (No) rs2107369 (No) rs8044328 (No)
rs4420638	Chr19:45,422,946	*BCAM*, *PVRL2*	No	
rs6029526	Chr20:39,672,618	*PLCG1*	No	
rs6065906	Chr20:44,554,015	*WFDC3*	Yes	rs6065905 (No) **rs12185764** (Yes)

B-FOXA2

Sentinel lipid SNP	Chr:Position (hg19)	*FOXA2* target gene(s)	Is there a *FOXA2* ENCODE peak at the locus?	SNPs in LD with sentinel lipid SNP that overlap with ENCODE *FOXA2* peaks (Do they disrupt predicted *FOXA2* binding sites?)
rs629301	Chr1:109,818,306	*SORT1*	Yes	rs7528419 (No) rs12740374 (No) rs660240 (No)
rs12328675	Chr2:165,540,800	*COBLL1*	Yes	rs7607980 (No)
rs2290159	Chr3:12,628,920	*PPARG*	Yes	rs60448371 (No) rs55762590 (No)
rs6450176	Chr5:53,298,025	*ARL15*	Yes	rs6889847 (No) rs6876198 (No) rs3776712 (No) rs1541681 (No) rs1541680 (No) rs3776707 (No) rs3776706 (No) rs3776705 (No) rs3776703 (No) **rs3776702** (Yes)
rs1800562	Chr6:26,093,141	*SLC17A3*	Yes	rs115740542 (No)
rs1169288	Chr12:121,416,650	*P2RX7*	Yes	rs6489786 (No) rs1169288 (No)
rs7206971	Chr17:45,425,115	*ITGB3*	Yes	rs4793978 (No) rs11079784 (No)
rs4129767	Chr17:76,403,984	*BIRC5*	Yes	**rs4969182** (Yes)
rs181362	Chr22:21,932,068	*HIC2*, *SDF2L1*	No	

in the absence of human eQTL information, or to complement such datasets, we used gene expression profiling in the mouse to prioritize candidate functional genes, and subsequently candidate functional variants. We reasoned that if a transcription factor binds preferentially at lipid loci (ENCODE ChIP-seq data), disruption of the mouse homolog could identify target genes that may be important for lipid levels variation in humans. Indeed, although it is known that transcription factors from different species bind different DNA motifs [27], the transcriptional target genes are often conserved across species [4,6]. This strategy allowed us to highlight the role of the *FOXA2* and *HNF4A* transcriptional networks in lipid metabolism. Importantly, we validated one of our predictions experimentally: a lipid sentinel SNP located 171 kb from *BIRC5*, a FOXA2 target gene in the mouse liver, is in LD with a marker that interferes with FOXA2 binding and modulates the enhancer activity of the DNA sequence (Fig. 2). We did not validate whether *BIRC5* plays a role in lipid metabolism; there are other potential candidate genes at the locus, although none are FOXA2 target genes based on the mouse data. Other candidates include *PGS1*, a gene involved in the biosynthesis of the anionic phospholipids phosphatidylglycerol and cardiolipin.

Fine-mapping may sometimes point to a candidate functional gene that will be different than what would be expected based on the known biology of the genes located within the locus. We have such an example in our analysis of eQTL data from human livers. rs16942887 is associated with HDL-cholesterol levels in humans [28], and is located 46 kb from *LCAT*, which encodes an important enzyme involved in cholesterol transport. Whereas common knowledge would suggest *LCAT* as the likeliest causal gene at the locus, genotypes at rs1692887 are associated with expression levels of *NFATC3* in human livers, a gene located

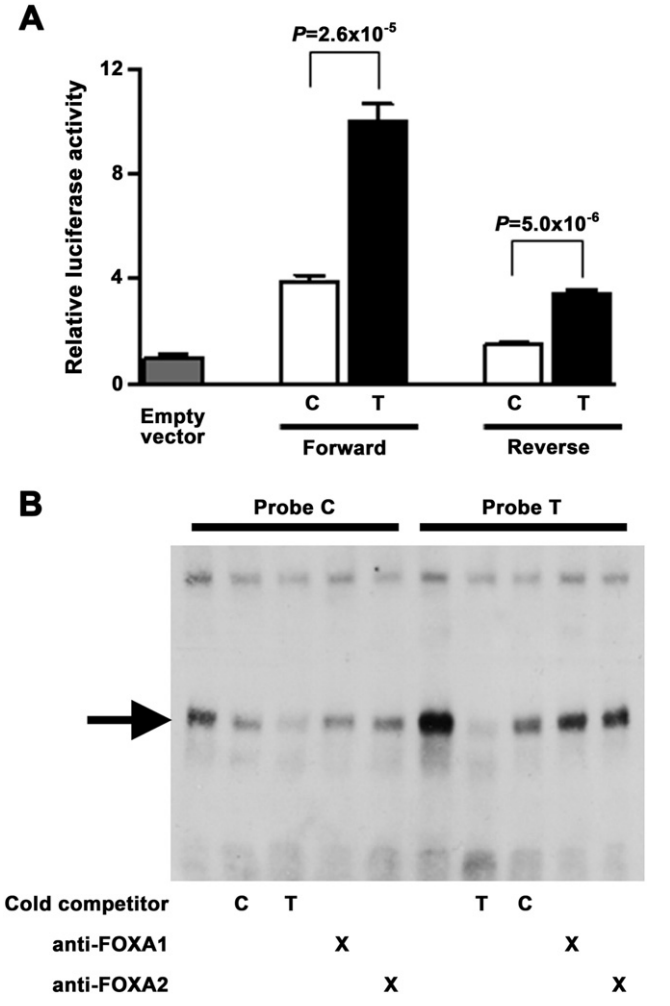

Fig. 2. Allelic differences in regulatory activity at rs4969182. (A) Differential transcriptional enhancer reporter activity in HepG2 cells. The T-allele, found in FOXA consensus binding motifs, showed significantly increased luciferase activity compared to C-allele in both orientations and with respect to a minimal promoter vector. Error bars represent standard error of five independent clones for each allele. Results are expressed as fold change compared to empty vector control. P-values were calculated by a two-sided *t*-test. (B) Electrophoretic mobility shift assay (EMSA) using HepG2 nuclear extract shows differential protein-DNA binding of rs4969182 alleles. The probe containing the T-allele shows increased protein binding (arrow A) compared to the probe containing the C-allele. Excess unlabeled specific probe containing the T-allele (T-comp) more efficiently competed away allele-specific binding than the unlabeled C-allele (C-comp). Incubation with FOXA1 and FOXA2 antibody reduced the DNA–protein complex (arrow). To enhance visualization of protein complexes, free biotin-labeled probe is not shown.

191 kb downstream. Furthermore, epigenomic characterization of this locus in HepG2 highlights rs7188085, a SNP in strong LD with rs1692887 ($r^2 = 0.85$) and located only 5.3 kb from the *NFATC3* transcription start site (Fig. 1B). *NFATC3* encodes a gene involved in immune responses. *LCAT* is critical for lipid metabolism in humans, but there is currently no functional evidence that suggests that the SNPs at this locus mediate their effect on HDL-C levels through *LCAT* itself, *NFATC3*, or both.

Several studies have proposed to use epigenomic annotations to prioritize DNA sequence variants at GWAS loci for functional testing [8–10,16–18,23,26,29,32]. We extended these methods and also developed a novel paradigm by proposing to integrate mouse transcriptomic data as an additional filter to prioritize candidate functional variants and genes. As with the other bioinformatic methods, ours also have limitations that are inherent to the type of data available. For instance, genomic regions that are difficult to sequence using next-generation DNA sequencers are less likely to be annotated by the ENCODE Project and might thus escape detection using such methods. Equally important is the fact that most epigenomic marks catalogued by the ENCODE Project are associated with transcriptional activation. Thus, functional genetic variants that relieve transcriptional repression are less likely to be found using these strategies. Finally, if the transcription factors tested by ChIP-seq do not have a mouse ortholog (unlikely since 99% of human genes have a mouse equivalent [19]) or if the mouse knockout models do not exist, our second strategy is not applicable.

In conclusion, we presented two simple strategies that combine epigenomic and transcriptomic profiling to prioritize functional genes and variants at GWAS loci. These methods should be applicable to prioritize rare genetic variants as well because they rely on the annotation of physical positions and are independent of allele frequency. The predictions from our approaches, which are statistically supported through enrichment analysis, are readily testable in the laboratory. These methods should be applicable to characterize genetic markers associated with many complex diseases and traits, and in particular those related to immune or hematological phenotypes as relevant tissues are easier to access. Combining human genetic findings with epigenomic characterization and gene expression data from mouse knockouts offer an alternative solution, in particular when human tissues are not accessible. Finally, as the repertoire of epigenomic annotations in various human tissues continue to expand, we anticipate that our strategies will become amenable to most human complex phenotypes.

4. Methods

4.1. ENCODE enrichment analysis

The enrichment pipeline strategy is summarized graphically in Supplementary Fig. 1. For each epigenomic annotation, peak coordinates were identified using software developed for the ENCODE Project (http://encodeproject.org/ENCODE/encodeTools.html). We obtained epigenomic annotations in the form of peak calls mapped onto the human genome (build hg19) directly from the ENCODE Project website (accessed June 2012). In total, we considered in our analysis 116, 147, 111, 177 different epigenomic annotations files for HepG2, GM12878, H1-hESC and K562, respectively. To quantify the enrichment of SNPs associated with a specific complex disease or trait, we developed a four step strategy: First, we generated sets of variants (with replacement) that are matched with the sentinel variants based on allele frequency ($\pm 4\%$), gene proximity (± 100 kb) and linkage disequilibrium (LD; all SNPs within the same set have $r^2 \leq 0.5$). For our analysis of the lipid loci, we generated 5000 sets of 95 SNPs using information from European individuals from the 1000 Genomes Project. Second, for each variant in the seed and matched sets, we retrieved all other variants in LD ($r^2 \geq 0.8$) using the 1000 Genomes Project European population genotypes and the PLINK software [24]. Third, we annotated all variants and their LD proxies for overlap with specified epigenomic annotations. Finally, we assessed statistical enrichment by computing empirical P-values for each epigenomic annotation by counting the number of matched set with more SNP-epigenomic annotation overlaps than found in the set of sentinel variants. We provide a step-by-step description of our methods in the Supplementary Information.

4.2. Gene expression datasets

Human liver eQTL results ($P \leq 5 \times 10^{-8}$) were available from previous reports [25,28]. The list of genes differentially expressed in liver-specific knockout *Foxa1*$^{-/-}$ and *Foxa2*$^{-/-}$ mice were obtained from a previous report (fold-change $\geq \pm 1.5$, false discovery rate = 15%) [3]. To identify the list of genes differentially expressed in *Hnf4a*$^{-/-}$ liver mice compared to wild-type animals, we recovered the corresponding dataset from NCBI Gene Expression Omnibus (GSE34581) [5] and analyzed the data with the GEO2R module, correcting for multiple testing using the Benjamini & Hochberg procedure (adjusted $P \leq 0.05$). We

converted mouse gene symbols to human gene symbols assuming a one-to-one homolog (Supplementary Information).

4.3. Luciferase transcriptional reporter assays

HepG2 hepatocellular carcinoma cells were cultured in MEM-alpha (Invitrogen) supplemented with 10% FBS, 1 mM sodium pyruvate and 2 mM L-glutamine. A 181 bp fragment (hg19 chr17: 76,392,913-76,393,093) surrounding the SNP rs4969182 was PCR-amplified using primers 5′-TGGAACACAGCCACACTCAT-3′ and 5′-ACTTGCACTCAGGTCCGTTT-3′ from DNA of individuals homozygous for either allele and cloned in both orientations into the multiple cloning sites of the minimal promoter-containing firefly luciferase reporter vector pGL4.23 (Promega, Madison, WI). Fragments are designated as 'forward' or 'reverse' based on their orientation in the genome with respect to the *BIRC5* coding sequence. Five independent clones for each allele for each orientation were isolated, verified by sequencing and transfected in duplicate into HepG2 cell line. Luciferase assays were performed as previously described [11].

4.4. Electrophoretic mobility shift assay (EMSA)

Nuclear cell extract was prepared from HepG2 cells using the NE-PER nuclear and cytoplasmic extraction kit (Thermo Scientific) as described [11]. 17 base-pair oligonucleotides were designed to the sequence surrounding rs4969182 alleles: Sense 5' biotin-ATATTTAC[**T/C**]CTCTGGCC-3', antisense 5'-biotin-GGCCAGAG[**G/A**]GTAAATAT-3' (SNP alleles in bold). For supershift assays, before adding labeled probe, 2 μg of polyclonal antibody against FOXA1 (ab23738; from ABCAM) or 4 ug of FOXA2 (ENCODE ChIP-seq antibody, SC-6554X; from Santa Cruz Biotechnology) was added to the binding reaction and incubated for 25 min. EMSAs were carried out on a second independent day and yielded comparable results.

Author contributions

Conceived and designed experiments: KSL, SV, MPF, KLM, GL
Performed experiments: KSL, SV
Directed the study: KLM, GL
All authors analyzed results and wrote the manuscript.

Disclosure declaration

The authors declare no conflicts of interest.

Acknowledgments

We thank investigators from the ENCODE Project for making the data publicly available. We also thank Cameron Palmer for sharing code to allow SNP matching. This work was funded by grants from the Centre of Excellence in Personalized Medicine (CEPMed), the "Fondation de l'Institut de Cardiologie de Montréal", the Canada Research Chair Program, the "Fonds de la Recherche en Santé du Québec" (to GL), and NIH grants DA027040 and DK072193.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2014.04.006.

References

[1] 1000 Genomes Project Consortium, A map of human genome variation from population-scale sequencing, Nature 467 (7319) (2010) 1061–1073.
[2] 1000 Genomes Project Consortium, An integrated map of genetic variation from 1,092 human genomes, Nature 491 (7422) (2012) 56–65.
[3] I.M. Bochkis, J. Schug, D.Z. Ye, S. Kurinna, S.A. Stratton, M.C. Barton, K.H. Kaestner, Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2, PLoS Genet. 8 (6) (2012) e1002770.
[4] S.F. Boj, J.M. Servitja, D. Martin, M. Rios, I. Talianidis, R. Guigo, J. Ferrer, Functional targets of the monogenic diabetes transcription factors HNF-1alpha and HNF-4alpha are highly conserved between mice and humans, Diabetes 58 (5) (2009) 1245–1253.
[5] J.A. Bonzo, C.H. Ferry, T. Matsubara, J.H. Kim, F.J. Gonzalez, Suppression of hepatocyte proliferation by hepatocyte nuclear factor 4alpha in adult mice, J. Biol. Chem. 287 (10) (2012) 7345–7356.
[6] E.T. Chan, G.T. Quon, G. Chua, T. Babak, M. Trochesset, R.A. Zirngibl, J. Aubin, M.J. Ratcliffe, A. Wilde, M. Brudno, et al., Conservation of core gene expression in vertebrate tissues, J. Biol. 8 (3) (2009) 33.
[7] O. Corradin, A. Saiakhova, B. Akhtar-Zaidi, L. Myeroff, J. Willis, R. Cowper-Sal lari, M. Lupien, S. Markowitz, P.C. Scacheri, Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits, Genome Res. 24 (1) (2014) 1–13.
[8] R. Cowper-Sal lari, X. Zhang, J.B. Wright, S.D. Bailey, M.D. Cole, J. Eeckhoute, J.H. Moore, M. Lupien, Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression, Nat. Genet. 44 (11) (2012) 1191–1198.
[9] I. Dunham, A. Kundaje, S.F. Aldred, P.J. Collins, C.A. Davis, F. Doyle, C.B. Epstein, S. Frietze, J. Harrow, R. Kaul, et al., An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57–74.
[10] J. Ernst, P. Kheradpour, T.S. Mikkelsen, N. Shoresh, L.D. Ward, C.B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, et al., Mapping and analysis of chromatin state dynamics in nine human cell types, Nature 473 (7345) (2011) 43–49.
[11] M.P. Fogarty, T.M. Panhuis, S. Vadlamudi, M.L. Buchkovich, K.L. Mohlke, Allele-specific transcriptional activity at type 2 diabetes-associated single nucleotide polymorphisms in regions of pancreatic islet open chromatin at the JAZF1 locus, Diabetes 62 (5) (2013) 1756–1762.
[12] M.B. Gerstein, A. Kundaje, M. Hariharan, S.G. Landt, K.K. Yan, C. Cheng, X.J. Mu, E. Khurana, J. Rozowsky, R. Alexander, et al., Architecture of the human regulatory network derived from ENCODE data, Nature 489 (7414) (2012) 91–100.
[13] L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, T.A. Manolio, Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, Proc. Natl. Acad. Sci. U. S. A. 106 (23) (2009) 9362–9367.
[14] D. Hnisz, B.J. Abraham, T.I. Lee, A. Lau, V. Saint-Andre, A.A. Sigova, H.A. Hoke, R.A. Young, Super-enhancers in the control of cell identity and disease, Cell 155 (4) (2013) 934–947.
[15] A.G. Holleboom, H. Karlsson, R.S. Lin, T.M. Beres, J.A. Sierts, D.S. Herman, E.S. Stroes, J. M. Aerts, J.J. Kastelein, M.M. Motazacker, et al., Heterozygosity for a loss-of-function mutation in GALNT2 improves plasma triglyceride clearance in man, Cell Metab. 14 (6) (2011) 811–818.
[16] L. Jia, G. Landan, M. Pomerantz, R. Jaschek, P. Herman, D. Reich, C. Yan, O. Khalid, P. Kantoff, W. Oh, et al., Functional enhancers at the gene-poor 8q24 cancer-linked locus, PLoS Genet. 5 (8) (2009) e1000597.
[17] K.J. Karczewski, J.T. Dudley, K.R. Kukurba, R. Chen, A.J. Butte, S.B. Montgomery, M. Snyder, Systematic functional regulatory assessment of disease-associated variants, Proc. Natl. Acad. Sci. U. S. A. 110 (23) (2013) 9607–9612.
[18] M.T. Maurano, R. Humbert, E. Rynes, R.E. Thurman, E. Haugen, H. Wang, A.P. Reynolds, R. Sandstrom, H. Qu, J. Brody, et al., Systematic localization of common disease-associated variation in regulatory DNA, Science 337 (6099) (2012) 1190–1195.
[19] C. Mouse Genome Sequencing, R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (6915) (2002) 520–562.
[20] K. Musunuru, A. Strong, M. Frank-Kamenetsky, N.E. Lee, T. Ahfeldt, K.V. Sachs, X. Li, H. Li, N. Kuperwasser, V.M. Ruda, et al., From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus, Nature 466 (7307) (2010) 714–719.
[21] S. Neph, J. Vierstra, A.B. Stergachis, A.P. Reynolds, E. Haugen, B. Vernot, R.E. Thurman, S. John, R. Sandstrom, A.K. Johnson, et al., An expansive human regulatory lexicon encoded in transcription factor footprints, Nature 489 (7414) (2012) 83–90.
[22] S.C. Parker, M.L. Stitzel, D.L. Taylor, J.M. Orozco, M.R. Erdos, J.A. Akiyama, K.L. van Bueren, P.S. Chines, N. Narisu, N.C.S. Program, et al., Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants, Proc. Natl. Acad. Sci. U. S. A. 110 (44) (2013) 17921–17926.
[23] M.M. Pomerantz, N. Ahmadiyeh, L. Jia, P. Herman, M.P. Verzi, H. Doddapaneni, C.A. Beckwith, J.A. Chan, A. Hills, M. Davis, et al., The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer, Nat. Genet. 41 (8) (2009) 882–884.
[24] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, et al., PLINK: a tool set for whole-genome association and population-based linkage analyses, Am. J. Hum. Genet. 81 (3) (2007) 559–575.
[25] E.E. Schadt, C. Molony, E. Chudin, K. Hao, X. Yang, P.Y. Lum, A. Kasarskis, B. Zhang, S. Wang, C. Suver, et al., Mapping the genetic architecture of gene expression in human liver, PLoS Biol. 6 (5) (2008) e107.
[26] M.A. Schaub, A.P. Boyle, A. Kundaje, S. Batzoglou, M. Snyder, Linking disease associations with regulatory information in the human genome, Genome Res. 22 (9) (2012) 1748–1759.
[27] D. Schmidt, M.D. Wilson, B. Ballester, P.C. Schwalie, G.D. Brown, A. Marshall, C. Kutter, S. Watt, C.P. Martinez-Jimenez, S. Mackay, et al., Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding, Science 328 (5981) (2010) 1036–1040.

[28] T.M. Teslovich, K. Musunuru, A.V. Smith, A.C. Edmondson, I.M. Stylianou, M. Koseki, J.P. Pirruccello, S. Ripatti, D.I. Chasman, C.J. Willer, et al., Biological, clinical and population relevance of 95 loci for blood lipids, Nature 466 (7307) (2010) 707–713.
[29] G. Trynka, C. Sandor, B. Han, H. Xu, B.E. Stranger, X.S. Liu, S. Raychaudhuri, Chromatin marks identify critical cell types for fine mapping complex trait variants, Nat. Genet. 45 (2) (2013) 124–130.
[30] L.D. Ward, M. Kellis, HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants, Nucleic Acids Res. 40 (2012) D930–D934 (Database issue).
[31] T.W. Whitfield, J. Wang, P.J. Collins, E.C. Partridge, S.F. Aldred, N.D. Trinklein, R.M. Myers, Z. Weng, Functional analysis of transcription factor binding sites in human promoters, Genome Biol. 13 (9) (2012) R50.
[32] X. Zhang, R. Cowper-Sal lari, S.D. Bailey, J.H. Moore, M. Lupien, Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus, Genome Res. 22 (8) (2012) 1437–1446.