Genomics 107 (2016) 248–254

Contents lists available at ScienceDirect

Genomics

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

Allele-specific transcription factor binding in liver and cervix cells unveils many likely drivers of GWAS signals

Marco Cavalli [a], Gang Pan [a], Helena Nord [a,1], Emelie Wallén Arzt [a,2], Ola Wallerman [a,b], Claes Wadelius [a,*]

[a] Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
[b] Science for Life Laboratory, Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 18 April 2016
Accepted 24 April 2016
Available online 26 April 2016

Keywords:
Allele-specific regulation
Association to GWAS/eQTLs
Functional variants

ABSTRACT

Genome-wide association studies (GWAS) point to regions with associated genetic variants but rarely to a specific gene and therefore detailed knowledge regarding the genes contributing to complex traits and diseases remains elusive. The functional role of GWAS-SNPs is also affected by linkage disequilibrium with many variants on the same haplotype and sometimes in the same regulatory element almost equally likely to mediate the effect.

Using ChIP-seq data on many transcription factors, we pinpointed genetic variants in HepG2 and HeLa-S3 cell lines which show a genome-wide significant difference in binding between alleles. We identified a collection of 3713 candidate functional regulatory variants many of which are likely drivers of GWAS signals or genetic difference in expression. A recent study investigated many variants before finding the functional ones at the *GALNT2* locus, which we found in our genome-wide screen in HepG2. This illustrates the efficiency of our approach.

1. Introduction

The aim of the human genome project was to learn more about human biology in general and disease biology in particular. Both goals have been met and in particular knowledge has increased dramatically on gene variants that predispose to common diseases. The hope has been that such knowledge could be used to improve health by preventive measures or development of new drugs. It is well known that the failure rate is high in the drug development pipeline but a recent study [1] provides reason for hope. They showed that at the preclinical stages few drugs are acting on targets supported by genetic studies but among approved drugs such targets are significantly more prevalent and they estimate that selecting targets based on genetics could double the success rate in clinical development. For such a strategy to be successful, knowledge has to be gained on the genes and pathways mediating the effect. A genome-wide association study (GWAS) points to a region with associated genetic variants but rarely to a specific gene and therefore detailed knowledge regarding the genes contributing to complex diseases has increased slowly. It has been estimated that 85% of the functional variants that drive the associations are located in regulatory elements and lead to different activity from the two alleles causing a difference in activity of the gene it regulates. The current GWAS catalog [2] contains >16.000 SNPs associated to disease but to our knowledge in <30 cases has a clear connection been made from the associated SNP to gene e.g. the ones regulating *SORT1* [3], *RFX6* [4] and *TOX3* [5]. This despite large efforts from international consortia, like ENCODE [6], Epigenome Roadmap and GTEx [7] [8], that have generated basic information on functional DNA sequences and genes.

A GWAS identifies SNPs with significant association to a disease or trait but due to linkage disequilibrium (LD) there can be many other variants that are almost equally likely to mediate the effect. Such variants are rarely located in coding regions and the widely accepted assumption is that most variants driving the association are located in a regulatory element that affects the activity of a gene nearby. A second problem is that regulatory elements like enhancers and silencers can act on genes over a considerable distance making it difficult to predict target genes. However, if the functional regulatory element has been identified, the regulated gene can be found experimentally by changing the sequence e.g. by using CRISPR, by overexpressing or knocking down transcription factors binding to the element or by studying 3 dimensional interactions [9–11]. Finding the functional regulatory elements is thus an important step towards defining disease mechanisms.

It is therefore desirable to have a collection of candidate regulatory variants and we set out to find them in a systematic way. A common feature for the published functional variants is that the two alleles of the driving SNP have different affinity for a transcription factor (TF). We

* Corresponding author at: Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, BMC, 75108 Uppsala, Sweden.
E-mail address: claes.wadelius@igp.uu.se (C. Wadelius).
[1] Present address: Department of Pre-Clinical Development, Galderma, Uppsala, Sweden.
[2] Present address: Department of Biosciences and Nutrition, Center for Biosciences, Karolinska Institute, Novum, Huddinge, Sweden.

http://dx.doi.org/10.1016/j.ygeno.2016.04.006

were the first to show that such signals can be detected using chromatin immunoprecipitation (ChIP) [12] and we [13] and others [14] have shown that they can be detected in data from large-scale sequencing (ChIP-seq). Here we established the genetic variants in the hepatocellular carcinoma cell line HepG2 and used the public sequence from the cervix cancer HeLa-S3 cell line. We then used public ChIP-seq data from these cell lines on many transcription factors and defined which variants that show a genome-wide significant difference in binding between alleles. This gave a collection of thousands of candidate functional regulatory variants many of which are likely drivers of GWAS signals or genetic difference in expression [15].

2. Material and methods

2.1. Genome sequencing

The HepG2 genome was sequenced on Illumina HiSeq to 10× coverage using 100 bp paired-end reads which in combination with reads from published HepG2 experiments [6] and from our ChIP-seq experiments and nucleosome sequencing [16] on the SOLiD system gave an average 55× genome coverage. We aligned all Illumina reads using BWA and SOLiD reads using BFAST and removed duplicate reads for each sample. SNP calling was done using the GATK unified genotyper. Sequencing, SNP calling and quality control is further described in Supplementary materials.

2.2. ChIP-seq sequences

Raw ChiP-seq reads (.fastq) were obtained from the ENCODE project database (ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/) selecting all the TF ChIP-seq data available at the time of download for HepG2 (data at Jul., 2012) and HeLa-S3 (Dec., 2013).

2.3. Genomic features

AS-SNPs collections were intersected and filtered with several publicly available databases: NHGRI GWAS catalog (Jan., 2014), collection of signal artifact blacklisted ENCODE regions [6], 1000 Genomes SNPs collection (1000 Genomes project, phase1_release_v3.20101123), liver tissue eQTLs collections (http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi# and http://eqtl.uchicago.edu/cgi-bin/gbrowse/eqtl/). The reference genome (G1) used was the UCSC hg19 assembly based on the Genome Reference Consortium Human genome build 37 (GRCh37) but excluding random and unplaced contigs. The alternative genomes (G2), when not available, were built for the different cell lines using the FastaAlternateReferenceMaker GATK utility that generates an alternative reference sequence replacing the reference bases at variation sites with the bases supplied by a cell-specific SNPs collection. The sources for cell-specific SNPs collections and alternative genomes were for HeLa-S3: NIHs dbGaP restricted access to hybrid assemblies of haplotype A and B on GRCh37 scaffold and SNPs calls, and for HepG2 SNP calls were made as described above.

2.4. AS-SNPs selection pipeline

The bioinformatics pipeline to identify AS-SNPs is described in details in Cavalli et al. [17]. The main steps (see Fig. 1) are summarized here for clarity:

(1) Alignment of ChIP-seq reads to the reference (G1) and alternative (G2) genome using ASAP (http://www.bioinformatics.babraham.ac.uk/projects/ASAP/).

(2) Reads mapped specifically to G1 or G2 were counted at the heterozygous SNPs.

(3) To determine whether the G1/G2 read counts difference was statistically significant a binomial test was applied against the null hypothesis of an equal G1:G2 coverage. After correcting for multiple testing (Benjamini & Hochberg or FDR), AS-SNPs with $P < 0.05$ were selected.

(4) AS-SNPs were then intersected with the 1000 Genomes SNPs collection in order to retrieve AFs.

(5) Extensive filtering of the selected AS-SNPs was performed to minimize the false positives where the difference in read count could be influenced by the genomic abnormal location of the SNPs (centromeric or telomeric regions, blacklisted ENCODE regions or CNVs).

(6) Pruned AS-SNPs selections were finally intersected with collections of GWAS or eSNPs and SNPs in LD ($r^2 > 0.5$) with GWAS or eSNPs to select candidate functional AS-SNPs for experimental validation.

2.5. Cell cultures

HepG2 cells were cultured in RPMI 1640 medium supplemented with 10% non-inactivated FBS, L-glutamine and a solution stabilized, with 10,000 units penicillin and 10 mg streptomycin/mL, sterile-

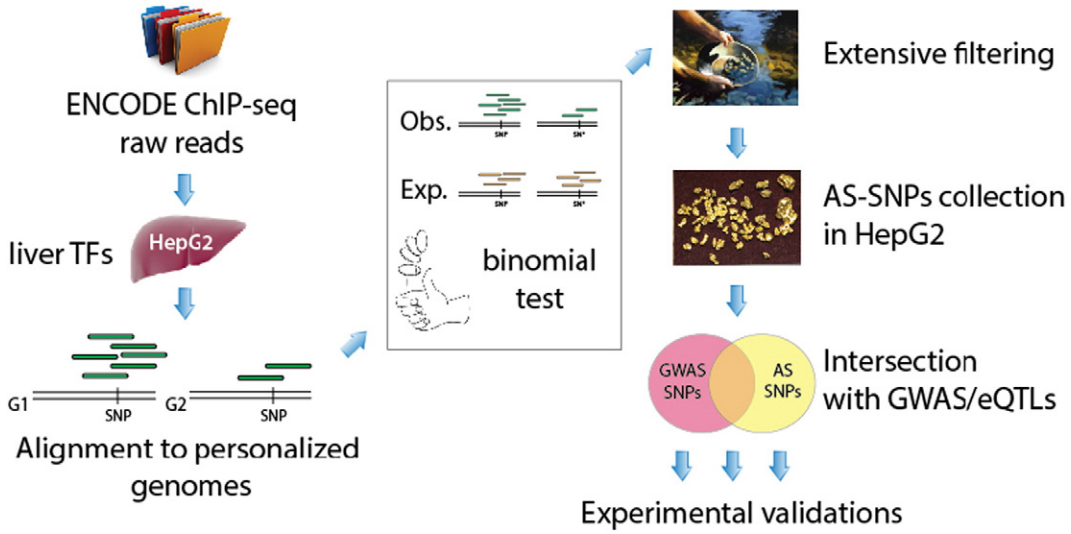

Fig. 1. Graphical summary of the AS-SNP discovery pipeline for the HepG2 cell line.

filtered, BioReagent, suitable for cell culture (Sigma-Aldrich) at 37 °C with 5% CO_2.

2.6. Construction of cloning plasmids and luciferase report assays

All the luciferase expression constructs were built based on pGL4.23 from Promega. The ccdB expression cassette was inserted into *Kpn*I and *Eco*RV sites of pGL4.23 to construct pGL4.23-ccdB, which was used as a basal vector to diminish false positive signal during the cloning process. Genomic sequences surrounding AS-SNPs were amplified by Phusion Hot Start Flex DNA polymerase (NEB) using HepG2 genomic DNA as template (see Table S3). The amplified fragments were purified by QIAquick Gel Extraction Kit (QIAGEN) and inserted upstream of the minimal promoter sequence of pGL4.23 by SLiCE cloning methods [18]. To get both of the alleles of all the AS-SNPs tested, multiple individual clones were picked up and subjected to Sanger sequencing. HepG2 cells were transfected one day after plating with approximately 90% confluence in 96-well plate. All the transfection reactions were carried out with X-tremeGENE HP DNA transfection reagent (Roche). Each well was transfected with 100 ng of firefly luciferase reporter vector harboring respective AS-SNPs alleles together with 1 ng of renilla luciferase reporter vector pGL4.74, which was used to normalize the transfection and lysis efficiency. Twenty-four hours after transfection, the cells were harvested and lysed in 1X passive lysis buffer (Promega) on a rocking platform for 45 min at room temperature. Firefly and luciferase activity were measured by Dual-Luciferase® Reporter (DLR™) Assay System (Promega) on an Infinite® M200 pro reader (TECAN) following instructions provided by the manufacturer. The ratios of firefly luciferase activity to renilla luciferase activity were calculated and expressed as Relative Luciferase Units (RLU) in the figures. All data came from four to six replicate wells, and *p*-values comparing RLU difference between AS-SNPs alleles were calculated using two-tailed *t*-test.

3. Results

3.1. Sequence and allele-specific signals in HepG2 and HeLaS3 cell lines

We sequenced the genomes of the liver cell line HepG2 and established that the allele calls were reliable using a genotyping array and downloaded public data on the genome sequence of the HeLa-S3 cervix cancer cell line [19]. The ENCODE project has generated ChIP-seq data for 55 and 57 TFs respectively from these cell lines and this data was downloaded. We used the Allele Specific Alignment Pipeline (ASAP) (http://www.bioinformatics.babraham.ac.uk/projects/ASAP/) to align the reads to the reference (G1) and alternative (G2) alleles, respectively. We counted the number of reads mapping to the G1 and G2 alleles at all heterozygous positions and those with a genome-wide statistically significant difference in the number of reads were identified after correcting for multiple testing and copy number variation (CNV) (see Material and methods and Fig 1). We investigated the number of reads mapping to the two genomes and found only small differences and in line with previous studies [14] concluding that reference and other alignment bias are well controlled for. We filtered the data to remove potential false positives in repeated and ENCODE "blacklisted" sequences.

In HepG2 we found 3001 SNPs with an allele-specific signal (AS-SNP) and 712 in HeLaS3 cells. Only 34 AS-SNPs were shared between the cells. This indicates that many regulatory elements are unique to each cell and that common elements rarely show functional genetic variation.

To validate the allele specific binding results from ChIP-seq, we tested 39 AS-SNPs detected in HepG2 in luciferase assays (see Material and methods). The AS-SNPs were either 1) randomly chosen AS-SNPs with common AF or 2) AS-SNPs associated to expression or 3) GWAS traits as explained below. An allele specific difference in activity was verified for 27/39 (69%) of the AS-SNPs (Fig. 2). We also tested 9 SNPs without a significant difference in ChIP-seq reads numbers between alleles and in no case was there a difference in activity. We concluded that variants considered to be AS-SNPs according to our definition are highly likely to be functional based on their location at regulatory elements, difference in TF binding between alleles in ChIP-seq and validation in functional tests such as luciferase assays.

3.2. AS-SNPs associated to disease and gene expression

Not all entries in the GWAS catalog can have a molecular explanation in the cells we study here. Genetic predisposition that could be mediated by the cells investigated are for example metabolic diseases for HepG2, and diseases of cervix for HeLaS3 (Table S1). We therefore searched the GWAS catalog for these traits. The SNP with the strongest association (GWAS top hit) was collected from the catalog and SNPs in high ($r^2 > 0.8$) or more relaxed ($r^2 > 0.5$) LD were identified and intersected with the collection of AS-SNPs. For 24 liver specific traits, we found 69 unique AS-SNPs that were candidates to explain GWAS signals with only 3 being the particular SNP reported in the GWAS catalog and the other 66 were in the defined LD intervals (Table 1, Fig. 3A and Table S5). We grouped the SNPs in 1 Mb loci and found candidate functional SNPs at 37 unique loci in HepG2.

Less cervical cancer specific traits are available in the GWAS catalog so we compared the list of AS-SNPs in HeLa-S3 cells to the full GWAS SNP collection (Fig. 3A and Table S8) as reported below.

We also observed a pattern where several GWAS-SNPs were associated to the same AS-SNP which is compatible with the fact that GWAS in different populations often show strongest association to different SNPs (Fig 3A and B). We also found that several AS-SNPs were detected at one locus e.g. seven SNPs located in different regulatory elements at the SLC7A5 gene associated to blood metabolite levels (Table S5). This suggests that SNPs at distinct regulatory elements could regulate activity of the same gene, which is supported by recent data [20,21].

Furthermore, we took the top hits from the whole GWAS catalog including SNPs in LD ($r^2 > 0.5$) and compared to the list of AS-SNPs. In this wide search we found 337 AS-SNPs (Tables S6 and S8) and some of them may be functional due to pleiotropic effects. We found novel putative functional variants that may explain GWAS SNPs using liver tissue. One example is the response to temozolomide, which is an alkylating agent causing methylation of guanine residues that lead to single and double-strand breaks in DNA and is used for the treatment of astrocytoma and glioblastoma, for which the top GWAS hits are located in *MGMT*. This gene encodes a methylguanine-DNA methyltransferase involved in DNA repair that is expressed in most tissues including the liver. In HepG2 cells, the two AS-SNPs rs577227 and rs524545 are located at regulatory elements 7 kb apart in the second intron of *MGMT* and are thus good candidates to drive the effect of the GWAS top hit rs477692. Most cells are likely to respond in the same way to temozolomide so the results from HepG2 may be representative of many cell types.

We also searched for AS-SNPs that may explain eQTL signals i.e. SNPs associated to gene expression in liver tissue [15]. We examined whether AS-SNPs identified in HepG2 are better candidates to drive the allelic difference in expression. We took SNPs in LD ($r^2 > 0.5$) with 3238 eSNPs for liver and investigated how many that are AS-SNPs in HepG2. In HepG2 we found 4 AS-SNPs that are eSNPs, and 244 AS-SNPs that are in high LD with an eSNP (Table S7). The AS-SNPs are located in regulatory elements and show evidence of being functional so we think that they are good candidates to drive the allele-specific expression variation. In the same way as for disease associated AS-SNPs we find that only a small fraction of eSNPs show tentative functional effect and that many more candidates are in high LD. This suggests that the eSNPs themselves are not good proxies as drivers of GWAS signals.

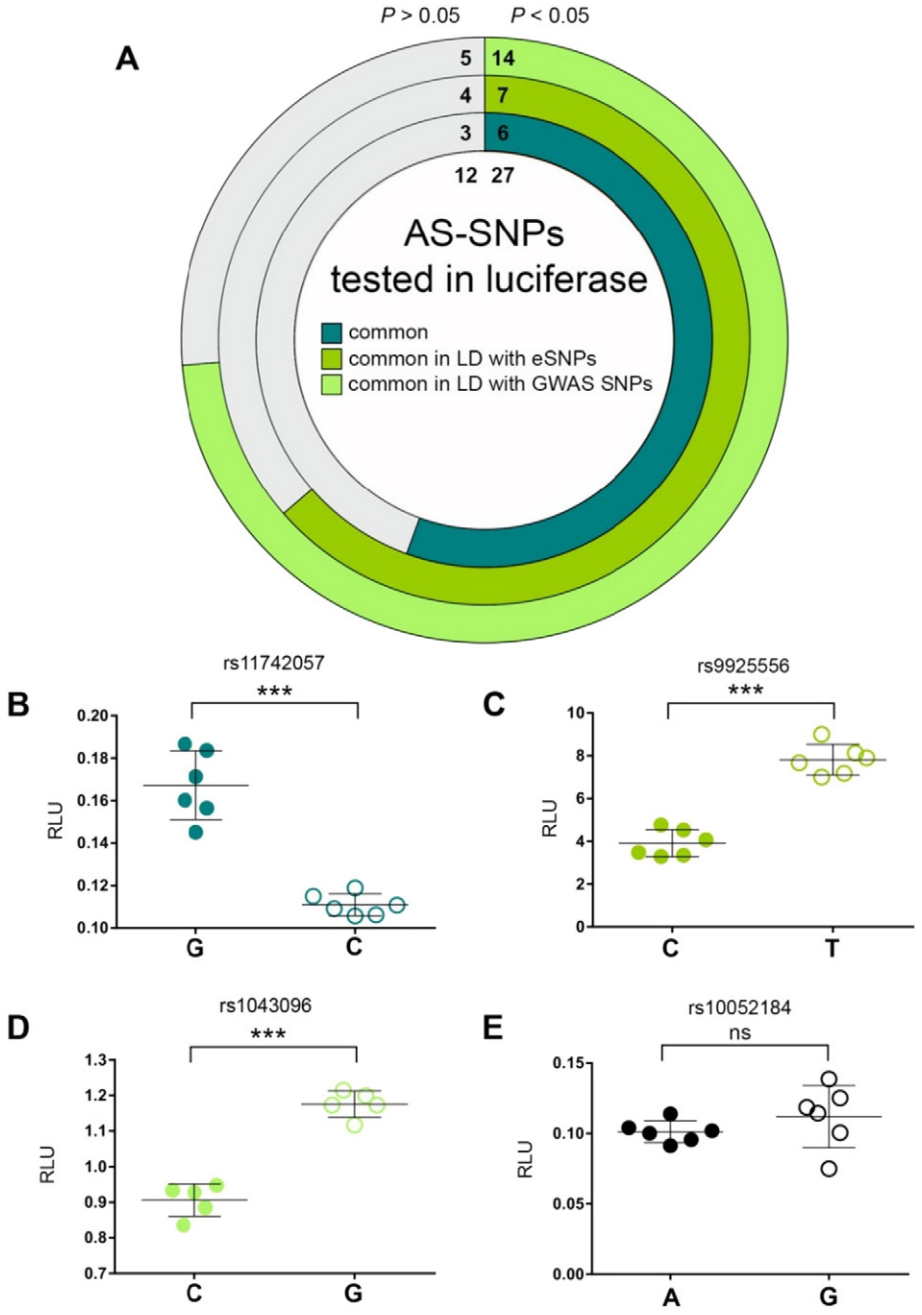

Fig. 2. Dual luciferase assays detecting a statistical significant difference in activity between the alleles. (A) Numbers of tested variants in 3 different categories of AS-SNPs. Examples are shown for common AS-SNPs (B) AS-SNPs in LD with eSNPs (C) or GWAS SNPs (D) and for non-AS-SNPs (E). Each dot represents a technical replica. Full scatter plots for all the 39 SNPs tested are present in Supplementary material Fig. S1–S5. *** = $P < 0.0001$; ** = $P < 0.001$; * = $P < 0.01$; ns = not significant.

3.3. AS-SNPs as suggested or verified drivers of GWAS signals

We found several AS-SNPs that may be causative in genetic associations. GWAS has found different SNPs related to levels of HDL-C which is a quantitative trait for which low levels are strongly associated to coronary artery disease. We found that the AS-SNP rs4846913 in the first intron of *GALNT2* is in LD ($1 > r^2 > 0.777$) in relation to several GWAS SNPs and therefore a good candidate to explain the associations. We verified the functional effect of rs4846913 and rs2144300 located in the same regulatory element using luciferase and EMSA [22]. In a parallel project Roman et al. thoroughly tested 25 candidate SNPs in luciferase assays identifying rs4846913, rs2144300 as well as rs2281721 located nearby as the functional variants mediating the allelic effect on GALNT2 [23]. Our approach allowed us to pinpoint the same regulatory element without the need to screen many SNPs before arriving at the functional one(s), emphasizing the precision and effectiveness of our strategy.

MERTK is a proto-oncogene and alleles of this gene are likely to predispose to Hepatitis C induced liver fibrosis which shows genetic association to rs4374383 [24] with 72 SNPs in high LD. Liver fibrosis predisposes to hepatocellular carcinoma suggesting that this pathway might contribute to both diseases. It is not reasonable to make functional tests of all 72 SNPs in the high-LD region. The AS-SNP rs6726639 is in the 8th intron and rs13394651 in intron 7 are good candidates to drive the associations which need to be validated in separate experiments.

ELOVL2 is a member of the elongase enzymes referred to as Elongation of very-long-chain fatty acids proteins (ELOVLs) that selectively acts on polyunsaturated fatty acids (PUFAs). Several recent studies

Table 1
AS-SNPs detected in HepG2 associated to liver-specific GWAS traits.

GWAS cell specific associated traits	Number of AS-SNPs in LD $r^2 > 0.8$ (0.5)	Number of AS-loci $r^2 > 0.8$ (0.5)	Number of reported loci*	% of reported loci with LD $r^2 > 0.8$ AS-SNPs
Blood metabolite levels	10 (16)	5 (9)	107	4,7
HDL cholesterol	6 (8)	3 (4)	78	3,8
Triglycerides	5 (9)	4 (5)	42	9,5
Primary biliary cirrhosis	4 (5)	2 (2)	24	8,3
Liver enzyme levels (alkaline phosphatase)	4 (4)	2 (2)	14	14,3
Metabolic syndrome	3 (4)	2 (2)	34	5,9
Fibrinogen	3 (4)	2 (2)	23	8,7
Blood metabolite ratios	2 (3)	2 (3)	47	4,3
Warfarin maintenance dose	2 (2)	1 (1)	6	16,7
Metabolite levels	2 (3)	2 (3)	209	1,0
Hepatitis C induced liver fibrosis	2 (3)	1 (1)	6	16,7
Cataracts in type 2 diabetes	2 (2)	1 (1)	2	50,0
Phospholipid levels (plasma)	2 (2)	1 (1)	35	2,9
Type 2 diabetes	1 (3)	1 (2)	133	0,8
Nonalcoholic fatty liver disease	1 (1)	1 (1)	38	2,6
Liver enzyme levels (gamma-glutamyl transferase)	1 (2)	1 (2)	24	4,2
Insulin-like growth factors	1 (1)	1 (1)	6	16,7
LDL cholesterol	1 (3)	1 (3)	59	1,7
Metabolic traits	0 (4)	0 (4)	94	0
Hepatitis B vaccine response	0 (1)	0 (1)	1	0
Drug-induced liver injury (amoxicillin-clavulanate)	0 (3)	0 (1)	2	0
Fasting insulin-related traits (interaction with BMI)	0 (1)	0 (1)	18	0
Liver enzyme levels	0 (1)	0 (1)	55	0
Homocysteine levels	0 (2)	0 (2)	24	0

* Loci defined as GWAS SNPs within 1 Mb regions.

have shown that ELOVL2 has an essential role in the synthesis of several fatty acid species in vivo and determines the lipid profile of PUFAs in the whole body. Independent GWAS studies have reported that variants located in the *ELOVL2* gene were associated with levels of PUFAs. Employing our AS-SNP pipeline, rs953413 and rs3798713 which show allele bias in TFs binding in ChIP-seq reads were identified as likely candidates to mediate the effect.

4. Discussion

GWAS has successfully identified thousands of variants associated to many diseases and traits. Despite large-scale efforts like the ENCODE and the Epigenome Roadmap projects, progress has been slow in identifying the exact SNP that is driving the effect. The most widely accepted approach in the search for regulatory variants relies on the

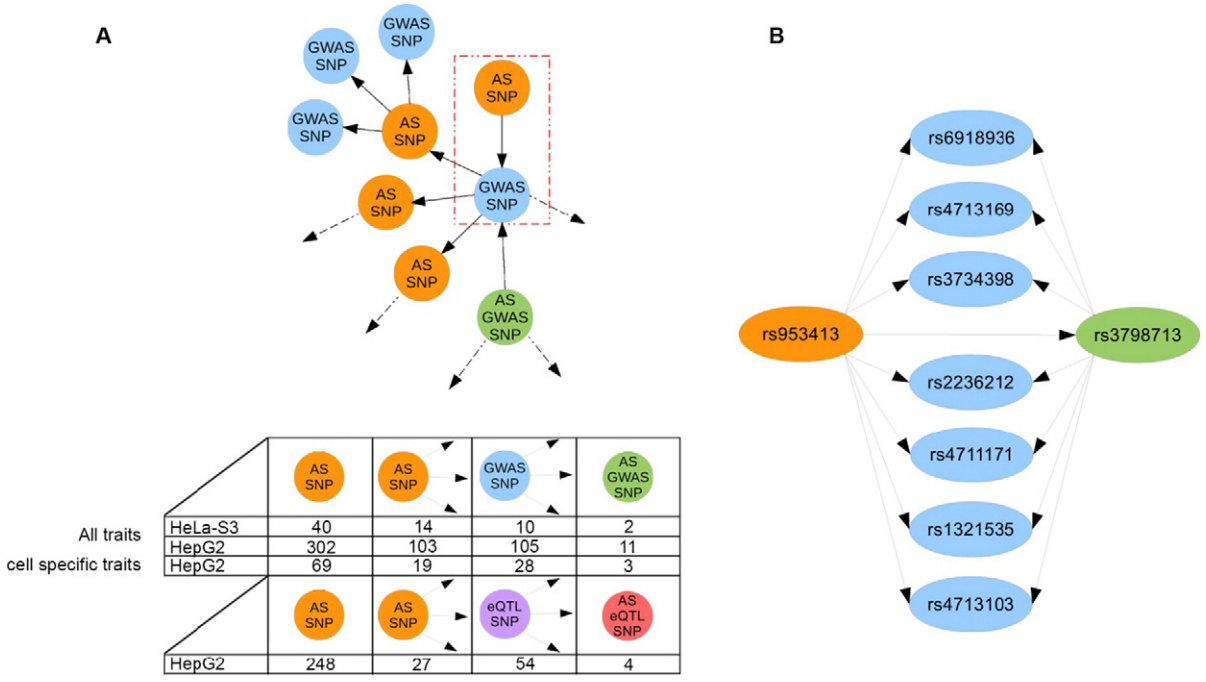

Fig. 3. AS-SNPs associated to GWAS SNPs. (A) Model representation of the networks of interactions observed between AS-SNPs and GWAS (or eQTL) SNPs. The dotted red box highlights the simplest scenario with one AS-SNP in LD with a GWAS SNP. The tables report the numbers of instances observed in each cell line where one AS-SNP is in LD with several GWAS (or eQTL) SNPs or one GWAS (or eQTL) SNPs is in LD with different AS-SNPs, or where GWAS (or eQTL) SNPs were also AS-SNPs. (B) Experimental example of interactions network. AS-SNP rs953413 identified in HepG2 (orange) is in LD with 8 different GWAS SNPs (cyan), one of which is an AS-SNP itself (green), associated to plasma lipid levels. The genomic location of the SNPs is presented in Fig. S6.

identification of SNPs altering a TF binding motif. Several databases, such as RegulomeDB [25] and HaploReg [26], have been built following this concept with SNPs defined as regulatory based on the accumulation of genomics evidences like ChIP-seq signals from TFs and histone modifications and overlap to specific TF binding motifs. The effect of a SNP on a TF motif represent the core of several bioinformatics approaches, among others sTRAP [27], which calculate the DNA binding affinity based on biophysical models or rSNP MAPPER [28], which score the effect of SNPs on TFBS on a large scale.

Several other pipelines (e.g. AlleleSeq [14], iASeq [29], ALEA [30]) have been developed to exploit RNA-seq and ChIP-seq datasets in order to gather information about allele specific expression (ASE) and binding (ASB).

In this work we followed an approach based on the principle that the information regarding the preferential binding of a TF to one allele is intrinsically "written" in the ChIP-seq reads of cell specific TFs. The ASB to an allele, measured in terms of ChIP-seq reads density at heterozygous positions, was used to isolate candidate regulatory SNPs regardless of their direct alteration of a TF binding motif.

We therefore searched systematically for such events in two cell lines and found thousands of candidate functional variants. Such variants located in GALNT2 have been characterized in detail by us [22] to show their functionality. The efficacy and quality of our genomic screening was further confirmed by an independent elegant study by Roman et al. where a labor intensive testing of 25 SNPs was necessary to identify the same regulatory variants. In another locus associated to Hepatitis C induced liver fibrosis there are 72 SNPs in high LD that potentially can drive the effect and at such loci a screening procedure clearly is needed to prioritize which SNPs that should be tested experimentally. We have found AS-SNPs in high LD to the GWAS SNPs that are likely drivers of the effect.

The liver is an important organ where several biochemical events unfold which can contribute to many diseases and traits that are central to the metabolism. Here we present a list of 3001 candidate functional SNPs that are likely to contribute to inter-individual variation. In the present study we investigate the allele-specific effect only at heterozygous positions which according to the Hardy-Weinberg law is 33% of variants that are common in the population. If ChIP-seq data was generated for additional cell lines or tissues a larger portion of common variation could be investigated for difference in TF binding.

The data we present here are important for several reasons. It provides a collection of candidate functional variants which can be investigated further for contribution to disease processes. The importance of developing new strategies to pinpoint functional variants is exemplified by variants in the *FTO* gene which are associated to body mass index [31]. Initially it was suspected that *FTO* itself was the culprit but Claussnitzer et al. [32] showed that in one regulatory element rs1421085 disrupts a motif for the repressor ARID5B leading to doubling of expression of *IRX3* and *IRX5* located 516 and 1164 kb away. Such information is crucial for a detailed understanding of disease processes. Since drugs acting on genetically supported targets are twice as likely to work compared to drugs acting on other targets this information will also aid pharmaceutical companies in developing new therapies.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

The authors thank the ENCODE project for generating ChIP-seq data. The genome sequence described/used in this research was derived from a HeLa cell line (http://www.ncbi.nlm.nih.gov/gap). Henrietta Lacks, and the HeLa cell line that was established from her tumor cells without her knowledge or consent in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her family members for their contributions to biomedical research. This study was reviewed by the NIH HeLa Genome Data Access Working Group. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b2010003 and b2011107. Genotyping and sequencing was performed at the SNP&Seq platform at SciLife Lab, Uppsala University. The gold nuggets image in Fig. 1 was originally posted to Flickr by jsj1771 at http://flickr.com/photos/47445767@N05/16884410748. It was reviewed on 3 May 2015 by the FlickreviewR robot and was confirmed to be licensed under the terms of the cc-by-2.0.

This work was supported by the Swedish Research Council (CW) (2010-3505), the Swedish Diabetes Foundation (CW) (2015-064), Diabetes Wellness Network Sweden (CW) (4445/2011SW), the Family Ernfors Fund (CW), Uppsala University (CW) and The Swedish Government's strategic research area EXODIAB (Excellence of Diabetes Research in Sweden) (CW).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ygeno.2016.04.006.

References

[1] M.R. Nelson, et al., The support of human genetic evidence for approved drug indications, Nat. Genet. 47 (8) (2015) 856–860.
[2] L. Hindorff, et al., A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 2014.
[3] K. Musunuru, et al., From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus, Nature 466 (7307) (2010) 714–719.
[4] Q. Huang, et al., A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding, Nat. Genet. 46 (2) (2014) 126–135.
[5] I.R. Cowper-Sal, et al., Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression, Nat. Genet. 44 (2012) 1191–1198.
[6] The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome, Nature 489 (7414) (2012) 57–74.
[7] The Genotype-Tissue Expression (GTEx) project, Nat. Genet. 45 (6) (2013) 580–585.
[8] The Roadmap Epigenomics Consortium, Integrative analysis of 111 reference human epigenomes, Nature 518 (7539) (2015) 317–330.
[9] S.S.P. Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 159(7): p. 1665–1680.
[10] Z. Tang, et al., CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 163(7): p. 1611–1627.
[11] P. Sahlén, et al., Genome-wide mapping of promoter-anchored interactions with close to single-enhancer resolution, Genome Biol. 16 (1) (2015) 156.
[12] A. Ameur, et al., Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP, Nucleic Acids Res. 37 (2009).
[13] M. Motallebipour, et al., Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq, Genome Biol. 10 (11) (2009) R129.
[14] J. Rozowsky, et al., AlleleSeq: analysis of allele-specific expression and binding in a network framework, Mol. Syst. Biol. 7 (2011) 522.
[15] E.E. Schadt, et al., Mapping the Genetic Architecture of Gene Expression in Human Liver, PLoS Biol. 6 (5) (2008), e107.
[16] S. Enroth, et al., Nucleosome regulatory dynamics in response to TGFβ, Nucleic Acids Res. (2014).
[17] M. Cavalli, et al., Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. Hum. Genet., 2016: p. 1–13.
[18] Y. Zhang, U. Werling, W. Edelmann, SLiCE: a novel bacterial cell extract-based DNA cloning method, Nucleic Acids Res. 40 (8) (2012), e55.
[19] A. Adey, et al., The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line, Nature 500 (7461) (2013) 207–211.
[20] O. Corradin, et al., Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits, Genome Res. 24 (2014) 1–13.
[21] N. Kumasaka, A. Knights, D. Gaffney, Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. bioRxiv, 2015.
[22] M. Cavalli, et al., Looking beyond GWAS: allele-specific transcription factor binding drives the association of GALNT2 to HDL-C plasma levels, Lipids Health Dis. 15 (1) (2016) 1–5.
[23] T.S. Roman, et al., Multiple hepatic regulatory variants at the GALNT2 GWAS locus associated with high-density lipoprotein cholesterol, Am. J. Hum. Genet. 97 (6) (2015) 801–815.

[24] E. Patin, et al., Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection, Gastroenterology 143 (2012) 1244–1252, e1-12.
[25] A.P. Boyle, et al., Annotation of functional variation in personal genomes using RegulomeDB, Genome Res. 22 (9) (2012) 1790–1797.
[26] 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 Research) 2011.
[27] T. Manke, M. Heinig, M. Vingron, Quantifying the effect of sequence variation on regulatory interactions, Hum. Mutat. 31 (4) (2010) 477–483.
[28] V.D. Marinescu, I.S. Kohane, A. Riva, MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes, BMC Bioinformatics 6 (1) (2005) 1–20.
[29] Y. Wei, et al., iASeq: integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets, BMC Genomics 13 (1) (2012) 1–19.
[30] H. Younesy, et al., ALEA: a toolbox for allele-specific epigenomics analysis, Bioinformatics 30 (8) (2014) 1172–1174.
[31] T.M. Frayling, et al., A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity, Science 316 (5826) (2007) 889–894.
[32] M. Claussnitzer, et al., FTO obesity variant circuitry and adipocyte browning in humans, N. Engl. J. Med. 373 (10) (2015) 895–907.