BMJ Open Diabetes Research & Care

Two novel type 2 diabetes loci revealed through integration of TCF7L2 DNA occupancy and SNP association data

Matthew E Johnson,[1] Jianhua Zhao,[1] Jonathan Schug,[2] Sandra Deliard,[1] Qianghua Xia,[1] Vanessa C Guy,[1] Jesus Sainz,[3] Klaus H Kaestner,[2] Andrew D Wells,[4] Struan F A Grant[1,2,5]

To cite: Johnson ME, Zhao J, Schug J, *et al*. Two novel type 2 diabetes loci revealed through integration of TCF7L2 DNA occupancy and SNP association data. *BMJ Open Diabetes Research and Care* 2014;**2**:e000052. doi:10.1136/bmjdrc-2014-000052

► Additional material is available. To view please visit the journal online (http://dx.doi.org/10.1136/bmjdrc-2014-000052).

MEJ and JZ contributed equally.

Received 5 August 2014
Revised 4 September 2014
Accepted 3 October 2014

For numbered affiliations see end of article.

Correspondence to
Dr Struan FA Grant;
grants@email.chop.edu

ABSTRACT

Background: The transcription factor 7-like 2 (*TCF7L2*) locus is strongly implicated in the pathogenesis of type 2 diabetes (T2D). We previously mapped the genomic regions bound by TCF7L2 using ChIP (chromatin immunoprecipitation)-seq in the colorectal carcinoma cell line, HCT116, revealing an unexpected highly significant over-representation of genome-wide association studies (GWAS) loci associated primarily with endocrine (in particular T2D) and cardiovascular traits.

Methods: In order to further explore if this observed phenomenon occurs in other cell lines, we carried out ChIP-seq in HepG2 cells and leveraged ENCODE data for five additional cell lines. Given that only a minority of the predicted genetic component to most complex traits has been identified to date, plus our GWAS-related observations with respect to TCF7L2 occupancy, we investigated if restricting association analyses to the genes yielded from this approach, in order to reduce the constraints of multiple testing, could reveal novel T2D loci.

Results: We found strong evidence for the continued enrichment of endocrine and cardiovascular GWAS categories, with additional support for cancer. When investigating all the known GWAS loci bound by TCF7L2 in the shortest gene list, derived from HCT116, the coronary artery disease-associated variant, rs46522 at the *UBE2Z-GIP-ATP5G1-SNF8* locus, yielded significant association with T2D within DIAGRAM. Furthermore, when we analyzed tag-SNPs (single nucleotide polymorphisms) in genes not previously implicated by GWAS but bound by TCF7L2 within 5 kb, we observed a significant association of rs4780476 within *CPPED1* in DIAGRAM.

Conclusions: ChIP-seq data generated with this GWAS-implicated transcription factor provided a biologically plausible method to limit multiple testing in the assessment of genome-wide genotyping data to uncover two novel T2D-associated loci.

Key messages

- Across multiple data sets derived from various cell lines, there is consistent evidence of a highly significant over-representation of genome-wide association study (GWAS)-implicated loci within the list of genes harboring a transcription factor 7-like 2 (TCF7L2) occupancy site.
- Given this TCF7L2 genome-wide occupancy behavior, we observe that through cross-referencing GWAS-derived statistics with specific ChIP (chromatin immunoprecipitation)-seq data, one can facilitate biologically plausible limitations to multiple testing and thus aid gene discovery efforts.
- We reveal *UBE2Z-GIP-ATP5G1-SNF8* and *CPPED1* as novel type 2 diabetes loci using this approach.

INTRODUCTION

The repertoire of genes already established to play a role in the pathogenesis of type 2 diabetes (T2D) has grown substantially as a consequence of results from recent genome-wide association studies (GWAS). One of the strongest T2D associations to date, based on risk conferred, is with variation within the transcription factor 7-like 2 (*TCF7L2*) gene.[1–6] Indeed, the common intronic variant at this locus is strongly associated with the disease in all major racial groups.[7]

In order to better understand the functional role of TCF7L2, we previously performed a ChIP (chromatin immunoprecipitation)-seq experiment with this transcription factor to elucidate its binding repertoire genome wide.[8 9] This approach performed well with the human colorectal carcinoma cell line, HCT116, where the TCF7L2 protein is abundantly expressed. Unexpectedly, and despite employing a carcinoma cell line, our initial data suggested that the gene list corresponding to TCF7L2 occupancy sites was strongly enriched for pathway categories related to metabolic-related functions and traits.

To our surprise, we also observed a highly significant over-representation of GWAS-implicated loci within the list of genes harboring a TCF7L2 occupancy site; indeed, the primary GWAS categories enriched were

for endocrine, in particular T2D, and cardiovascular traits. Our observations are supported by the recent report that classically defined transcription factors operating in the β-cell cluster around variants reported in GWAS.[10]

As only a minority of the predicted genetic component to most complex traits has been identified to date, termed the 'missing heritability',[11] there is potential for using knowledge of TCF7L2 occupancy to aid further gene discovery for T2D. The rationale behind this is that if one restricted association analyses to just the genes occupied by TCF7L2, one could limit the extent of correction for multiple testing that typically blights GWAS analyses.

In order to elucidate this possibility, we first elected to expand on our initial findings to investigate if this intriguing pattern holds across multiple cell lines, using the algorithm HOMER (Hypergeometric Optimization of Motif EnRichment); indeed, we have already reported our use of this program when analyzing ChIP-seq data for other GWAS-implicated transcription factors, namely MEF2C[12] and FOXA2.[13] To that end, we meshed our in-house-derived ChIP-seq data sets, both for HCT116 and HepG2, with those made available by the ENCODE project.[14] In addition to analyzing these data sets separately to further investigate possible GWAS locus enrichment, we postulated that many novel genes on the TCF7L2 target list could be relevant to T2D; as such, we analyzed the shortest ChIP-seq-derived gene list, generated in HCT116, in the context of GWAS data itself to investigate if novel T2D loci could be revealed when restricting testing to just the loci derived from this approach.

METHODS

Cell culture and reagent

The HepG2 hepatocarcinoma cell line was purchased from the American Type Cell Center (ATCC, Manassas, Virginia, USA). Cells were cultured at 37°C, 95% humidity, and supplied with 5% CO_2 in ATCC-formulated Eagle's Minimum Essential Medium supplemented with 10% fetal bovine serum (Sigma, St. Louis, Missouri, USA), 2 mM L-glutamine (Gibco Invitrogen, Carlsbad, California, USA), 100 units/mL penicillin/100 µg/mL streptomycin (Cellgro, Manassas, Virginia, USA). On the basis of previous papers outlining *TCF7L2* isoforms,[15–17] we chose from antibodies that were raised to antigen at the most constant region among TCF7L2 isoforms, that is, the amino acids encoded by exons 1–3 (Cat.05-511; Millipore, Billerica, Massachusetts, USA) as described previously.[8]

Chromatin Immunoprecipitation

ChIP was performed in triplicate following the instructions provided by the suppliers of the EZ-ChIP kit (Cat.17-371; Millipore, Billerica, Massachusetts, USA) and as described previously.[18] Cells were sonicated on ice for 12 cycles of 15 s on and 45 s off at setting 3 (2100XL ultrasonic liquid processors, Misonix, Farmingdale, New York, USA). Sonicated chromatin was primarily in the 100–500 bp range, averaging 200–300 bp.

After overlaying all reads from two independent experiments for HepG2, a total of 3810 binding sites were observed at a false discovery rate of 1%, cumulative Poisson p value of 0.0001, and fold coverage threshold of four times normalized sequence tags in the target experiment comparable with random background sequence tags using the HOMER[19] analysis package. The TCF7L2 ChIP signal was clearly distinct from the pseudo-ChIP signal as identified by GLITR[18] (online supplementary figure S1). In addition, we chose 17 sites with a variable binding score for validation purposes by real-time PCR, all of which showed clear evidence of enrichment (online supplementary figure S2 and table S1).

Sequencing

The sequencing library was prepared as per Illumina's instructions (http://www.illumina.com, San Diego, California, USA). Sequencing on the Illumina Genome Analyzer and subsequent analyses were performed at the Functional Genomics Core at the University of Pennsylvania.

DNA libraries were assessed for size, purity, and quantity using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, California, USA), followed by sequencing using an Illumina GA-II according to the manufacturer's instructions, and have been described previously.[8] The ENCODE ChIP-seq ((HepG2, HeLa-S3 (×2), HEK293, MCF7, and PANC1) and input raw sequence files were downloaded from the UCSC database (http://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html). HOMER[19] was utilized to determine TCF7L2 binding sites and their association with RefSeq transcripts aligned to hg19 was downloaded from UCSC via HOMER. The candidate target gene was the closest gene regardless of the direction from the binding site. In all cases, the transcription start site of the aligned transcript was used as the anchor point for distance measurements.

TCF7L2 ChIP-seq in ENCODE

We processed the ENCODE TCF7L2 ChIP-seq data for the HepG2, HeLa-S3, HEK293, MCF7, and PANC1 cell lines[14] and also reanalyzed our HCT116 data,[8] using HOMER.

We observed a wide range in the occupancy site number, location, and nearest unique gene. HCT116 was found to harbor the lowest number of binding sites (n=865) and corresponding genes (n=750). ENCODE's HeLa-S3 (exons 1–3) revealed the highest number of binding sites (n=11 817) and corresponding genes (n=6451; online supplementary table S2).

We elected to reanalyze all of ENCODE's and our own generated TCF7L2 ChIP-seq data with the same HOMER peak parameters described above to eliminate any threshold effects that would be caused by comparing the occupancy sites between different peak finding programs and threshold parameter settings. The number of

placed sequence reads for all eight ChIP-seq experiments varied from a low of 18 139 284 to a high of 57 938 025 (online supplementary table S3).

Pathway analysis

Data were analyzed through the use of Ingenuity Pathways Analysis (Ingenuity Systems, http://www.ingenuity.com, Redwood City, California, USA) specified for 'Human'. The genes that corresponded to at least one function or pathway annotation in the Ingenuity Knowledge Base were eligible for the analysis. The p value associated with functions and pathways was calculated using the right-tailed Fisher exact test.

GWAS category analyses

We based our analysis on all GWAS genes summarized in a freeze of the National Human Genome Research Institute (NHGRI) GWAS catalog (http://www.genome.gov/gwastudies) from 19 February 2013. Enrichment was investigated using a χ^2 analysis. Our method of scoring overlapping GWAS-implicated genes detected in the ChIP-seq data was to assign 1 point to a GWAS region where all the genes in the region were found in our list and a fraction of a point determine by how many genes where found in our gene list divide by the total genes in the GWAS region. For instance, this analysis model would give a GWAS region with 1 gene the same weight as a region harboring 8 genes.

Association analyses

We derived the list of genes bound by TCF7L2 within 5 kb of the transcription start site in the HCT116 cell line, as it yielded the smallest number of binding sites and thus the smallest list of corresponding genes. First, we derived the list of single nucleotide polymorphisms (SNPs) at GWAS-implicated loci on this gene list (n=40). Furthermore, we aimed to look at the remainder of the gene list, the members of which had not been previously implicated by GWAS, and in order to minimize multiple testing we elucidated which tag-SNPs represented on the basic Illumina Human Hap 550 BeadChip resided with our genes of interest (n=892). We then separately queried both lists against the publicly available GWAS meta-analysis data set generated by DIAGRAM (http://diagram-consortium.org/downloads.html)[20] to determine if any variants would yield a p value lower than the Bonferroni-corrected p value for the respective test, where the threshold for significance for the previously reported GWAS loci test was set at 1.25×10^{-3} and for the non-GWAS-implicated loci test it was set at 5.61×10^{-5}.

RESULTS

ChIP-seq data appraisal

To extend our previous genomic occupancy analyses for TCF7L2 in HCT116 cells,[8] we performed ChIP-seq in the human heptocarcinoma cell line, HepG2, to map DNA sequences bound by TCF7L2. Utilizing HOMER, the distribution of the binding sites was 1555 intronic, 1920 intergenic, and the remaining 335 in various other genic regions (online supplementary figure S3). We also processed similarly with HOMER TCF7L2 ChIP-seq data for the HepG2, HeLa-S3, HEK293, MCF7, and PANC1 cell lines from ENCODE,[14] plus our previously generated HCT116-derived data.[8]

We went on to employ the de novo motif discovery algorithm, also within HOMER, to derive the consensus binding site for these other seven ChIP-seq data sets compared with the consensus motif derived from our HCT116 ChIP-seq data and from previous work by others.[21] A similar 12 bp consensus was found in 24–53% of all binding sites (online supplementary figure S4). The majority of occupancy (>93%) fell within 5–500 kb of a RefSeq gene transcription start site in the remaining seven ChIP-seq data sets analyzed (online supplementary table S4).

We went on to perform pathway analyses for each of the eight ChIP-seq-derived gene sets. In HCT116, we observed pathways related to 'Factors Promoting Cardiogenesis in Vertebrates', 'Type II Diabetes Mellitus Signaling', and 'NF-κB Activation by Viruses,' respectively, making them the most significant annotations and readily surviving correction for multiple comparisons (see all categories that achieved a nominal p<0.05 in online supplementary table S5).We also observed that HeLaS3 (exons 1–3), HeLaS3 (exons 4–16), MCF7, and PANC1 yielded significant enrichment, following adjustment for multiple comparisons (uncorrected p value: 4.47×10^{-6}, 9.12×10^{-6}, 5.62×10^{-4}, and 9.33×10^{-4}) for genes in the 'Type II Diabetes Mellitus Signaling' category from the top 20 canonical pathway analyses (see all categories that achieved a nominal p<0.05 in online supplementary tables S6–S9). We observed consistent under-representation of members of the β-cell-related pathway in the 'Type II Diabetes Mellitus Signaling' category and over-representation of binding in other tissues within the same category across these data sets (see figure 1 for representative image derived from HCT116 in-house data).

Three of the data sets, two derived from the liver and one from the kidney, that is, two HepG2 and one HEK293, did not yield a significant enrichment of genes in the 'Type II Diabetes Mellitus Signaling' category from the canonical pathway analyses (see all categories that reached a nominal p<0.05 in online supplementary tables S10–S12).

In addition, our pathway analysis also determined consistent and significant enrichment of genes in the 'Wnt/β-catenin Signaling', 'Molecular Mechanisms of Cancer', and 'Factors Promoting Cardiogenesis in Vertebrates' categories from the top 20 canonical pathway analyses in all eight of the cell lines (see all categories that achieve an adjusted p<0.05 in online supplementary tables S5–S12).

GWAS category enrichment

Given that original HCT116 study suggested TCF7L2 occupancy was found more often at GWAS loci than

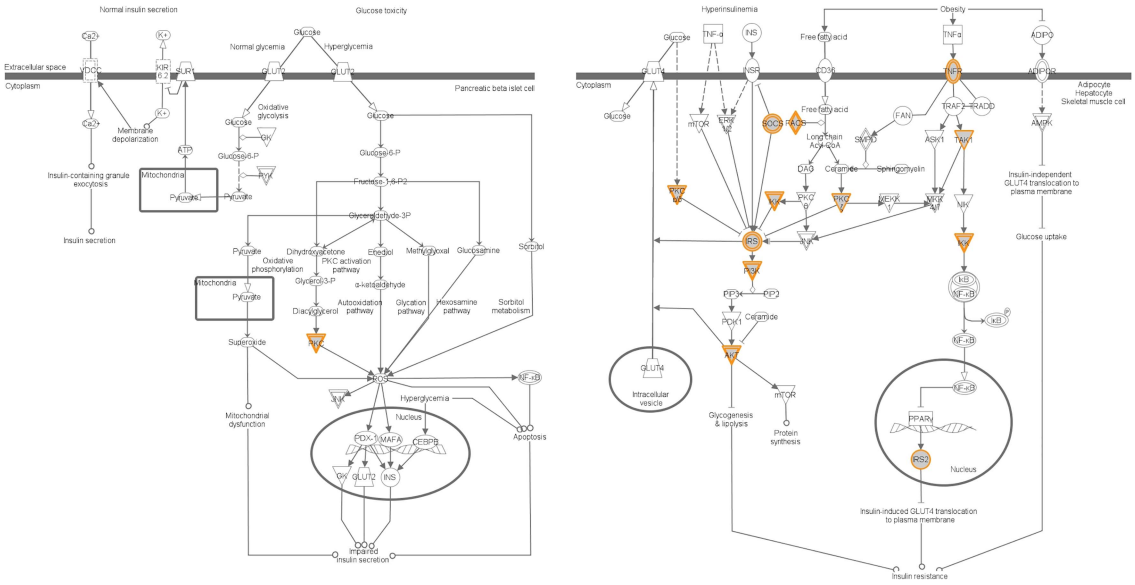

Figure 1 TCF7L2 ChIP-seq in HCT116 cells. The orange color genes in the 'Type II Diabetes Mellitus Signaling' pathway represents TCF7L2 binding sites candidate target genes that were the closest gene transcription start site to TCF7L2 binding sites regardless of the direction from the binding site. Data were analyzed through the use of Ingenuity Pathways Analysis (Ingenuity Systems, http://www.ingenuity.com, Redwood) City, California, USA) specified for 'Human'. The genes that contain at least one function or pathway annotation in the Ingenuity Knowledge Base were eligible for the analysis. ChIP, chromatin immunoprecipitation; TCF7L2, transcription factor 7-like 2; VDCC, voltage-dependent calcium channels; JNK, Jun N-terminal kinases; ROS, reactive oxygen species; PKC, protein kinase C; NF-κB, nuclear factor-κB; SUR, sulfonylurea receptor; TNFα, tumor necrosis factor α; INS, insulin; INSR, INS receptor; TRAF2, TNF receptor-associated factor 2; mTOR, mechanistic target of rapamycin; SMPD, sphingomyelin phosphodiesterases; DAG, diacylglycerol; PPAR, peroxisome proliferator-activated receptors; GK, glucokinase; PYK, pyruvate kinase; GLUT, glucose transporter; ADIPOR, adiponectin receptor.

expected by chance,[8] as did our work with MEF2C[12] and FOXA2,[13] we were motivated to query the results for each data set in turn against all GWAS signals reported, as derived from the NHGRI GWAS catalog.

Of the 2435 nearest genes with a TCF7L2 binding site in our in-house HepG2-generated data set, representing 12.8% of all RefSeq genes used in the overall analysis (n=19 015), there was a highly significant over-representation of loci implicated in disease susceptibility by GWAS (629.71 of 3607 (17.5%) loci; $p=1.09\times10^{-10}$; table 1). This observation was primarily driven by excess loci revealed from GWAS of endocrine ($p=6.18\times10^{-10}$), cardiovascular ($p=9.30\times10^{-10}$), and cancer ($p=1.04\times10^{-6}$) traits; specifically, T2D also showed an enrichment ($p=8.80\times10^{-3}$). In contrast, we observed only marginal or no significant enrichment of GWAS signals for neurological or inflammation-related traits.

Expanding our GWAS signal analyses to the data derived from the other six cell lines (HeLaS3 (exons 1–3), HeLaS3 (exons 4–16), MCF7, PANC1, HCT116, HEK293, and HepG2 (ENCODE)), we determined that there was also highly significant over-representation of GWAS loci for TCF7L2 targets in all seven cell lines (table 1). As seen in HepG2 (in-house), this observation was again primarily driven by excess loci from GWAS of endocrine, cardiovascular, and cancer traits. As demonstrated in HepG2 (in-house), T2D GWAS-implicated loci specifically are also generally enriched, although not statistically significant due to the relatively small list of GWAS-implicated T2D genes being queried. Neurological and inflammation-related GWAS signals were largely consistently shown to have marginal or no enrichment in the seven ChIP-seq data sets, comparable with what was seen in HepG2 (in-house; table 1).

To contrast with control data sets, we also generated a random list of 5000 genes from the 19 015 RefSeq genes used by HOMER to determine the nearest gene lists described above to ascertain if there was a bias of our data analysis. The randomly generated gene list showed no significant over-representation of GWAS genes in the random gene set; in fact, it showed a trend of under-representation of GWAS genes in the random HOMER gene set, primarily due to the fact that some gene names in the NHGRI GWAS catalog are not RefSeq annotations (online supplementary table S13).

Cross comparisons with genome-wide meta-analysis summary data

Given that only a minority of the predicted genetic component to most complex traits has been identified to date, plus the fact that our GWAS-implicated transcription factor of interest shows consistent statistically significant preferential binding to loci associated with complex traits, we investigated if restricting association analyses to just the genes uncovered from our ChIP-seq approach in order to reduce multiple testing could yield novel loci associated with T2D. When investigating all the known GWAS loci bound within 5 kb by TCF7L2 (most likely to

Table 1 Enrichment of GWAS signals for the nearest RefSeq genes to the TCF7L2 binding site in all eight cell lines

	Percentage of total hg19 gene list	Percentage of ChIP-seq gene list	p Values: χ^2	Percentage of total hg19 gene list	Percentage of ChIP-seq gene list	p Values: χ^2
	HCT116: 750 genes			HEPG2 (ENCODE): 1924 genes		
Endocrine	3.9 (750/19 015)	7.3 (64.66/888)	2.69×10^{-6}	10.1 (1924/19 015)	19.7 (175.34/888)	3.38×10^{-15}
T2D	3.9 (750/19 015)	11.0 (9/82)	0.0025	10.1 (1924/19 015)	21.7 (17.83/82)	0.0024
Cancer	3.9 (750/19 015)	10.7 (35.87/335)	5.81×10^{-9}	10.1 (1924/19 015)	22.0 (73.86/335)	9.17×10^{-10}
Cardiovascular	3.9 (750/19 015)	7.8 (36.19/463)	9.33×10^{-5}	10.1 (1924/19 015)	21.3 (98.41/463)	3.34×10^{-11}
Inflammation	3.9 (750/19 015)	7.2 (37.32/521)	0.00 062	10.1 (1924/19 015)	16.3 (85.1/521)	5.40×10^{-5}
Neuropsychiatric	3.9 (750/19 015)	5.7 (33/584)	0.048	10.1 (1924/19 015)	13.5 (78.67/584)	0.023
All	3.9 (750/19 015)	5.9 (212.25/3607)	5.06×10^{-7}	10.1 (1924/19 015)	15.0 (541.21/3607)	2.58×10^{-14}
	HeLa exons (4–16): 1983 genes			HEPG2 (in-house): 2435 genes		
Endocrine	10.4 (1983/19 015)	16.8 (148.86/888)	1.64×10^{-7}	12.8 (2435/19 015)	21.3 (188.73/888)	6.18×10^{-10}
T2D	10.4 (1983/19 015)	17.5 (14.33/82)	0.086	12.8 (2435/19 015)	23.8 (19.5/82)	0.0088
Cancer	10.4 (1983/19 015)	19.8 (62.21/335)	3.41×10^{-5}	12.8 (2435/19 015)	23.7 (79.29/335)	1.04×10^{-6}
Cardiovascular	10.4 (1983/19 015)	20.0 (92.50/463)	9.18×10^{-9}	12.8 (2435/19 015)	24.4 (113.02/463)	9.30×10^{-10}
Inflammation	10.4 (1983/19 015)	14.0 (73.13/521)	0.02	12.8 (2435/19 015)	15.4 (80.43/521)	0.14
Neuropsychiatric	10.4 (1983/19 015)	12.5 (72.94/584)	0.15	12.8 (2435/19 015)	18.2 (106/584)	0.0011
All	10.4 (1983/19 015)	14.0 (506.47/3607)	2.02×10^{-8}	12.8 (2435/19 015)	17.5 (629.71/3607)	1.09×10^{-10}
	HEK293: 3519 genes			MCF7: 3863 genes		
Endocrine	18.5 (3519/19 015)	32.5 (288.23/888)	6.46×10^{-16}	20.3 (3863/19 015)	31.9 (283.57/888)	8.36×10^{-11}
T2D	18.5 (3519/19 015)	37.2 (30.5/82)	0.00 057	20.3 (3863/19 015)	36.0 (29.5/82)	0.0053
Cancer	18.5 (3519/19 015)	32.0 (107.19/335)	9.26×10^{-7}	20.3 (3863/19 015)	32.0 (107.09/335)	4.97×10^{-5}
Cardiovascular	18.5 (3519/19 015)	36.6 (169.37/463)	4.74×10^{-14}	20.3 (3863/19 015)	35.6 (164.92/463)	7.28×10^{-10}
Inflammation	18.5 (3519/19 015)	25.9 (134.85/521)	0.00 058	20.3 (3863/19 015)	28.7 (149.47/521)	0.0002
Neuropsychiatric	18.5 (3519/19 015)	31.2 (182.4/584)	1.33×10^{-9}	20.3 (3863/19 015)	25.2 (147/584)	0.022
All	18.5 (3519/19 015)	27.4 (988.25/3607)	1.54×10^{-22}	20.3 (3863/19 015)	28.3 (1019.56/3607)	5.47×10^{-17}
	PANC1: 5123 genes			HeLa exons (1–3): 6451 genes		
Endocrine	26.9 (5123/19 015)	34.5 (306.16/888)	0.00 029	33.9 (6451/19 015)	42.0 (372.76/888)	0.00 073
T2D	26.9 (5123/19 015)	44.5 (36.5/82)	0.0087	33.9 (6451/19 015)	47.4 (38.83/82)	0.082
Cancer	26.9 (5123/19 015)	36.5 (122.21/335)	0.0047	33.9 (6451/19 015)	47.9 (160.32/335)	0.00 041
Cardiovascular	26.9 (5123/19 015)	35.1 (162.48/463)	0.0037	33.9 (6451/19 015)	49.9 (230.9/463)	2.13×10^{-6}
Inflammation	26.9 (5123/19 015)	31.1 (161.98/521)	0.12	33.9 (6451/19 015)	38.8 (202.03/521)	0.11
Neuropsychiatric	26.9 (5123/19 015)	29.0 (169.5/584)	0.38	33.9 (6451/19 015)	37.7 (220.2/584)	0.19
All	26.9 (5123/19 015)	30.3 (1094.66/3607)	0.0016	33.9 (6451/19 015)	38.0 (1369.44/3607)	0.0013

We based our analysis on all GWAS genes summarized in the NHGRI GWAS catalog (http://www.genome.gov/gwastudies) from 19 February 2013. Enrichment was investigated using a χ^2 analysis. Our method of scoring the GWAS ChIP-seq gene overlap was to assign 1 point to a GWAS region where all the genes in the region were found in our list, and a fraction of a point determined by how many genes were found in our gene list divided by the total genes in the GWAS region. This analysis model would equally weight a GWAS region with 1 gene the same as a region with 8 genes as a single region.
ChIP, chromatin immunoprecipitation; GWAS, genome-wide association studies; T2D , type 2 diabetes; TCF7L2, transcription factor 7-like 2.

be functional) in the shortest gene list in order to minimize multiple testing, derived from HCT116, apart from the known *TCF7L2* locus itself (rs7901695), the coronary artery disease-associated variant, the T allele of rs46522 within the *UBE2Z-GIP-ATP5G1-SNF8* locus, yielded significant and novel DIAGRAM-derived association with T2D risk (OR=1.07; p=3.20×10^{-4}) (table 2); indeed, the occupancy site was ~4 kb from the transcription start site for *GIP* in an intergenic region known to be a hub for binding proteins, H3K27Ac histone marks and open chromatin via a DNase I hypersensitive site. Furthermore, when we analyzed Illumina Human Hap 550 tag-SNPs within genes not previously implicated by GWAS but bound within 5 kb by TCF7L2 in HCT116, again due to the fact that it was the shortest gene list, we observed significant association within the DIAGRAM data set of the A allele of rs4780476 within the gene encoding calcineurin-like phosphoesterase domain-containing protein 1 (*CPPED1*) with T2D risk (OR=1.1, p=4.10×10^{-5}; table 2). Furthermore, the TCF7L2 occupancy site was in the immediate *CPPED1* promoter region.

DISCUSSION

Given that only a minority of the predicted genetic component to most complex traits has been identified to date, plus the fact that this GWAS-implicated transcription factor shows preferential binding to genes genetically associated with complex traits, we investigated if restricting association analyses to the genes yielded from our ChIP-seq approach in order to reduce multiple testing could yield novel loci associated with T2D. Indeed, we found that of the known GWAS loci for any trait bound by TCF7L2 within 5 kb in HCT116, the coronary artery disease-associated variant, rs46522, within the *UBE2Z-GIP-ATP5G1-SNF8* locus[22] yielded association that survived correction for multiple testing. Interestingly, rs46522 is in strong LD with two potential functional variants in the biologically plausible gene encoding gastric inhibitory polypeptide (*GIP*): p.Ser103Gly (rs2291725) and variant influencing the splice site of intron 3 (rs2291726) leading to a truncated transcript.[22] This is particularly notable as this observation implicates a variant playing a role in T2D after being found originally in another GWAS category, i.e. cardiovascular. It has long been thought that TCF7L2 may confer its T2D effect via incretins,[1] of which GIP is one, thus furthering the case for this line of investigation; indeed, the locus encoding the receptor for GIP (*GIPR*) has already been reported in relevant GWAS settings to be associated with body mass index[23–25] and to influence the glucose and insulin responses to an oral glucose challenge.[26]

Furthermore, when considering the non-GWAS-implicated loci bound by TCF7L2 within 5 kb in HCT116, we observed significant association with rs4780476 within *CPPED1*. This is an equally interesting observation, as only two papers have been published to date on this gene product, with one showing that

Table 2 DIAGRAM-derived association results for type 2 diabetes with respect to the loci on interest

SNP	CHROMOSOME	POSITION	RISK_ALLELE	OTHER_ALLELE	P_VALUE	OR	OR_95L	OR_95U	N_CASES	N_CONTROLS	Gene
Variants previously published for GWAS of any trait											
rs7901695	10	114 744 078	C	T	2.50E-65	1.37	1.32	1.42	12 171	56 862	*TCF7L2* *
rs46522	17	44 343 596	T	C	3.20E-04	1.07	1.03	1.11	9580	53 810	*UBE2Z-GIP-ATP5G1-SNF8*
Variants not previously reported by GWAS of type 2 diabetes											
SNP	CHROMOSOME	POSITION	RISK_ALLELE	OTHER_ALLELE	P_VALUE	OR	OR_95L	OR_95U	N_CASES	N_CONTROLS	Gene
rs4780476	16	12 769 508	A	C	4.10E-05	1.1	1.05	1.15	6634	49 797	*CPPED1*

These loci survived correction for multiple testing based on the constraints derived from working with the HCT116 ChIP-seq 5 kb gene list and integrating with tag-SNPs coinciding with genes on that list.
*Locus previously reported to be associated with type 2 diabetes.
ChIP, chromatin immunoprecipitation; SNP, single nucleotide polymorphism.

downregulation of *CPPED1* expression improves glucose metabolism in vitro in adipocytes[27] and another implicating it in syndromic obesity using array comparative genomic hybridization.[28]

The challenge of the increasing level of genetic data being generated in population-based cohorts, such as imputed genome-wide genotypes, exome and whole-genome data, is how one can derive true positive signals from the large amount of data, where the required stringent corrections for multiple testing at the genome level can easily miss true signals. Indeed, there have been great efforts to rationalize restricted testing to plausible regions of the genome to address a particular complex trait, most typically by leveraging previously reported linkage signals. However, linkage regions are often broad in terms of genomic regions covered and are therefore fraught with imprecision. Our limitation of multiple testing is based on biological plausibility, where a GWAS-implicated transcription factor is clearly pointing us to genes that are genetically associated with complex disease more often than expected by chance and thus may also be pointing us to novel genes where their strength of the association was at the level of noise at the genome-wide scale.

We carried out the meshing of GWAS-derived data with a ChIP-seq-derived gene list for a GWAS-implicated transcription factor in one of our cell lines. The rationale was that as all cell lines exhibited the same GWAS category enrichment characteristic, we would aim to carry out this investigation by narrowing the field as much as possible. As such, we elected to only leverage the gene names derived from the cell line that yielded the shortest gene list, namely HCT116. We also added the extra constraint of the site being within 5 kb of the nearest gene, as this made them the most biologically plausible, and thus limiting our testing further. We also limited our testing by only considering tag-SNPs used on a conventional genotyping array. Of course, we recognize that these cut-offs are completely arbitrary and that further testing with additional genotype and phenotype (we only considered T2D due to the obvious *TCF7L2* connection) data sets should be the subject of subsequent studies to refine this data-mining approach.

In conclusion, our study has further characterized loci bound by TCF7L2, which has in turn reinforced our previous observation that TCF7L2 has a statistically significant preference to occupy loci previously implicated by GWAS. By cross-referencing the loci at these occupancy sites with GWAS results in order to restrict correction for multiple testing, *UBE2Z-GIP-ATP5G1-SNF8* and *CPPED1* have been uncovered as T2D-associated loci. This approach has potential utility for the discovery process of novel therapeutic targets for diabetes and related traits in the future.

Author affiliations
[1]Division of Human Genetics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA
[2]Department of Genetics and Institute of Diabetes, Obesity and Metabolism, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Institute of Biomedicine and Biotechnology of Cantabria (IBBTEC), Spanish National Research Council (CSIC), Santander, Spain
[4]Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA
[5]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Contributors JZ, JS, SD, QX, VCG, KHK, ADW, and SFAG generated the data; MEJ, JS, KHK, ADW, and SFAG analyzed the data; and MEJ, JZ, SD, QX, VCG, JS, KHK, ADW, and SFAG reviewed the manuscript and contributed to the writing.

Funding This work was supported by NIH grant number P30 DK 19525. This work was supported by Institutional Development Funds and the Ethel Brown Foerderer Fund for Excellence from the Children's Hospital of Philadelphia.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All sequence reads from the ChIP-seq data sets generated at the Children's Hospital of Philadelphia and the University of Pennsylvania are available on request.

REFERENCES

1. Grant SF, Thorleifsson G, Reynisdottir I, *et al*. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 2006;38:320–3.
2. Sladek R, Rocheleau G, Rung J, *et al*. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007;445:881–5.
3. Saxena R, Voight BF, Lyssenko V, *et al*. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007;316:1331–6.
4. Scott LJ, Mohlke KL, Bonnycastle LL, *et al*. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007;316:1341–5.
5. Zeggini E, Weedon MN, Lindgren CM, *et al*. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007;316:1336–41.
6. Voight BF, Scott LJ, Steinthorsdottir V, *et al*. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010;42:579–89.
7. Cauchi S, El Achhab Y, Choquet H, *et al*. TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *J Mol Med* 2007;85:777–82.
8. Zhao J, Schug J, Li M, *et al*. Disease-associated loci are significantly over-represented among genes bound by transcription factor 7-like 2 (TCF7L2) in vivo. *Diabetologia* 2010;53:2340–6.
9. Deliard S, Zhao J, Xia Q, *et al*. Generation of high quality chromatin immunoprecipitation DNA template for high-throughput sequencing (ChIP-seq). *J Vis Exp* 2013(74):e50286.
10. Pasquali L, Gaulton KJ, Rodriguez-Segui SA, *et al*. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* 2014;46:136–43.
11. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53.
12. Johnson ME, Deliard S, Zhu F, *et al*. A ChIP-seq-defined genome-wide map of MEF2C binding reveals inflammatory pathways associated with its role in bone density determination. *Calcif Tissue Int* 2014;94:396–402.
13. Johnson ME, Schug J, Wells AD, *et al*. Genome-wide analyses of ChIP-seq derived FOXA2 DNA occupancy in liver points to genetic networks underpinning multiple complex traits. *J Clin Endocrinol Metab* 2014;99:E1580–5.
14. Consortium EP, Dunham I, Kundaje A, *et al*. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74.

15. Mondal AK, Das SK, Baldini G, *et al*. Genotype and tissue-specific effects on alternative splicing of the transcription factor 7-like 2 gene in humans. *J Clin Endocrinol Metab* 2010;95:1450–7.
16. Osmark P, Hansson O, Jonsson A, *et al*. Unique splicing pattern of the TCF7L2 gene in human pancreatic islets. *Diabetologia* 2009;52:850–4.
17. Prokunina-Olsson L, Welch C, Hansson O, *et al*. Tissue-specific alternative splicing of TCF7L2. *Hum Mol Genet* 2009;18:3795–804.
18. Tuteja G, White P, Schug J, *et al*. Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res* 2009;37:e113.
19. Heinz S, Benner C, Spann N, *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 2010;38:576–89.
20. Morris AP, Voight BF, Teslovich TM, *et al*. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 2012;44:981–90.
21. Hatzis P, van der Flier LG, van Driel MA, *et al*. Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol* 2008;28:2732–44.
22. Li Z, Gadue P, Chen K, *et al*. Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* 2012;151:1608–16.
23. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* 2011;9:e1001046.
24. Weedon MN, Ellard S, Prindle MJ, *et al*. An in-frame deletion at the polymerase active site of POLD1 causes a multisystem disorder with lipodystrophy. *Nat Genet* 2013;45:947–50.
25. Speliotes EK, Willer CJ, Berndt SI, *et al*. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 2010;42:937–48.
26. Saxena R, Hivert MF, Langenberg C, *et al*. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. *Nat Genet* 2010;42:142–8.
27. Manning AK, Hivert MF, Scott RA, *et al*. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 2012;44:659–69.
28. Vuillaume ML, Naudion S, Banneau G, *et al*. New candidate loci identified by array-CGH in a cohort of 100 children presenting with syndromic obesity. *Am J Med Genet A* 2014;164A:1965–75.