Disease-associated loci are significantly over-represented among genes bound by transcription factor 7-like 2 (TCF7L2) in vivo

J. Zhao · J. Schug · M. Li · K. H. Kaestner ·
S. F. A. Grant

Received: 28 March 2010 / Accepted: 24 June 2010 / Published online: 17 July 2010
© Springer-Verlag 2010

Abstract

Aims/hypothesis Transcription factor 7-like 2 (TCF7L2) has been strongly implicated in type 2 diabetes and cancer. Our goal was to identify the DNA sequences bound by this transcription factor in vivo.

Methods We applied chromatin immunoprecipitation and sequencing to globally identify and map human DNA sequences bound by TCF7L2 in the colorectal carcinoma cell line, HCT116, where it is abundantly expressed.

J. Zhao and J. Schug contributed equally to this work

Electronic supplementary material The online version of this article (doi:10.1007/s00125-010-1852-3) contains supplementary material, which is available to authorised users.

J. Zhao · S. F. A. Grant
Division of Human Genetics,
The Children's Hospital of Philadelphia,
Philadelphia, PA, USA

J. Schug · K. H. Kaestner · S. F. A. Grant
Department of Genetics and Institute of Diabetes, Obesity
and Metabolism, University of Pennsylvania School of Medicine,
Philadelphia, PA, USA

M. Li
Department of Biostatistics and Epidemiology,
University of Pennsylvania,
Philadelphia, PA, USA

S. F. A. Grant
Department of Pediatrics,
University of Pennsylvania School of Medicine,
Philadelphia, PA, USA

S. F. A. Grant (✉)
Room 1216F, Abramson Research Center,
3615 Civic Center Blvd,
Philadelphia, PA 19104, USA
e-mail: grants@chop.edu

Results We identified 1,095 discrete binding sites across the genome, of which a subset were within 5 kb of 548 annotated NCBI Reference Sequence (RefSeq) genes. Despite using a cancer cell line, the most significant functions represented using pathway analysis software were related to diabetes, genetic disorders and coronary artery disease. As one of the enriched categories was related to genetic disorders, we queried our results against all published genome-wide association studies (GWAS) and found a highly significant over-representation of reported loci from among the genes bound by TCF7L2 within 5 kb ($p=7.50 \times 10^{-15}$). This observation was primarily driven by excess loci revealed from GWAS of metabolic and cardiovascular traits; however, there was no or only minor enrichment of GWAS-derived loci for cancer, and inflammatory or neurological diseases. Of the specific traits, the most enriched loci were for type 2 diabetes and height. When defining the distance from genes at 50 kb or 500 kb, this enrichment pattern persisted, with some additional evidence for enrichment of cancer-related loci.

Conclusions/interpretation A highly significant proportion of genes bound by TCF7L2 are known disease-associated loci. These findings suggest that TCF7L2 is a central node in the regulation of human diabetes and other disease-associated genes.

Keywords ChIP-sequencing · GWAS · TCF7L2 · Type 2 diabetes

Abbreviations

ChIP	Chromatin immunoprecipitation
GLTR	Global Identifier of Target Regions
GWAS	Genome-wide association study
RefSeq	Reference Sequence
SAGE	Serial analysis of gene expression

siRNA	Small interfering RNA
TCF7L2	Transcription factor 7-like 2
TSS	Transcription start site

Introduction

The number of genes already established to play a role in the pathogenesis of type 2 diabetes has grown substantially as a consequence of results from recent genome-wide association studies (GWAS). One of the strongest type 2 diabetes associations to date, based on risk conferred, is with variation in the gene encoding transcription factor 7-like 2 (TCF7L2) [1], which has already been extensively and independently replicated and reviewed [2], and observed in multiple GWAS reports [3–6].

The primary site of action for the TCF7L2 gene product in the context of type 2 diabetes is still unknown. It was initially speculated that TCF7L2 operated in tandem with insulin to influence blood glucose homeostasis by altering levels of glucagon-like peptide 1 [1, 7]. However, other studies have shown that *TCF7L2* variation is associated with decreased insulin secretion, possibly implicating the pancreatic beta cell, although an indirect effect on insulin secretion by TCF7L2 action in another tissue could not be excluded [8–10]. Furthermore, electrophoretic mobility shift assays have shown that several protein complexes bind differently depending on *TCF7L2* genotype in multiple cell lines, including 3T3L1 (adipocyte), HepG2 (hepatocarcinoma) and β -TC3 (beta cell) [11].

TCF7L2 has also been previously linked to cancer risk [12, 13]. Indeed, this connection intensified following reports that the 8q24 locus revealed by GWAS of a number of cancers, including colorectal carcinomas, was due to an extreme upstream TCF7L2-binding element driving the transcription of *MYC* [14, 15].

There is intense interest to determine which genes are transcriptionally regulated by TCF7L2, because these targets may be more amenable to therapeutic intervention than the transcription factor itself. Information on transcription factor binding-site specificity is often incomplete or biased by the prediction methods used. It is therefore clear that experimental studies are required to characterise the many genes transcriptionally regulated by TCF7L2.

Efforts have been made with respect to characterising the binding sites for TCF7L2 (formerly TCF4) through the use of chromatin immunoprecipitation (ChIP)-serial analysis of gene expression (SAGE) [12] and ChIP-on-chip [16]. However, ChIP-sequencing, which combines ChIP and massively parallel sequencing, can be used to directly identify the DNA sequences bound by transcription factors, including strength of binding, in vivo [17]. Here, we

performed global DNA sequence analysis of triplicate TCF7L2 ChIP samples in the human colorectal cell line, HCT116, where TCF7L2 is abundantly produced. Strikingly, we found that genes previously implicated in genetic disorders, primarily metabolic and cardiovascular, are highly enriched among the genes with nearby TCF7L2 binding sites.

Methods

Cell culture and reagent The HCT116 colorectal carcinoma cell line was purchased from the American Type Cell Center (ATCC; Manassas, VA, USA). Cells were cultured at 37°C and 95% humidity, and supplied with 5% CO₂ in ATCC-formulated McCoy's 5a Medium Modified supplemented with 10% by volume fetal bovine serum (Sigma, St Louis, MO, USA), 2 mmol/l L-glutamine (Gibco Invitrogen, Carlsbad, CA, USA), 100 units/ml penicillin and 100 µg/ml streptomycin (Cellgro, Manassas, VA, USA). Based on previous papers outlining TCF7L2 isoforms [18–20], we chose from antibodies that were raised to antigen at the most constant region among TCF7L2 isoforms, i.e. the amino acids encoded by exons 1 to 3 (Catalogue number 05-511; Millipore, Billerica, MA, USA), which have also been previously reported on in successful ChIP-on-chip [16] and SAGE studies [12] of TCF7L2. In addition, our western blot result demonstrates our antibodies' specificity for the main known isoforms (Electronic supplementary material [ESM] Fig. 1).

Chromatin immunoprecipitation Chromatin immunoprecipitation was performed in triplicate following the instructions provided by the suppliers of the EZ-ChIP kit (Millipore) and as described previously [21]. 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, NY, USA). Sonicated chromatin was primarily in the 100 to 500 bp range, averaging 200 to 300 bp.

Sequencing library preparation The sequencing library was prepared as instructed by Illumina (www.illumina.com) (for details, see ESM Methods). Sequencing on the Illumina Genome Analyzer and subsequent analyses was performed at the Functional Genomics Core at the University of Pennsylvania.

Sequencing DNA libraries were assessed for size, purity and quantity using a bioanalyser (2100 B; Agilent Technologies, Santa Clara, CA, USA) followed by sequencing using an Illumina GA-II according to the manufacturer's instructions. TCF7L2 binding sites were associated with NCBI Reference Sequence (RefSeq) transcripts aligned to

the hg18 genome sequence downloaded from UCSC (<http://genome.ucsc.edu>, accessed 28 March 2010). The candidate target gene was the closest gene within a set distance of 5 kb, 50 kb or 500 kb, as well as any other genes that were no more than 50% further than the closest gene or any transcripts that spanned the binding site. In all cases the transcription start site (TSS) of the aligned transcript was used as the anchor point for distance measurements (ESM Methods).

Pathway analysis Data were analysed by Ingenuity Pathways Analysis (www.ingenuity.com) specified for ‘Human’. Genes containing at least one function or pathway annotation in the Ingenuity knowledge base were eligible for the analysis. In our particular case, 529 of 548 genes with a nearby TCF7L2 binding site met the criteria for 5 kb, plus 830 of 866 for 50 kb and 1,007 of 1,067 for 500 kb. The *p* value associated with functions and pathways was calculated using the right-tailed Fisher’s exact test.

GWAS data analysis We based our analysis on all GWAS signals summarised in the NHGRI GWAS catalogue (www.genome.gov/gwastudies) from 8 March 2010. Enrichment was investigated using a two-tailed Fisher’s exact test.

Small interfering RNA knockdown of TCF7L2 and microarray expression analysis To determine the extent of impact on expression of genes bound by TCF7L2 in HCT116 cells, we carried out small interfering RNA (siRNA)-mediated knockdown of TCF7L2. We transfected HCT116 cells with either TCF7L2 siRNA (s13881; Invitrogen, Carlsbad, CA, USA) or siCTRL (12935111; Invitrogen) using Lipofectomine RNAiMAX (1377815; Invitrogen) in four independent transfections. TCF7L2 knockdown efficiency was assessed by western blot using whole-cell lysate extracted after 72 h transfection (ESM Fig. 2). RNA was extracted from the same HCT116 samples and differentially expressed genes in these two settings were examined on an array (G2519F Whole Human Gene Expression V2; Agilent). The overlap between ChIP-sequencing-determined loci and differentially expressed genes was then assessed.

Results

We performed global DNA sequence analysis of triplicate TCF7L2 ChIP samples in the human colorectal cell line, HCT116, where TCF7L2 is abundantly present, with the goal of identifying and mapping human DNA sequences bound by this transcription factor in vivo. Based on previous reports outlining TCF7L2 splicing [18–20], we chose an antibody that was raised to the most constant

region among such isoforms, i.e. the amino acids encoded by exons 1 to 3.

To determine the reproducibility and reliability of our ChIP-sequencing method, we compared the triplicates by scatter plot analysis. The high degree of correlation among the samples indicates ChIP-sequencing was a reliable method for studying TCF7L2 binding genome-wide in this cell-line (Fig. 1). In addition, of the 29 sites we selected for validation purposes with real-time PCR, 27 showed clear evidence of enrichment (ESM Fig. 3), further supporting the reliability of the ChIP-sequencing method. The two sites that did not provide evidence of enrichment had low binding strength scores; however, these sites had been previously reported on in the ChIP-on-chip study [16], where it was suggested that these two sites were undetected because they were below the sensitivity threshold for the quantitative PCR method.

After overlaying all reads from the ChIP-sequencing experiments, 1,095 binding sites were observed at a false discovery rate of 1% using the Global Identifier of Target Regions (GLITR) analysis package [21] (ESM Table 1). The TCF7L2 ChIP signal was clearly distinct from the pseudo-ChIP signal (ESM Fig. 4). The distribution of these binding sites included 37.35% that were intronic (ESM Fig. 5). The majority of binding events were single events within 100 kb of a RefSeq gene (*n*=876), but there were

Fig. 1 Reproducibility of replicates. Plot of the number of reads per million (RPM) from each replicate that overlap with the GLITR regions identified from reads pooled from all samples. Grey lines, direction of the first principal component of the data, i.e. the direction of their correlation. The correlation value is indicated in each plot

105 double events, 22 events with three events, five with four, four with five and one with seven events.

We used the de novo motif discovery algorithm, MDscan [22], to derive the consensus binding site within windows of different lengths, i.e. 7 bp, 11 bp and 15 bp; indeed, when compared with the work of Hatzis et al. [16], the 7 bp consensus motif agrees exactly, while the longer motifs show very strong similarities (ESM Fig. 6).

The TCF7L2-bound regions were within 5 kb (i.e. directly upstream or downstream) of 548 known genes (ESM Table 2). Indeed, approximately 20% of the regions were within 5 kb from the transcriptional start sites of genes, with the remainder being more distant from the promoter, either upstream or within the gene of interest (Fig. 2). We went on to query the SAGE expression library for HCT116 (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM383862) and found that of the 6898 distinct gene symbols assigned, 234 of our TCF7L2 binding sites were within the 5 kb distance setting of 398 of such assigned genes; as such, this indicates that 72.63% of genes bound (i.e. 394 of 548) within this distance were expressed.

Our data also confirm the findings of a prior study that employed serial analysis of TCF7L2 chromatin occupancy in the same cell line [12], i.e. in the neighbourhood of genes including *MYC*, *CCND1*, *CAMK2D*, *HDAC4*, *OBFC1*, *DST*, *UAP1L1*, *RUNX1*, *CUX1*, *GAN*, *DOTIL* and *ATXN1*; however, we did not observe binding in the proximity of two TCF7L2 targets, namely *TGFBR3* and *E2F4*. Our study greatly expands our knowledge of TCF7L2 binding sites in this cell line.

Next, we performed a comprehensive pathway analysis including 529 of the 548 genes within 5 kb of a TCF7L2 binding site that were well annotated functionally. Data

were analysed by Ingenuity Pathways Analysis specified for ‘Human’. The genes containing at least one function or pathway annotation in the Ingenuity knowledge base were eligible for the analysis. Of these, 137 are known in the context of ‘diabetes’, which was ranked the most significant annotation ($p=5.60 \times 10^{-13}$). Other annotations were ‘endocrine system disorder’ (141 genes; $p=1.10 \times 10^{-12}$), ‘genetic disorder’ (304 genes; $p=1.68 \times 10^{-11}$), ‘type 2 diabetes mellitus’ (92 genes; $p=1.28 \times 10^{-10}$) and ‘coronary artery disease’ (83 genes; $p=1.77 \times 10^{-10}$). These 137 genes were also highly enriched among the TCF7L2 targets and easily survived correction for multiple testing (see the top ten annotations in Table 1 and all annotations with adjusted $p < 0.05$ in ESM Table 3). Annotations for cancer were not present at the top of the list, which was somewhat surprising due to the well-known role of TCF7L2 in cancer and the fact that we had studied binding in a colorectal carcinoma cell line. DAVID (<http://david.abcc.ncifcrf.gov>) integrates Biocarta (www.biocarta.com), Panther (www.pantherdb.org), GO (www.geneontology.org) and KEGG (www.genome.jp/kegg), and when we entered our gene list into that package, the most enriched gene set for genetic association was also for metabolism, irrespective of distance from TSS, reflecting our findings with Ingenuity.

As one of the highly enriched categories with Ingenuity was ‘genetic disorder’, we queried the results against all GWAS signals reported to date, as derived from the NHGRI GWAS catalogue (8 March 2010). Of the 548 genes with TCF7L2 binding sites within 5 kb, which represents 2.55% of all RefSeq genes ($n=21,457$), there was a highly statistically significant over-representation of GWAS loci (83 of 1243 [6.68%] loci; $p=7.50 \times 10^{-15}$) (Table 2). This observation was primarily driven by excess loci revealed from GWAS of metabolic (26 of 250 [10.04%] loci; $p=7.31 \times 10^{-9}$) and cardiovascular (17 of 141 [12.06%] loci; $p=4.57 \times 10^{-7}$) traits. Of the specific traits, the most enriched loci were for type 2 diabetes (6 of 21 [28.57%] loci; $p=4.40 \times 10^{-5}$), i.e. *ADAMTS9*, *CDC123*, *CDKAL1*, *FTO*, *TCF7L2* and *WFS1*, and height (15 of 93 [16.13%] loci; $p=7.56 \times 10^{-8}$), i.e. *ANAPC13*, *ANKFN1*, *BCAS3*, *BMP6*, *CDK6*, *DNM3*, *DYM*, *HMGAI*, *HMGAI2*, *HOMER1*, *MAP3K3*, *PXMP3*, *SOCS2*, *ZBTB38* and *ZDHHC7*. Relatively marginal or no significant enrichment of GWAS signals for cancer, neurological or inflammation-related traits was observed.

When defining the distance from genes at 50 kb or 500 kb, this enrichment pattern persisted (all Ingenuity annotations with adjusted $p < 0.05$ in ESM Table 3), with some additional evidence of enrichment of GWAS loci for cancer-related traits when the distance from genes was set at 500 kb (ESM Table 4).

Knockdown of *TCF7L2* using siRNA revealed that a proportion of the genes identified by ChIP-sequencing have

Fig. 2 Cumulative distribution of distances (in basepairs) from TCF7L2 binding sites to the start of genes in the HCT116 cell line. Approximately 20% of the TCF7L2 sites are within 5 kb. Overall, 1,067 of the 1,095 sites are within 500 kb of a TSS

Table 1 Ingenuity pathway analysis: top ten annotations for genes within 5 kb of a TCF7L2 binding site in the HCT116 cell line

Category	Function	Function annotation	<i>p</i> value	Adjusted <i>p</i> value	Genes (<i>n</i>)
Endocrine system disorders	Diabetes	Diabetes	5.60×10^{-13}	8.61×10^{-10}	137
Metabolic disease	Diabetes	Diabetes	5.60×10^{-13}	8.61×10^{-10}	137
Endocrine system disorders	Endocrine system disorder	Endocrine system disorder	1.10×10^{-12}	8.61×10^{-10}	141
Genetic disorder	Genetic disorder	Genetic disorder	1.68×10^{-11}	8.80×10^{-9}	304
Endocrine system disorders	Type 2 diabetes mellitus	Type 2 diabetes mellitus	1.28×10^{-10}	5.03×10^{-8}	92
Metabolic disease	Type 2 diabetes mellitus	Type 2 diabetes mellitus	1.28×10^{-10}	5.03×10^{-8}	92
Genetic disorder	Type 2 diabetes mellitus	Type 2 diabetes mellitus	1.28×10^{-10}	5.03×10^{-8}	92
Genetic disorder	Coronary artery disease	Coronary artery disease	1.77×10^{-10}	5.55×10^{-8}	83
Cardiovascular disease	Coronary artery disease	Coronary artery disease	1.77×10^{-10}	5.55×10^{-8}	83
Cardiovascular disease	Atherosclerosis	Atherosclerosis	3.38×10^{-10}	8.85×10^{-8}	85

Data were analysed using Ingenuity Pathways Analysis (Ingenuity Systems) specified for ‘Human’. Genes containing at least one function or pathway annotation in the Ingenuity knowledge base were eligible for the analysis. In our particular case, 529 of 548 genes with a nearby TCF7L2 binding site met the criteria

The *p* value associated with functions and pathways was generated by the software using the right-tailed Fisher’s exact test; all *p* values survived adjustment for multiple testing

perturbed expression, i.e. 17.0% at 5 kb, 18.0% at 50 kb and 21.2% at 500 kb (ESM Table 5).

Discussion

Using ChIP-sequencing technology, we performed a global location analysis for binding sites of the transcription factor TCF7L2, the gene of which has previously been implicated in type 2 diabetes [1–6] and various cancers [12–15]. We uncovered 1,095 TCF7L2 binding sites in the proximity of 1,067 genes at up to 500 kb distance from the TSS. However, beyond the original goal of characterising the global binding pattern of TCF7L2 in HCT116 cells, the target gene range revealed an unexpected and striking enrichment of loci already implicated in disease, in particular in the metabolic and cardiovascular disease areas. The fact that cancer genes were less abundant among the putative TCF7L2-bound loci may have implications for the

recent claim that TCF7L2 has growth inhibitory or tumour-suppressor functions [23].

Based on data suggesting that TCF7L2 influences insulin secretion from pancreatic islets [10], we had initially aimed to apply this ChIP-sequencing approach in that tissue; however, our efforts were in vain due to the very low TCF7L2 protein levels in islets (ESM Fig. 1).

We compared existing ChIP-on-chip data [16] for TCF7L2 from a different colorectal cancer cell line, namely LS174T. From that existing dataset, it was clear that TCF7L2 (formerly TCF4) binds to over six times more sites in LS174T than in HCT116 cells, i.e. 6,868 vs 1,095 sites. This would suggest that TCF7L2 binding is more specific in HCT116 cells and therefore more degenerate in LS174T cells, although the different platforms employed in the respective studies must also be considered. In addition the choice of antibody is crucial for this transcription factor, as it is heavily alternatively spliced. Despite this large difference in apparent binding specificity and the fact that

Table 2 Enrichment of GWAS signals for RefSeq genes within 5 kb of a TCF7L2 binding site in the HCT116 cell line

Variables	Total genes	Genes from TCF7L2 ChIP-sequencing	%	Fisher’s exact <i>p</i> value
All RefSeq Genes	21,457	548	2.55	
All Genes found by GWAS	1,243	83	6.68	7.50×10^{-15}
Metabolism	250	26	10.04	7.31×10^{-9}
Type 2 diabetes	21	6	28.57	4.40×10^{-5}
Height	93	15	16.13	7.56×10^{-8}
Cancer	102	9	8.82	0.00184
Cardiovascular	141	17	12.06	4.57×10^{-7}
Inflammation	180	9	5.00	0.0560
Neurology	281	15	5.34	0.0124

We based our analysis on GWAS signals summarised in the NHGRI GWAS catalogue (www.genome.gov/gwastudies)
Enrichment was investigated using two-tailed Fisher’s exact test

there is evidence of abundant binding in LS174T cells and potentially a high degree of non-specificity (within 5 kb of 13.09% of RefSeqs and within 500 kb of 25.53% of RefSeqs), we were able to investigate the publically available dataset to see if enrichment also occurred in this cell line. Indeed, irrespective of the setting for distance from TSS, ‘cardiovascular disorder’, ‘diabetes’, ‘endocrine system disorder’ and ‘genetic disorder’ were consistently among the highly significant top ten annotations in Ingenuity runs (ESM Table 6). In addition, although the fold enrichment was not so marked, we also observed evidence of an over-representation of GWAS loci in this dataset. Indeed, when we analysed the overlap of the genes in the Ingenuity categories ‘endocrine system disorders/metabolic disease’ in our dataset with the genes of the same category in Hatzis et al. [16], there was extensive overlap of 62.8% at 5 kb, 62.7% at 50 kb and 69.8% at 500 kb (ESM Table 7), which further supports our observations.

It is well established that TCF7L2 is benignly bound to many regions in the genome and it is only when wnt signalling is activated that transcription itself is activated. It is highly likely that the various genes that are bound by TCF7L2 and then go on to be transcriptionally activated will differ markedly between cell types. The main goal of this study was to map out the binding sites of TCF7L2 using ChIP-sequencing in HCT116 cells. However, we found that knock-down experiments with siRNA for *TCF7L2* affected approximately a fifth of the genes bound by TCF7L2 at the expression level in this particular cell line.

The study by Hatzis et al. [16] compared the 293 β -catenin binding sites identified by Yochum et al. [12] in HCT116 cells represented on the NimbleGen genome-wide arrays, where 52 (18%) overlapped with the 6,868 sites observed by Hatzis et al.; it was therefore deemed greater than would be expected with random genomic sequences. In fact, we found the proportion to be substantially higher when comparing the Hatzis et al. [16] dataset with our ChIP-sequencing dataset, with 337 of the 1,095 binding sites identified (31%) overlapping with their 6,868 sites. Together with our quantitative PCR validation results, these observations lend credibility to the list of binding sites in HCT116 cells outlined in this paper.

Within 500 kb from the TSS, there was also evidence of enrichment of cancer-related GWAS loci. Indeed, the 8q24 locus, found in GWAS to be associated with a number of cancers, including colorectal carcinomas, probably causes disease through mutation of an upstream TCF7L2-binding element driving transcription of *MYC* [14, 15]. Thus, *TCF7L2* appears to be a key player in type 2 diabetes and cancer, an observation supported by the data we present here. Although there are apparent genetic links between cancer and type 2 diabetes, the mechanism is still not fully

understood. Curiously, many of the GWAS-derived risk-conferring alleles for type 2 diabetes have been shown to protect against prostate cancer [24]. Our findings suggest a possible connection between these previous observations.

One caveat to the pathway analyses is that the functional annotations are often highly correlated, e.g. there is a large overlap for the diabetes genes and the endocrine system disorder genes, so such categories cannot be considered completely independent. However, it is clear that genes implicated in endocrine related disorders are strongly enriched in our gene list compared with other disorders, in particular cancer, despite the use of a cancer cell line. The Ingenuity results also prompted us to investigate the GWAS literature with respect to our gene list.

Genes near TCF7L2 binding sites identified in this study are primarily significantly over-represented with respect to GWAS signals for metabolic and cardiovascular disorders. The fact that there was little or no enrichment for GWAS loci in the fields of inflammatory or neurological disease, despite the number of known loci for the former being similar to those for metabolism, quells concerns that the observed over-representations are artefacts. Indeed, these diseases are further from the disease areas where TCF7L2 has been primarily implicated and could be considered closer to the baseline for enrichment considerations.

Our data suggest that the influence of TCF7L2, with respect to genetic disease, extends much further than previous knowledge about this transcription factor’s role in disease. This therefore poses the question: what do the genes on this list that have not been previously implicated by human genetics in disease processes represent? One could postulate that many novel genes on the TCF7L2 target list are relevant to disease and may harbour the much sought-after rare variants contributing to diabetes, cardiovascular disease and potentially other related disorders such as cancer. As such, they represent good candidates for a re-sequencing exercise in multiple disease settings. Indeed, one of the conclusions in a recent paper investigating chromatin signatures was that heritable transcription factor binding differs on the basis of genetic variation and may underlie phenotypic variation in humans [25]; as such, our observations resonate with this finding.

In summary, this project applied a combination of two techniques, namely chromatin immuno-precipitation and ultra-high-throughput sequencing (ChIP-sequencing), in order to identify and map human DNA sequences that are bound by the TCF7L2 protein in vivo. Unexpectedly, we observed a highly significant over-representation of loci already implicated in disease. We have revealed novel candidate therapeutic targets for type 2 diabetes, and potentially for other diseases, which could be followed up in functional characterisation experiments. Importantly, our data suggest that TCF7L2 is a central node in the genetic

basis for type 2 diabetes, regulating multiple genes also associated with disease risk. This finding might also explain why TCF7L2 confers the strongest risk for diabetes among all loci identified thus far, as any alteration in TCF7L2 function will affect an entire network of diabetes-associated genes.

Acknowledgements We would like to thank G. S. Yochum for providing the primers for the known TCF7L2 binding sites in HCT116 cells and G. Tuteja for tutoring on sequencing library preparation. This work was supported by the Ethel Brown Foerderer Fund for Excellence from the Children's Hospital of Philadelphia, a Penn Center for Musculoskeletal Disorders pilot award and the University of Pennsylvania Diabetes and Endocrinology Research Center (NIH DK 19525). We would also like to acknowledge support from the Institute for Diabetes, Obesity, and Metabolism (IDOM) Functional Genomics Core NIDDK Research Center grant P30DK19525. Human islet samples were provided by the Islet Cell Resource Centers Basic Science Islet Distribution Program.

Duality of interest The authors declare that there is no duality of interest associated with this manuscript

References

- Grant SF, Thorleifsson G, Reynisdottir I et al (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat Genet* 38:320–323
- Cauchi S, El Achhab Y, Choquet H et al (2007) TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis. *J Mol Med* 85:777–782
- Sladek R, Rocheleau G, Rung J et al (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445:881–885
- Saxena R, Voight BF, Lyssenko V et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331–1336
- Scott LJ, Mohlke KL, Bonnycastle LL et al (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341–1345
- Zeggini E, Weedon MN, Lindgren CM et al (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316:1336–1341
- Yi F, Brubaker PL, Jin T (2005) TCF-4 mediates cell type-specific regulation of proglucagon gene expression by beta-catenin and glycogen synthase kinase-3beta. *J Biol Chem* 280:1457–1464
- Florez JC, Jablonski KA, Bayley N et al (2006) TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program. *N Engl J Med* 355:241–250
- Saxena R, Gianniny L, Burt NP et al (2006) Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes* 55:2890–2895
- Lyssenko V, Lupi R, Marchetti P et al (2007) Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes. *J Clin Invest* 117:2155–2163
- Cauchi S, Choquet H, Gutierrez-Aguilar R et al (2008) Effects of TCF7L2 polymorphisms on obesity in European populations. *Obesity (Silver Spring)* 16:476–482
- Yochum GS, McWeeney S, Rajaraman V, Cleland R, Peters S, Goodman RH (2007) Serial analysis of chromatin occupancy identifies beta-catenin target genes in colorectal carcinoma cells. *Proc Natl Acad Sci USA* 104:3324–3329
- Duval A, Busson-Leconiat M, Berger R, Hamelin R (2000) Assignment of the TCF-4 gene (TCF7L2) to human chromosome band 10q25.3. *Cytogenet Cell Genet* 88:264–265
- Pomerantz MM, Ahmadiyeh N, Jia L et al (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* 41:882–884
- Tuupanen S, Turunen M, Lehtonen R et al (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nat Genet* 41:885–890
- Hatzis P, van der Flier LG, van Driel MA et al (2008) Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol Cell Biol* 28:2732–2744
- Robertson G, Hirst M, Bainbridge M et al (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Meth* 4:651–657
- Mondal AK, Das SK, Baldini G et al (2010) Genotype and tissue-specific effects on alternative splicing of the transcription factor 7-like 2 gene in humans. *J Clin Endocrinol Metab* 95:1450–1457
- Osmark P, Hansson O, Jonsson A, Ronn T, Groop L, Renstrom E (2009) Unique splicing pattern of the TCF7L2 gene in human pancreatic islets. *Diabetologia* 52:850–854
- Prokunina-Olsson L, Welch C, Hansson O et al (2009) Tissue-specific alternative splicing of TCF7L2. *Hum Mol Genet* 18:3795–3804
- Tuteja G, White P, Schug J, Kaestner KH (2009) Extracting transcription factor targets from ChIP-Seq data. *Nucleic Acids Res* 37:e113
- Li W, Meyer CA, Liu XS (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics (Oxford, England)* 21(Suppl 1):i274–i282
- Tang W, Dodge M, Gundapaneni D, Michnoff C, Roth M, Lum L (2008) A genome-wide RNAi screen for Wnt/beta-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer. *Proc Natl Acad Sci USA* 105:9697–9702
- Frayling TM, Colhoun H, Florez JC (2008) A genetic link between type 2 diabetes and prostate cancer. *Diabetologia* 51:1757–1760
- McDaniell R, Lee BK, Song L et al (2010) Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328:235–239