Leveraging cross-species transcription factor binding site patterns: from diabetes risk Loci to disease mechanisms.

Claussnitzer, Melina; Dankel, Simon N; Klocke, Bernward; Grallert, Harald; Glunk, Viktoria; Berulava, Tea; Lee, Heekyoung; Oskolkov, Nikolay; Fadista, Joao; Ehlers, Kerstin; Wahl, Simone; Hoffmann, Christoph; Qian, Kun; Rönn, Tina; Riess, Helene; Müller-Nurasyid, Martina; Bretschneider, Nancy; Schroeder, Timm; Skurk, Thomas; Horsthemke, Bernhard; Spieler, Derek; Klingenspor, Martin; Seifert, Martin; Kern, Michael J; Mejhert, Niklas; Dahlman, Ingrid; Hansson, Ola; Hauck, Stefanie M; Blüher, Matthias; Arner, Peter; Groop, Leif; Illig, Thomas; Suhre, Karsten; Hsu, Yi-Hsiang; Mellgren, Gunnar; Hauner, Hans; Laumen, Helmut

Published in:
Cell

DOI:
10.1016/j.cell.2013.10.058

2014

Title: Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms

Melina Claussnitzer[1,2,3,4*], Simon N. Dankel [6,7], Bernward Klocke [5], Harald Grallert [8], Viktoria Glunk [1,2,3,4], Tea Berulava[9], Heekyoung Lee [1,2,3,4], Nikolay Oskolkov [10], Joao Fadista [10], Kerstin Ehlers [1,2,3,4], Simone Wahl[8], Christoph Hoffmann [2,11], Kun Qian [1,2,3,4], Tina Rönn [10], Lena Riess [12-14], Martina Müller-Nurasyid [12-14], Nancy Bretschneider[5], Thomas Skurk [1,2], Bernhard Horsthemke[9], DIAGRAM+ Consortium, Derek Spieler [15,16], Martin Klingenspor [2,11], Martin Seifert[5], Michael J. Kern [17], Niklas Mejhert [10], Ingrid Dahlman [10], Ola Hansson [10], Stefanie M. Hauck [19], Matthias Blüher [18], Peter Arner 23, Leif Groop [10], Thomas Illig [20], Karsten Suhre, Yi-Hsiang Hsu, Gunnar Mellgren [6,7], Hans Hauner [1,2,3,4,20], Helmut Laumen [1,2,3,4,21*]

1Else Kroener-Fresenius-Centre for Nutritional Medicine, Chair of Nutritional Medicine, Technical University München, 85350 Freising-Weihenstephan, Germany
2ZIEL - Research Centre for Nutrition and Food Sciences, Technical University München, 85350 Freising-Weihenstephan, Germany
3German Center for Diabetes Research (DZD)
4Clinical Cooperation Group Nutrigenomics and Type 2 Diabetes, Helmholtz Zentrum München and Technical University München, 85350 Freising-Weihenstephan
5Genomatix Software GmbH, 80335 Munich, Germany
6Department of Clinical Science, University of Bergen, Bergen 5021, Norway
7Hormone Laboratory, Haukeland University Hospital, Bergen 5021, Norway
8Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
9Institut für Humangenetik, Universitätsklinikum Essen, 45147 Essen, Germany
10Diabetes and Endocrinology Research Unit, Department of Clinical Sciences, Lund University, Malmö 20502, Sweden
11Else Kroener-Fresenius-Centre for Nutritional Medicine, Chair of Molecular Nutritional Medicine, Technical University München, 85350 Freising-Weihenstephan, Germany
12Department of Medicine I, University Hospital Grosshadern, Ludwig-Maximilians-Universität, 81377 Munich, Germany
13Chair of Genetic Epidemiology, Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, 81377 Munich, Germany
14Institute of Genetic Epidemiology, Helmholtz Zentrum München-German Research Center for Environmental Health, 85764 Neuherberg, Germany
15Institute of Human Genetics, Helmholtz Zentrum München, 85764 Neuherberg, Germany
16Institute of Human Genetics, Technical University München, Munich, Germany
17Department of Regenerative Medicine and Cell Biology, Medical University of South Carolina, Charleston, SC 29425, USA
18Department of Medicine, University of Leipzig, 04103 Leipzig, Germany.
19Research Unit Protein Science, Helmholtz Zentrum München, 85764 Neuherberg, Germany
20Hanover Unified Biobank, Hanover Medical School, 30625 Hannover, Germany
21Else Kroener-Fresenius-Centre for Nutritional Medicine, Klinikum rechts der Isar, Technical University München, 81675 Munich, Germany
22Institute of Experimental Genetics, Helmholtz Zentrum München, Neuherberg 85764, Germany
23 Department of Medicine, Karolinska Institutet,Center for Endocrinology and Metabolism, Karolinska University Hospital Huddinge, SE-141 86 Stockholm, Sweden *Correspondence: melina.claussnitzer@tum.de (M.C.), helmut.laumen@tum.de (H.L.).

SUMMARY

Genome-wide association studies revealed numerous risk loci associated with diverse diseases. However, identification and mechanistic elucidation of the disease-causing variants within association loci remains a major challenge. Divergence in gene expression due to cis-regulatory variation is central to disease risk. We show that integrative analysis of cross-species conserved transcription factor binding site (TFBS) patterns can identify cis-regulatory variants and elucidate the mechanisms mediating their role in diseases. Analysis of established type 2 diabetes risk loci revealed a striking clustering of distinct homeobox TFBSs. We unveiled a novel activity of the PRRX1 homeobox transcription factor as a repressor of PPARγ2 expression, and showed its adverse effect on lipid metabolism and in human adipose tissue samples from rs4684847 risk allele carriers, resulting from SNP-mediated increase in PRRX1 binding affinity. Thus, cross-species conservation analysis at the level of co-occurring TFBSs provides a valuable contribution to the translation of genetic association signals to disease-related molecular mechanisms.

HIGHLIGHTS

►Cross-species analysis of co-occurring TFBSs predicts cis-regulatory variants ►Analysis of diabetes-associated loci reveals clustering of distinct homeobox TFBSs ►rs4684847 at the PPARG diabetes-locus influences binding of the homeobox TF PRRX1 ►PRRX1 represses adipose PPARγ2 transcript levels in rs4684847 risk allele carriers

INTRODUCTION

Recent advances in genome-wide association studies (GWAS) have yielded a plethora of loci associated with diverse human diseases and traits (Hindorff LA). However, signals emerging from GWAS, which identify typically dozens of variants in linkage disequilibrium (LD), have rarely been traced to the disease-causing variants and even more rarely to the mechanisms by which they may increase disease risk. The majority of common genetic variants are located in non-coding regions (1000 Genomes Project Consortium, 2012), and disease-associated loci are enriched for eQTLs (Nica et al., 2010), DHSseq and ChIPseq peaks (Maurano et al., 2012; The ENCODE Project Consortium, 2012), suggesting that variants modulating gene regulation are major contributors to common disease risk.

Experimental DHS-, RNA-, and ChIPseq approaches have been used to prioritize candidate *cis*-regulatory variants (Maurano et al., 2012; The ENCODE Project Consortium, 2012; Ward and Kellis, 2012b). However, such experimental approaches require access to appropriate human tissues and are hampered by the spatial, temporal, environmental and epigenetic complexity of gene regulation. These limitations emphasize the need for bioinformatics approaches that reliably assess the regulatory role of non-coding variants. So far, phylogenetic conservation has been a common denominator in the search for non-coding regulatory regions (Chinwalla et al., 2002; Pennacchio et al., 2006; The ENCODE Project Consortium, 2007; Visel et al., 2009b; Blow et al., 2010; Lindblad-Toh et al., 2011; The ENCODE Project Consortium, 2012). Unfortunately, intra- and cross-species differences in gene expression are often driven by changes in transcription factor binding sites (TFBSs), and their rapid evolutionary turnover results in lineage-specific regulatory regions that are functionally conserved but have low phylogenetic conservation (Ward and Kellis, 2012a), thus challenging the use of these algorithms. Importantly, gene regulatory regions in eukaryotes tend to be organized in *cis*-regulatory modules (CRMs), comprising complex

patterns of co-occurring TFBSs for the combinatorial binding of transcription factors (TFs) (Arnone and Davidson, 1997; Pennacchio et al., 2006; Visel et al., 2013). CRMs integrate upstream signals to regulate the expression of coordinated gene sets, making them a prime target to achieve phenotypic changes as a result of adaptive evolution (Junion et al., 2012). Despite the critical importance of CRMs, no algorithms have so far been developed to harness the potential power of conserved TFBS patterns within CRMs to predict regulatory variants in disease genetics.

Here, by phylogenetic module complexity analysis (PMCA), we show that cross-species conservation at the level of the CRMs – rather than at the level of the regulatory sequence that comprises them – identifies *cis*-regulatory variants within disease-associated GWAS loci. We applied PMCA to type 2 diabetes (T2D) risk loci, for which the specific causal *cis*-regulatory variants have rarely been pinpointed (Stitzel et al., 2010).

RESULTS

Cross-species analysis of TFBS modularity discovers *cis*-regulatory SNPs at T2D risk loci

We developed a method, PMCA, which leverages conserved co-occurring TFBS patterns within CRMs to predict *cis*-regulatory variants, i.e. variants affecting gene expression (Figure 1A, detailed description of the Procedure in Extended Experimental Procedures). To systematically identify the *cis*-regulatory variants at GWAS risk loci, we extracted the GWAS tagSNPs, and consequently all non-coding (nc) SNPs that are in high LD with these tagSNPs. PMCA individually tests each nc variant by analyzing the flanking region for cross-species conserved TFBS patterns, regardless of global sequence conservation. This requires first the extraction of the region surrounding a nc SNP (±60bp) from the human genome, and

consequent identification of orthologous regions in 15 vertebrate species. Within each SNP-specific set of orthologous regions, phylogenetically conserved TFBSs, TFBS modules (a cross-species conserved pattern of two or more TFBSs occurring in the same order and in a certain distance range) and TFBSs in those TFBS modules were identified and then counted. SNP-surrounding regions with a significant enrichment of phylogenetically conserved TFBS modules are classified as *complex regions*, as compared to *non-complex regions* (example in Figure 1B) wherein the occurrence of TFBS modules does not exceed expectation by chance. To compute this enrichment we estimated background probabilities using randomizations of orthologous sets (details on scoring cut-offs in Extended Experimental Procedures).

We applied PMCA to eight GWAS T2D risk loci (*MTNR1B*, *TCF7L2*, *PPARG*, *CENTD2*, *FTO*, *GCK*, *CAMK1D*, *KLF14*) (Dupuis et al., 2010; Voight et al., 2010) (Figure 1C) covering strong and weaker GWAS signals, and reflecting the different T2D features, i.e. insulin resistance and impaired insulin secretion (Doria et al., 2008). Using non-coding sequence information we defined 200 SNPs in LD with the tagSNPs ($r^2 \geq 0.7$, 1000G, Figure S1A-H,). PMCA predicted 64 complex and 136 non-complex regions (Table S1, Figure 1C-G). We ranked complex regions based on the count of TFBSs in conserved TFBS modules (Table S2), and examined the allele-dependent *cis*-regulatory potential of the 25% highest scoring SNPs using electrophoretic mobility shift assays (EMSA) and reporter assays. As predicted, SNPs in complex regions significantly differed in allele-dependent *cis*-regulatory activity compared to control non-complex regions (Figure 1H,I, Table S3). Indeed, the regulatory variants revealed effects ranging from 3.1- to 101-fold change in DNA-protein binding and 1.3- to 3.5-fold change in reporter activity. We further verified that the identified variants operate in a cell type-specific manner (Figure S1I).

To examine if the identified *cis*-regulatory variants associate with T2D *in vivo*, we performed look-ups in the population-based MAGIC and DIAGRAM cohorts (Dupuis et al.,

2010; Voight et al., 2010). The *cis*-regulatory variants in complex regions revealed a similar or stronger association compared to the initial GWAS signal (Table S4), and a look-up in a recent fine-mapping study (Maller et al., 2012) reassured that our *cis*-regulatory SNPs belong to the predicted disease causal SNP set. To further assess the predictive power of PMCA more generally, we analyzed GWAS signals for 18 human diseases(Hindorff LA) and confirmed an enrichment of SNPs in complex regions relative to random SNPs matched for MAF and genomic localization from the 1,000G Project ($P=1.9 \times 10^{-4}$, binominal, Table S5).

Moreover, we applied PMCA on reported *cis*-regulatory SNPs associated with diverse disease-related traits, including cancer, myocardial infarction, thyroid hormone resistance, hypercholesterolemia and adiponectin levels (*MYC* Pomerantz et al., 2009 *MDM2* Post et al., 2010 *PSMA6* Ozaki et al., 2006 *THRB* Alberobello et al., 2011 *SORT1* Musunuru et al., 2010 *APM2* Laumen et al., 2009). Consistent with the functional proof from the original publications, our analysis informed on all but one of the *cis*-regulatory SNPs (Table S6). The highest scores inferred from PMCA predicted the reported myocardial infarction risk variant, which was shown to regulate hepatic *SORT1* expression (Musunuru et al., 2010). Together, these results demonstrate the utility of TFBS modularity information within CRMs to elucidate functionality of GWAS signals in the non-coding genome.

Clustering of distinct homeobox TFBSs is a specific feature of T2D-related complex regions

Considering that TFBS turnover is characteristic for CRM evolution (Blow et al., 2010; Ward and Kellis, 2012a), the utility of sequence conservation in deciphering *cis*-regulatory variants may be limited. To assess the power of harnessing TFBS patterns, which allows sequence variability, beyond conventional sequence conservation, we performed PMCA on all 47 autosomal T2D risk loci (Hindorff LA accessed June 2012; 1,465 SNPs; $r^2 \geq 0.7$; Figure S2A-

E,H; Table S4) and tested the resulting 487 complex and 978 non-complex regions for correlations with evolutionary constrained elements detected by the SiPhy-π-method (Lindblad-Toh et al., 2011). We observed that non-complex regions are depleted of constrained elements in their close proximity (Figure 2A). Conversely, complex regions were enriched for nearby constrained elements, consistent with a 1.37-fold enrichment of GWAS SNPs relative to HapMap SNPs (Lindblad-Toh et al., 2011). Strikingly, however, though we found a 1.88-fold enriched overlap with complex relative to non-complex regions ($p=2.4 \times 10^{-9}$, hypergeometric distribution, right sided), the majority of complex regions lacked an overlap with constrained elements (Figure 2B, Table S8). This lack of overlap was true for all variants that we experimentally characterized as *cis*-regulatory (example in Figure 2C). In essence, considering sequence conservation helps to prioritize genomic regions that harbor potential causal variants, yet seems insufficient to pinpoint them. This underscores the importance of exploiting conservation in terms of a complexity assessment of co-occurring TFBSs, in the search for *cis*-regulatory variants involved in human diseases.

To further support PMCA predictions with functional genomics data we compiled chromatin state and TF binding data from the ENCODE consortium (2011). We found complex regions highly enriched for both DHS and ChIP-seq peaks ($p=3.52 \times 10^{-10}$ $p=4.68 \times 10^{-6}$, respectively, hypergeometric distribution, right sided, Figures 2D,E, Figures S2F,G, Table S9). Additionally, crossing our regulatory predictions for T2D SNPs with a recently published analysis of multiple types of functional ENCODE data (Schaub et al., 2012) confirmed that complex regions are significantly enriched for functionality ($P=3 \times 10^{-24}$, hypergeometric distribution, right sided, Table S10).

Next we sought evidence for a discerning T2D functional feature. TFBS clustering relative to transcription start sites indicates biological significance (FitzGerald et al., 2004), and TFBS combination coupled with the TFs recruited to a CRM determines CRM function

(Zinzen et al., 2009). Given a SNP genomic region we used positional bias analysis, scanning 1,000bp with the SNP at midposition for the occurrence of putative TF binding sequences (883 TFBS matrices grouped in 192 TFBS matrix families, Table S11). For the eight T2D risk loci analyzed above, we observed a significant positional bias for distinct TFBS families (-log10(P)>6) exactly at SNP position of complex (Figure 3A) contrary to non-complex regions (Figure S3A). The striking SNP-directed overrepresentation in T2D complex regions was restricted to specific TFBSs in the superfamily of homeobox TFs, including the matrix families CART (-log10(P)=6.52) and PDX1 (-log10 (P)=6.18) (Table S12A). These findings were reproduced in the set of 47 T2D risk loci (Table S12B) which showed clustering at SNP position exclusively in complex regions and again co-localization of T2D risk SNPs with homeobox TFBS matrices (Figure 3B,C). In this extended analysis of all T2D risk loci, we again found the CART (-log10(P)=13.00) and PDX1 families (-log10(P)=6.78) together with the homeobox matrix families NKX6, HOMF, HBOX and BCDF (-log10(P)=8.50, 8.94, 8.54 and 7.24; respectively). No other TFBS matrices showed a significant peak in the bias profile at SNP position (Table 12C). Importantly, performing PMCA for risk loci of T2D non-related traits, Crohn's disease (Schaub et al., 2012) and asthma (Moffatt et al., 2010) (Figures S3E-N, Table S13), revealed disease-specific TFBSs at SNP position (Tables S12C-D; Figure S3B,D for Crohn's). The specific clustering of the Early Growth Response Factor matrix family (EGRF, -log10(P)=8.50, Figures 3D,3E,S3C) for asthma risk SNPs in complex regions was in strong contrast to T2D and Crohn (Figure 3F, -log10(P)=3.97, 2.07, respectively). Of note, the EGRF-binding factor EGR1 regulates asthma-related IL13-induced inflammation (Cho et al., 2006).

Homeobox TFs are known to be involved in embryonic and tissue developmental processes including β-cell development (Jonsson et al., 1994; Harrison et al., 1999; Nekrep et al., 2008). However, except for the mature onset of diabetes gene *PDX1* (Fajans et al., 2001) and the common T2D-associated loci *HHEX1* and *ALX4* (Sladek et al., 2007), the inferred

homeobox factors have not been implicated in T2D pathogenesis. T2D is marked by insulin resistance and impaired insulin secretion (Doria et al., 2008). To evaluate a functional role of the inferred T2D-specific homeobox TFBS matrix families in T2D pathogenesis, we extracted data for insulin resistance (HOMA-IR) and impaired insulin secretion (HOMA-B) (Dupuis et al., 2010), to compute the enrichment of predicted *cis*-regulatory T2D risk SNPs that localize in close proximity to an inferred homeobox TFBS (+/-20bp, (permutations on the phenotypes, n=1,000, 95% confidence interval, Extended Experimental Procedures). We verified a significant enrichment of SNPs that localize ±20bp at inferred homeobox TFBS for both insulin resistance ($p=1.287 \times 10^{-7}$, $mean=9.45 \times 10^{-4}$, CI: $5.37 \times 10^{-4} - 1.34 \times 10^{-2}$) and impaired insulin secretion (p-value=3.281×10^{-4}, $mean=1.09 \times 10^{-6}$, CI: $9.59 \times 10^{-7} - 9.51 \times 10^{-3}$). Furthermore, we evaluated a possible effect of the binding TFs on impaired insulin secretion. By assessing mRNA levels in human islets from deceased donors with and without T2D (8 and 51, respectively) (RNA-seq, L. Groop, unpublished data) we found a marked mRNA expression difference for *RAX*, *PRRX2*, *BARX1*, *PITX1*, *EMX2*, *NKX6-3*, *BARX2*, *MSX2* and *PDX1* in islets from T2D patients compared to controls (FDR < 1%, Table S14). By genome-wide co-expression analysis, we found significantly co-regulated gene sets (Table S15, FDR < 5%). All but one of these co-regulated gene sets included the category "metabolic pathways" among the top 5 significantly enriched pathways (hypergeometric test, FDR 5%, other top 5 enriched pathways included insulin signaling, MAPK signaling, Notch signaling, Calcium signaling and pancreatic secretion, Figures S3O-T). Knock down of candidate homeobox TFs in pancreatic INS-1 β cells further confirmed significant perturbation of glucose-stimulated insulin secretion (Figure S3U). Strikingly, for all PMCA-inferred homeobox TFs except for PDX1 and MSX2 (FDR 5% corrected p-value=0.53 and 0.076, respectively), we found a significant co-expression with the insulin gene. Although the result for PDX1 was borderline non-significant it is a well-known regulator of insulin

expression (Brissova et al., 2002). The other identified homeobox TFs may be regarded as novel candidates for regulation of proinsulin production.

The T2D identified variant rs4684847 regulates *PPARG2* gene expression

To establish the informative value of TFBS pattern analysis we chose the *PPARG* locus for detailed study. PPARγ is crucial in adipogenesis, lipid metabolism and systemic insulin sensitivity (Rosen et al., 1999; Zhang et al., 2004), and exists as two isoforms: PPARγ1 (*PPARG1*, *PPARG3* mRNA) and PPARγ2 (*PPARG2* mRNA) (Fajas et al., 1998)(Figure 4A), the latter being mainly expressed in adipocytes (Tontonoz et al., 1994). Several studies have established a robust association of *PPARG* with T2D (Deeb et al., 1998; Heikkinen et al., 2009; Dupuis et al., 2010; Voight et al., 2010). Yet, results from these studies have appeared contradictory. The T2D GWAS association comes from an LD region mainly tagged by the coding missense mutation Pro12Ala. However, the minor 12Ala allele, associated with enhanced insulin sensitivity in humans, paradoxically blunts the transcriptional activity of the insulin-sensitizing PPARγ2 TF (Deeb et al., 1998). Hypothesizing that the elusive *PPARG* T2D signal arises from a regulatory variant which instead increases *PPARG2* expression, we applied PMCA to all 23 correlated non-coding variants and found six complex regions ($r^2 \geq 0.7$, Figure 4A). Luciferase assays showed that the *cis*-regulatory activity of each complex region significantly differed from non-complex regions (p=0.02, Figure S4A, Table S16). Indeed, qRT-PCR on human adipose stromal cells (hASCs) revealed a risk allele-dependent 3.8-fold decrease of *PPARG2* mRNA ($p=1.00 \times 10^{-3}$, Figure 4B), whereas *PPARG1* expression was unaffected (Figure 4C). Using allele-specific primer extension assay in heterozygous hASCs, we found a striking allelic imbalance with 5.4-fold lower *PPARG2* expression from the risk allele ($p=6.00 \times 10^{-4}$, Figure 4D). When we studied adipose tissue eQTL data we observed an up-regulation of total *PPARG* mRNA in risk allele carriers

(p=0.01, Figure S4B). Concurrent *PPARG2* decrease and total *PPARG* increase might be explained by the co-occurrence of activating or repressing variants within the analyzed haplotype. Indeed, reporter assays demonstrated either activating or repressing risk allele-dependent effects for all but one (rs35000407) predicted SNP (Figure 4E).

Interestingly, the risk allele-dependent suppression of *PPARG2* mRNA diminished with progression of adipocyte differentiation ($p < 0.001$, Figure S4C). We therefore integrated genome-wide H3K27ac data of hASCs undergoing adipogenesis(Mikkelsen et al., 2010) with complex regions, and observed H3K27ac temporal density distributions consistent with the cell stage-dependent regulatory effect for the variant rs4684847 (Figure S4D). To prove that rs4684847 explains *PPARG2* mRNA suppression we first performed reporter assays demonstrating a 5.2-fold decrease in transcriptional activity for the risk allele in 3T3-L1 preadipocytes ($p=1.0 \times 10^{-4}$, Figure 4F). This effect was independent of 5´- vs. 3´-orientation to the reporter gene (p=0.03) and forward vs. reverse orientation (p=0.03) (Figure S4E), suggesting enhancer function for the non-risk allelic complex region. Consistent with the GWAS signal for insulin resistance rather than insulin secretion (Voight et al., 2010), we observed rs4684847 cell type-specific effects in 3T3-L1 adipose cells, C2C12 myocytes and Huh7 hepatocytes, whereas pancreatic INS-1 β-cells and 293T cells lacked allelic activity (Figure S2F). Last, using EMSA we found rs4684847 risk allele-specific DNA-protein binding (Figure 4H).

To remind, exploiting cross-species TFBS patterns at T2D loci unveiled distinct homeobox TFBS families including the CART matrix family (-log10(P)=13.0, Figure 3B). For the PPARG locus, we relate this specific TFBS matrix clustering to a binding sequence in the CART matrix family, which harbors the rs4684847 *cis*-regulatory variant and which is predicted to bind the paired-related homeobox protein-1 (PRRX1) (Figure 4G). By affinity chromatography and LC-MS/MS we demonstrate a 2.3-fold increased binding of PRRX1 to

the rs4684847 risk relative to non-risk allele (Extended Experimental Procedures). Competition EMSA and supershift experiments confirmed that the identified TF PRRX1 was responsible for allele-specific DNA-protein binding (Figure 4I). Perturbing the PRRX1 consensus sequence without affecting SNP position itself fully abrogated the risk allelic repression of reporter gene activity (Figure 4F), whereas overexpressing PRRX1 enhanced it ($p=2 \times 10^{-4}$, Figure 4J). Because the rs4684847 is in near-perfect LD with 23 non-coding variants, we tested if the rs4684847 risk allele – independent of correlated sequence variants – causes the suppression of endogenous *PPARG2* expression (Figure 4B-D). We used an adopted CRISPR/Cas homology-directed repair genome editing approach (Wang et al., 2013) to introduce the rs4684847 non-risk allele in human SGBS preadipocytes, replacing the endogenous risk allele. Notably, the rs4684847 non-risk allele was sufficient to increase *PPARG2* transcript levels by 5.4-fold (p=0.005, Figure 4K), whereas *PPARG1* mRNA was unaffected (Figure S4G). In parallel we performed PRRX1 knockdown and confirmed that 1) risk allele-driven suppression of *PPARG2* expression was reversed by PRRX1 silencing (p=0.005) and 2) PRRX1 silencing did not affect *PPARG2* expression in non-risk allele cells (Figure 4K).

rs4684847 via PRRX1 binding affects FFA homeostasis and insulin sensitivity

Finally, we sought to elucidate the *in vivo* mechanism by which rs4684847 might confer T2D risk. Analyzing hASCs isolated from BMI-matched subjects revealed a strong inverse correlation of *PRRX1* and *PPARG2* mRNA levels in homozygous but not in heterozygous risk allele carriers ($\beta=-0.815$, $p=1.4 \times 10^{-8}$). PRRX1 knockdown was sufficient to restore the risk allelic *PPARG2* mRNA suppression ($p=3.3 \times 10^{-15}$, Figure 5A), with no effect on *PPARG1* (Table 1). These data implicate PRRX1 as the mediator of the rs4684847 risk allele effect. To inform on the cellular processes by which PRRX1 may contribute to the T2D association, we

studied the impact of PRRX1 on PPARγ-regulated genes in hASCs from homozygous rs4684847 risk allele carriers by microarray analysis. We found 2,258 transcripts regulated by PRRX1 knockdown ($q < 0.2$), 336 of which were reversely regulated by concomitant PPARG knock-down (Figure 5B). Gene Set Enrichment Analysis (GSEA) highlighted an enrichment of those anti-regulated genes among the most differentially expressed genes after PRRX1 knockdown (Figure 5C), revealing that PPARγ2 mediated the primary PRRX1 effect on global gene expression. Ingenuity Pathway Analysis (IPA) showed the strongest enrichment for lipid metabolism ($p=2.81 \times 10^{-14}$) followed by adipose tissue function, glucose homeostasis, nutritional disease and insulin resistance (Figure 5D). Accordingly, an inverse relationship between PRRX1 and adipocyte triglyceride (TG) accumulation was observed in PRRX1-overexpressing SGBS adipocytes (Figure 5F).

Next, by qPCR we confirmed rs4684847 allele-dependent dysregulation of genes in those biological pathways. Notably, the gene with the strongest risk allele-dependent decrease in mRNA levels was *PEPCKC* (2.76-fold, $p=1.62 \times 10^{-10}$, Table 1). The top scoring IPA interaction network reinforced a central role for *PEPCKC* (Figure 5E). PEPCK-C is the enzyme controlling the first committed step of glyceroneogenesis (GNG), a crucial metabolic process in adipocytes regulating the re-esterification of free fatty acids (FFA) to TG (Ballard et al., 1967). GNG limits FFA release from adipocytes in the fasting state thereby controlling systemic FFA homeostasis and insulin sensitivity (Millward et al., 2010). In a cohort of 67 BMI- and body fat-matched obese subjects we confirmed rs4684847 risk allele association with increased serum FFAs levels ($p=0.049$) and a risk allele-dependent association of *PRRX1* mRNA with FFA levels ($p=0.015$, Table 2). To prove that rs4684847, by determining PRRX1 binding, affects GNG and subsequent FFA release, we monitored pyruvate incorporation in TG (Ballard et al., 1967). We confirmed a PRRX1-dependent suppression of GNG in homozygous risk allele carriers, marked by a robust correlation with PRRX1 mRNA levels (Figure 5G) and a risk allele-dependent increase of FFA release (Figure 5H). In the same

samples we show risk allele-dependent resistance to insulin-stimulated 2-deoxyglucose (2DG) uptake, with PRRX1 knockdown being sufficient to prevent cellular insulin resistance (Figure 5I). Importantly, Rosiglitazone (Rosi), a synthetic ligand of PPARγ2 (Lehmann et al., 1995), pharmacologically promotes insulin sensitivity largely via control of FFA homeostasis through GNG (Cadoudal et al., 2007), and (Kang et al., 2005) reported an impaired therapeutic Rosi-response in *PPARG* risk haplotype carriers. In our analysis of GNG in hASCs we observed an impaired response to Rosi-mediated suppression of FFA release in homozygous relative to heterozygous risk allele carriers (Figure 5J). Strikingly, PRRX1 silencing in homozygous risk-allele patient samples was sufficient to abolish the reduced Rosi responsiveness, making PRRX1 a potential target for pharmacological genotype-specific T2D intervention.

In GWAS the *PPARG* risk genotype associates with increased BMI, increased fasting insulin, and decreased insulin sensitivity (Deeb et al., 1998; Voight et al., 2010). Further supporting the dependency on PRRX1 *in vivo*, we found significant associations of *PRRX1* mRNA levels with BMI and insulin resistance, assessed by TG/HDL ratio and HOMA-IR (Table 2). Importantly, the BMI-adjusted significant association with HOMA-IR strongly depended on the rs4684847 risk allele. We confirmed this genotype-dependent link by highly sensitive euglycemic hyperinsulinemic clamp studies measuring the glucose infusion rate (GIR) in the cohort of 67 BMI- and body fat-matched patients (Table 2, Figure S4I).

In summary, the specific homeobox TFBS clustering at T2D risk SNPs inferred at the genetic level unveiled a novel role of *PRRX1* as a repressor of *PPARG2*. We establish the *cis*-regulatory SNP rs4684847 as a determinant of *PPARG2* mRNA expression by changing *PRRX1* binding to its complex regulatory region, thereby provoking dysregulation of FFA turnover and insulin sensitivity (Figure 5K).

DISCUSSION

We have developed a bioinformatics approach, PMCA, which enables the extraction of *cis*-regulatory variants that may mechanistically contribute to human disease, by dysregulation of gene expression. In line with our approach to exploit conservation in terms of co-occurring TFBS patterns, (Visel et al., 2013) has recently shown that combination of TFBSs, rather than single TFBS, via combinatorial TF binding governs spatial enhancer activity in the developing telencaephalon. Further, tissue-specific enhancers were recently accurately detected by *in vivo* mapping of the enhancer-associated proteins p300, in addition to comparative genomics approaches (Visel et al., 2009a; Blow et al., 2010).

Using T2D as a showcase we demonstrate PMCA's utility in the generic prediction of specific homeobox TFBSs at T2D risk SNPs, which is important for understanding disease regulatory circuits when we consider that interactions in a regulatory network involve numerous genes and a rather small set of TFs (Califano et al., 2012). Pursuing the results emerging from our comprehensive T2D analysis, we show that identification of the *cis*-regulatory variant rs4684847 at the *PPARG* locus enabled linking the molecular upstream factor *PRRX1* to aberrant downstream mechanisms of impaired lipid handling and insulin sensitivity, explaining the GWAS association with T2D. Notably, *PRRX1* was recently implicated in adipogenesis (Du et al., 2013), yet the regulated genes remain elusive.

Here, we restricted the analysis to SNPs in LD with GWAS tagSNPs. However, the approach could be applied to any other kind of variability, such as somatic mutations in cancer, without loss of generality. Certain issues will require consideration, e.g. analyzing genomes of closely related species to refine scoring criteria and extending our analysis to whole genome sequencing studies including rare variant information, should further inform on the genetic underpinnings of phenotypic diversity in humans. Our *in silico* scoring results predict varying numbers of regulatory SNPs per LD block. Studies have now found evidence

for allelic heterogeneity (Maller et al., 2012; Schaub et al., 2012), yet the number of causal variants within a disease locus is elusive. We propose an integrative framework where computational TFBS modularity analysis may be synergistically combined with functional genomics and population genetics data.

In sum, our results demonstrate that the extension of sequence analysis to functional conservation integrates biological information with statistical signals, and our novel methodology should help clarify the role of inherited and somatic variability in altering gene regulatory networks, in both Mendelian and common human diseases.

Definition of LD blocks

SNPs in close LD ($r^2 \geq 0.7$) to GWAS tagSNPs (references in Tables S1,5,7,8) from 1000G Pilot 1 CEU data. For details see Extended Experimental Procedures.

Phylogenetic Module Complexity Analysis

Our bioinformatics method analyzes the presence of complex patterns of evolutionarily conserved TFBSs in a CRM, within genomic regions surrounding a SNP to predict its *cis*-regulatory functionality. The method is presented in results and Figure 1, a detailed description in the Extended Experimental Procedures.

Positional Bias Analysis

Genomic regions (SNP±500bp) were scanned for presence of TFBS family matches at SNP position, and positional bias of TFBS families was calculated using overlapping 50bp sliding windows in steps of 10bp. Positional bias (P) was calculated as binominal P value for each TFBS family and each window. For details see Extended Experimental Procedures.

Correlation with evolutionary constraint, DHSseq and ChIPseq regions

Genomic regions (SNP±500bp) were correlated to constrained regions or DHSseq and ChIPseq peaks. From midpoint of constrained regions (±500bp) as anchor, the overlapping

positions (correlation) with complex/non-complex regions were counted, and plotted vs. position relative to anchor. From complex and non-complex regions with SNP (±500bp) as anchor, the overlapping positions of DHSseq and ChIPseq regions (correlation) with complex/non-complex regions were counted and plotted vs. position relative to anchor. For details see Extended Experimental Procedures.

Primary human tissue and hASC

Human islets and adipose tissue were obtained with informed consent from each subject. The studies were approved by the local ethics committees. Primary hASCs (adipose-derived stem cells) were isolated from subcutaneous adipose tissue and differentiated *in vitro*. Genotyping was done by MassARRAY (Sequenom), Omni express (Illumina) or Sanger Sequencing. For details see Extended Experimental Procedures.

RNA Preparation and Expression Analysis

Total RNA was prepared by TRIzol (Invitrogen) or RNeasy Lipid Tissue Mini Kit (Qiagen), and gene expression was measured by qRT-PCR or microarrays (Affymetrix, Illumina). Allele-specific primer extension was performed with SNaPshotKit (ABI Prism). For details see Extended Experimental Procedures.

Cell Culture and Reporter Assays

Huh7, INS-1, 293T, C2C12, 3T3-L1 and SGBS cells were cultured using standard protocols. Genomic sequences surrounding SNPs were synthesized (MWG), cloned into pGL4.22 (Promega) with TK-promoter and transfected into cells (with renilla-luciferase for normalization) by Lipofectamine 2000 (Invitrogen), and luciferase activity was measured by LuminoscanAscent (Thermo). For details see Extended Experimental Procedures.

Gene knockdown by siRNA

All knockdowns were performed with ON-TARGETplus SMARTpool siRNA (Dharmacon) and HiPerFect (Qiagen). For details see Extended Experimental Procedures.

CRISPR/Cas genome editing

HDR genome editing in human SGBS preadipocytes by transfection of CRISPR/Cas9- and sgRNA (single-guide RNA targeting a NGG PAM sequence 5' of rs4684847) expression vectors (R. Kühn, München) and rs4684847 DNA donor vectors (T-allele to replace endogenous allele, C-allele control). Cell enrichment by MACS selected transfected cell selection kit (Miltenyi). rs4684847 sequence confirmed by sanger sequencing. For details see Extended Experimental Procedures.

EMSA

42bp allelic Cy5-labeled-DNAs (MWG) and nuclear protein were used for EMSA. Supershift experiments with αPRRX1 (M. Kern) or IgG control, competition with excess unlabeled probe, protein from pCMV-PRRX1-flag transfected 293T. For details and primers see the Extended Experimental Procedures.

DNA-Protein affinity chromatography, LC-MS/MS

DNA-protein affinity chromatography with streptavidin magnetic beads (Invitrogen) and allelic biotinylated DNA-probes (MWG) and Ultimate3000nano HPLC (Dionex) LC-MS/MS coupled to LTQ OrbitrapXL (Thermo Fisher Scientific). Data analysis with Progenesis software v2.5. For details see Extended Experimental Procedures.

Statistical Analysis

Statistical analyses were done using Graph Pad Prism v5.02, Pearl or R Software v2.14.2. For details see figure legends and Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, 15 tables, 5 figures and can be found with this article online.

ACKNOWLEDGEMENTS

The authors declare competing financial interests: details see full-text HTML version of the paper. Funding, acknowledgement and author contributions see Supplemental Information.

FIGURE LEGENDS

Figure 1. Discovery of *cis*-regulatory diabetes SNPs.

(A) Workflow of the PMCA methodology: (1) The flanking region of the nc SNP was extracted from the human reference genome; (2) orthologous regions were searched in the genomes of 15 vertebrate species; (3) TFBSs were identified in each orthologous sequence; (4) TFBS modules were identified in the set of orthologous sequences (TFBS modules defined as all, two or more TFBSs occurring in the same order and in certain distance range in all or a subset of the orthologous sequences); (5) phylogenetically conserved TFBS Ω_{TFBS}, TFBS modules $\Omega_{modules}$, and occurrences of TFBSs in TFBS modules $\Omega_{TFBS\_in\_module}$ were counted; (6) repeated counting for different numbers of input-sequences weighs the degree of cross species conservation and the number of TFBS in modules. Computation of conserved TFBS with more restricted parameters Ω_{restr_TFBS} accounts for genomic regions with low numbers of orthologs; (7) steps 3-6 were repeated using randomized input sequences (randomization of sequences is done using local shuffling in order to conserve local nucleotide frequency distributions) to estimate; (8) the probability p-est of observing a given Ω_{TFBS}, Ω_{restr_TFBS}, $\Omega_{modules}$, and $\Omega_{TFBS\_in\_modules}$ and to calculate the overall scoring criterion

and finally; (9) input sequences were categorized as complex and non-complex regions. Details see Extended Experimental Procedures.

(B) Representative complex region (rs4684847) and non-complex region (rs17036342). Conserved TFBS modules occurring in more than 2 vertebrate species are shown to illustrate conserved TFBS modularity.

(C-G) Classification of candidate SNP regions at eight T2D risk loci ($r^2 \geq 0.7$) in complex and non-complex regions. Box-Whisker plots (IQR 50%) show distributions for Ω_{TFBS} (C), $\Omega_{modules}$ (D) and $\Omega_{TFBS\_in\_modules}$ (E). (Note, at 47 T2D loci we find a Median/median of 354.5/470.46 and 310/382.35 for $\Omega_{TFBS\_in\_modules}$ in complex or non-complex regions, respectively, Table S4).

(F,G) PMCA scoring illustrated for p-est$_{TFBS}$ (F) and overall score S_{all} (G). Histograms show distribution of measures to randomly observe an equal or higher measure based on calculations from the random set. *Blue curve* empirical density function of the histogram data. *Red dashed line* cut-off scores separating complex from non-complex regions (-log10 p-est$_{TFBS}$=1.12, S_{all}=6.5), regions left of this line are non-complex. *Isolated peak at the right* data points at limit of p-est calculations.

(H,I) Experimental validation of *cis*-regulatory prediction at complex regions. Non-complex regions (controls) include regions matched for TFBS density of complex regions (median=88 TFBS). The allele-dependent change in DNA-binding activity from EMSAs (n=4) (H) and reporter activity (n=10) (I) is shown for each SNP. Mean±SD, p-value from linear mixed-effects model.

See also Figure S1 and Tables S1-3

Figure 2. Correlations of *cis*-regulatory predictions with evolutionary constraint elements and functionally annotated genomic regions.

(A) The occurrences of 487 complex and 978 non-complex T2D-associated regions within constrained regions from SiPhy-π algorithm (Lindblad-Toh et al., 2011). Localization of SNPs relative to transcription start site in Figure S2I,J, Table S9.

(B) The Venn diagram illustrates number of complex and non-complex regions that directly map to a constrained element (overlap).

(C) Experimentally validated *cis*-regulatory complex regions at the *PPARG* locus (Figure 4E) lack an overlap with constrained regions. *Zoom-in*: the rs4684847 *cis*-regulatory region does not map to a constrained region (393bp upstream of nearest constrained element). A representative TFBS module ($\Omega_{TFBS\_in\_module}$=3) is shown and its TFBS module conservation for a given quorum of five species is visualized by a sequence logo. The TFBS module harbors the PRRX1 homeobox TFBS matrix, member the of T2D-distinct CART family (Figure 3A,B) and involved in regulation of endogenous *PPARG2* expression (Figures 4,5).

(D,E) The occurrences of 487 complex and 978 non-complex T2D-associated SNP regions in vicinity to DHSseq (D) and ChIPseq (E) peaks is shown. T2D complex regions were significantly enriched for overlaps to DHSseq and ChIPseq regions (correlations with Crohn's-associated regions in Figure S5, Table S10).

See also Figure S2 and Tables S9,10

Figure 3. Positional bias of distinct homeobox TFBS families at T2D risk SNPs

(A-F) Distribution of TFBS matrices relative to SNP position (SNP±500bp) at eight T2D (A,F), 47 T2D (B,C) and eight asthma (D,E) risk loci, calculated from TFBS match occurrence for 192 TFBS families (sliding 50bp windows, binomial distribution model).

(A-C) Distribution of TFBS matrices relative to SNP position within complex and non-complex regions in a set of eight (A) and the set of 47 T2D loci (B,C). Positional bias profiles for TFBS family distribution matching the criteria of central SNP position (±20bp, grey dashed lines) and –log10(P)>6 reveals clustering of distinct homeobox TFBS matrix families within T2D complex regions at SNP position (including CART and PDX1). All TFBS families displayed equal distributions within T2D non-complex regions, represented by a subset of TFBS families (C).E) Distribution of TFBS matrices relative to SNP position within complex regions for asthma loci identifies specific clustering of the EGRF matrix family (E) as opposed to T2D loci (F). No positional bias of identified distinct T2D and Crohn´s disease TFBS families (Figure S3) was found in complex asthma regions (D).

See also Figures S3 and Tables 11-15

Figure 4. Genotype-dependent down-regulation of *PPARG2* mRNA by the homeobox TF PRRX1, inferred from the predicted *cis*-regulatory variant rs4684847

(A) LD plot of the *PPARG* locus. *Diamonds* show the tagSNP Pro12Ala and pair-wise correlation of SNPs in LD (MAF ≥ 1%) against genomic position. *Red lines* predicted *cis*-regulatory SNPs; *blue PPARG* gene and exons, *zoom-in* human *PPARG* gene, *PPARG1-3* mRNA isoforms (*coding exons* boxes; *untranslated exons* dashed boxes; *introns* lines; *promoters*: arrows).

(B-D) Genotype-dependent mRNA expression in undifferentiated hASCs genotyped for Pro12Ala and rs4684847 (r^2=1.0). qRT-PCR of *PPARG1* and *PPARG2* mRNA isoforms (standardized to HPRT, Hypoxanthin-Guanin-Phosphoribosyltransferase) in homozygous (n=9) and heterozygous risk allele carriers (n=5) normalized to mean in homozygotes (B,C). Allele-specific primer extension analysis in heterozygous subjects (n=6) normalized to mean risk allele levels (D). Mean±SD, p-values from Man Whitney U test.

(E) Validation of *cis*-regulatory predictions for complex regions (*red*) at the *PPARG* locus (*black* non-complex regions as control). Quantified change in reporter activity (comparing risk vs. non-risk allele) in 3T3-L1 adipocytes with constructs matching the respective alleles is shown for each SNP (log2 scale), representing an activating or repressing effect of the risk-allele on transcriptional activity. Mean±SD, n=3-14, p-values from paired t-tests.

(F) Reporter assays with constructs harboring rs4684847 risk and non-risk allele in 3T3-L1 preadipocytes. Truncation of the PRRX matrix without affecting rs4684847 reveals abrogated allelic *cis*-regulatory activity. Mean±SD, n=9, p-values from paired t-tests.

(G) The CART matrix family member PRRX1 matches the rs4684847 (C/T) variant. TFBS modularity at the complex region surrounding rs4684847 exemplary illustrated by one conserved TFBS module comprising putative TEF, LHXF (grey) and PRRX1 binding sequences (in consistent orientation and distance range across several species).

(H,I) Increased PRRX1 binding at the risk allele in EMSAs with rs4684847 allelic probes and 3T3-L1 preadipocyte nuclear extracts (H), confirmed by competition with cold PRRX1 probe (I, left panel) and PRRX1 antibody shift of protein–DNA complex in 293T with ectopically expressed PRRX1 (I, right panel).

(J) Inhibition of reporter activity (normalized to pCMV control) at the rs4684847 risk allele by ectopic expression of PRRX1 in 3T3-L1 preadipocytes. Mean±SD; n=9, p-values from paired t-tests.

(K) Regulation of *PPARG2* mRNA expression in SGBS adipocytes with homozygous risk or non-risk allele introduced by CRISPR/Cas9 genome-editing approach. siPRRX1 and siNT transfected concurrent with induction of differentiation, *PPARG2* mRNA assessed by qRT-PCR, standardized to *IPO8* mRNA. Mean±SD, n=12, p-values from t-test.

See also Figures S4 and Table S16

Figure 5. Altered binding of the homeobox TF PRRX1 at the rs468487 variant in human adipose cells regulates lipid metabolism and insulin sensitivity

(A,G-J) PRRX1 silencing in hASC from BMI-matched heterozygous (n=16) and homozygous (n=32) rs4684847 risk allele carriers. siPRRX1 and siNT transfected concurrent with induction of adipogenic differentiation.

(A) rs4684847-dependent *PPARG2* and *PRRX1* mRNA levels measured by qRT-PCR (standardized to HPRT mRNA) 72 hours after induction of adipogenic differentiation. *Left panel:* Pearson's correlation in the siNT set. *Right panel:* Box-Whisker plot comparing *PPARG2* mRNA in siNT vs. siPRRX1 treated cells, p-values from t-test.

(B,C) Global gene expression profiling by Illumina microarrays (q <0.2) in hASCs from homozygous rs4684847 risk allele carriers transfected with siPRRX1 (n=9, grey dots) and co-transfected with siPRRX1 and siPPARG (n=4, red dots) 72 hours after induction of adipogenic differentiation (B). Distribution of siPRRX1/siPPARG anti-regulated genes in all regulated genes (C). FC=fold change.

(D,E) Biological pathways associated with siPRRX1/siPPARG anti-regulated genes (D) and top scoring ranked interaction network (E) from Ingenuity Pathway Analysis Knowledge Base.

(F) Oil Red O lipid staining of human SGBS cells with lentiviral-transduced overexpression of flag-tagged PRRX1 (or control vector) 12 days after induction of adipocyte differentiation. Protein expression with αflag (PRRX1) and αACTB antibodies.

(G,H) rs4684847-dependent GNG rate measured by [1-^{14}C]-pyruvate incorporation (G) and FFA-release (H) in hASCs. *Left panel (G):* Pearson's correlation in the siNT set: *Right panel* Box-Whisker plot comparing siNT vs. siPRRX1 treated cells, p-values from t-test.

(I) rs4684847-dependent increase of 2-deoxyglucose (2DG) uptake following insulin stimulation in hASCs. Box-Whisker plot comparing siNT vs. siPRRX1 treated cells. p-values from t-test.

(J) rs4684847-dependent rosiglitazone-mediated suppression of FFA-release during GNG. Pearson's correlation comparing siNT vs. siPRRX1. Mean±SD, p-values from t-test.

See also Figure S4G,H and Table 1,2.

Table 1. Genotype-PRRX1-dependent regulation of PRXX1/PPARG anti-regulated genes in hASCs.

	siNT				**siPRRX1**				**siPRRX1 / siNT**			
	hetero	*homo*	*hetero/homo*		*hetero*	*homo*	*hetero/homo*		*hetero*		*homo*	
	Mean ±SD	Mean ±SD	FC	p	Mean ±SD	Mean ±SD	FC	p	FC	p	FC	p
PRRX1	0.52 ±0.18	0.51 ±0.19	1.01	0.92	0.11 ±0.05	0.12 ±0.06	0.90	0.56	0.25	**2.83×10^{-7}**	0.22	**4.02×10^{-8}**
PPARG2	4.32 ±1.07	0.79 ±0.08	0.18	**2.46×10^{-11}**	4.34 ±1.47	3.37 ±1.04	0.77	0.08	1.00	0.96	4.29	**7.24×10^{-11}**
PPARG1	1.07 ±0.26	1.04 ±0.33	1.03	0.79	1.18 ±0.35	1.20 ±0.49	0.98	0.90	1.15	0.35	1.10	0.41
PEPCKC	2.83 ±0.58	1.03 ±0.20	2.76	**1.62×10^{-10}**	2.66 ±0.50	2.98 ±0.42	0.89	0.09	0.94	0.43	2.90	**8.77×10^{-4}**
PDK4	2.01 ±0.88	0.74 ±0.18	2.73	**3.19×10^{-5}**	2.00 ±0.60	1.73 ±0.61	1.15	0.27	0.99	0.97	2.35	**8.01×10^{-6}**
LIPE	1.37 ±0.64	0.68 ±0.32	2.01	**2.00×10^{-3}**	1.30 ±0.32	1.21 ±0.45	1.08	0.56	0.95	0.74	1.77	**2.03×10^{-3}**
ADIPOQ	1.89 ±0.32	0.95 ±0.31	1.98	**7.92×10^{-8}**	1.85 ±0.44	1.75 ±0.61	1.05	0.66	0.98	0.81	1.84	**2.84×10^{-4}**
OPG	0.78 ±0.36	1.67 ±0.53	0.47	**3.91×10^{-5}**	0.84 ±0.28	1.09 ±0.38	0.77	0.07	1.08	0.61	0.65	**4.10×10^{-3}**
TIMP3	0.61 ±0.21	1.50 ±0.52	0.41	**6.45×10^{-6}**	0.83 ±0.33	1.00 ±0.39	0.83	0.23	1.36	0.06	0.67	**0.01**
BBOX1	2.16 ±0.48	0.96 ±0.30	2.26	**8.04×10^{-8}**	1.84 ±0.37	2.14 ±0.44	0.86	0.07	0.85	0.07	2.23	**3.09×10^{-8}**
GLUT4	1.57 ±0.35	0.99 ±0.24	1.58	**6.15×10^{-5}**	1.62 ±	1.50 ±0.31	1.09	0.26	1.03	0.67	1.50	**1.08×10^{-4}**
THRSP	0.99 ±0.28	1.61 ±0.39	0.61	**8.18×10^{-5}**	1.53 ±0.33	1.60 ±0.32	0.95	0.57	1.55	**1.38×10^{-4}**	0.99	0.93

PRRX1/PPARG anti-regulated genes were identified by Illumina microarray analysis in samples with PRRX1 knockdown and simultaneous PRRX1 and PPARG knockdown during adipogenic differentiation (Figure 5E). Confirmatory qRT-PCR was performed for these representative top regulated genes in hASC from BMI-matched heterozygous (hetero, n = 16) and homozygous (homo, n = 32) risk-allele carriers (genotyped for the *PPARG* locus *cis*-regulatory variant rs4684847 and the tagSNP rs1801282 Pro12Ala). PRRX1, Paired-related homeobox 1; PPARG, peroxisome proliferator-activated receptor gamma; PEPCKC, Phosphoenolpyruvate carboxylase cytosolic; PDK4, pyruvate dehydrogenase kinase, isozyme 4; LIPE, lipase, hormone-sensitive; ADIPOQ, adiponectin, C1Q and collagen domain containing; OPG, Osteoprotegerin; TIMP3, TIMP metallopeptidase inhibitor 3; BBOX1, butyrobetaine (gamma), 2-oxoglutarate dioxygenase (gamma-butyrobetaine hydroxylase); GLUT4, Glucose Transporter Type 4; THRSP, thyroid hormone responsive Spot 14 Protein; FC, fold change; p, p-value from unpaired t-test.

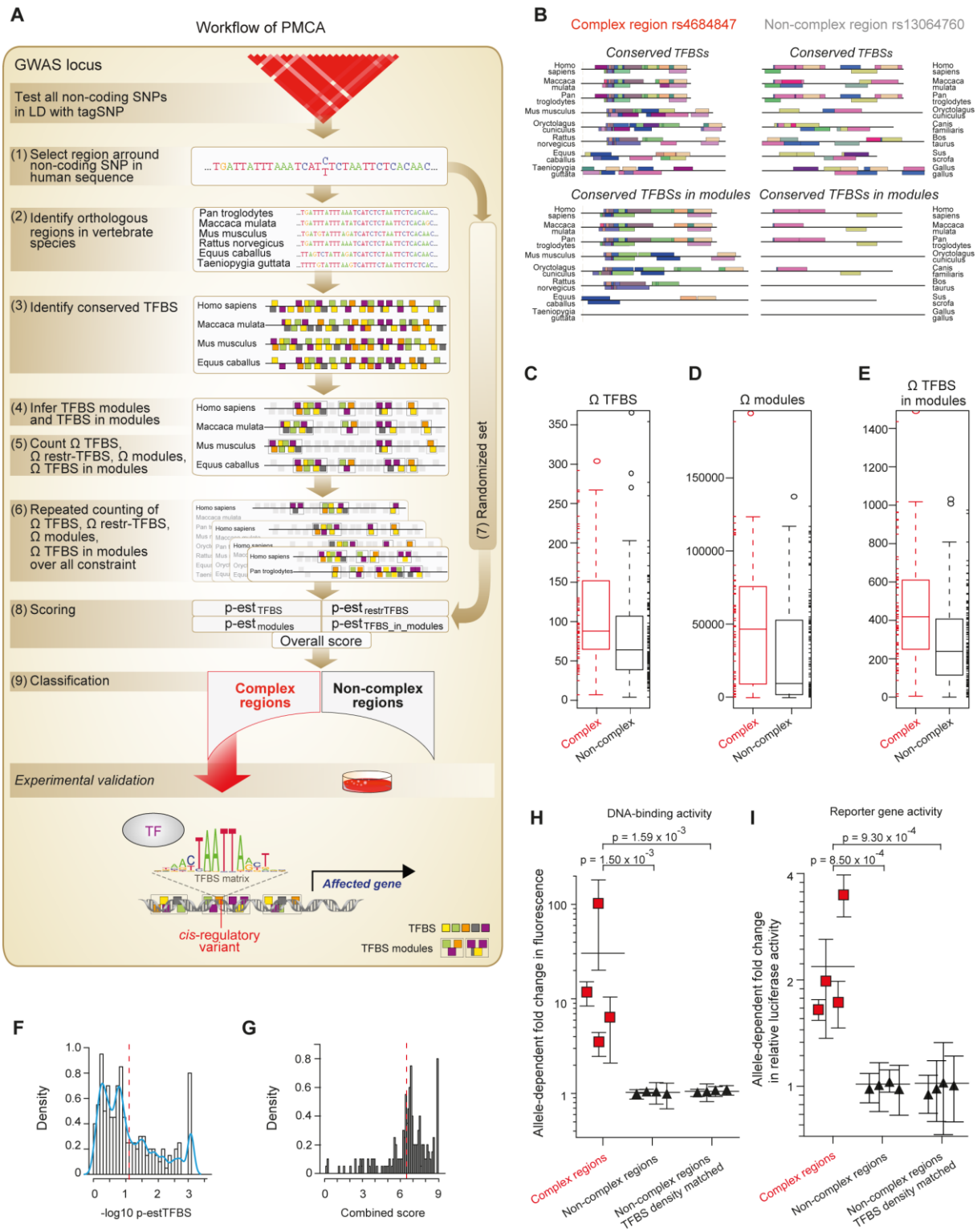
Table 2. PRRX1 mRNA expression levels in adipose tissue correlates with FFA, BMI, TG/HDL ratio and insulin resistance measures HOMA-IR and GIR.

		***PRRX1* mRNA**		***PRRX1* mRNA**		***PRRX1* mRNA**	
rs4684847 genotypes		**All**		**CC**		**CT and TT**	
		β-estimate	p-value	β-estimate	p-value	β-estimate	p-value
log(FFA)[a]	age/ BMI	0.25	**0.014**	0.27	**0.015**	-0.009	0.99
log(BMI)[b]	-	1.32	**0.05**	1.23	0.19	1.43	0.23
	age	1.45	**0.03**	1.23	0.19	1.96	0.09
log(TG/HDL)[b]	-	6.92	**7.54×10^{-4}**	6.40	**0.02**	6.35	0.07
	age	6.97	**7.36×10^{-4}**	6.14	**0.02**	6.81	0.07
	age/ BMI	4.86	**8.3×10^{-3}**	5.00	0.07	2.64	0.33
log(HOMAIR)[b]	-	2.77	**3.52×10^{-3}**	3.13	**8.3×10^{-3}**	1.80	0.29
	age	2.77	**3.77×10^{-3}**	3.12	**8.6×10^{-3}**	1.70	0.34
	age/ BMI	1.41	**0.028**	2.1	**4.6×10^{-3}**	-0.55	0.63
log(GIR)[a]	age/ BMI	-0.51	**1.83×10^{-7}**	-0.78	**3.30×10^{-8}**	-0.38	0.28

rs4684847-dependent PRRX1 adipose tissue mRNA expression data from a) patients undergoing a hyperinsulinemic euglycemic clamp measured by qRT-PCR analysis (BMI-matched study sample, risk allele n = 54, non-risk allele n = 13) and b) a cohort with PRRX1 expression data from microarrays (risk allele n = 20, non-risk allele n= 18). rs4684847 risk-allele and non-risk allele genotypes were determined by Sequenom-assay. FFA, free fatty acids; GIR, glucose infusion rate of hyperinsulinemic euglycemic clamp; BMI, body mass index; HOMA-IR, homeostasis model assessment of insulin resistance; TG, triglyceride; HDL, high density lipoprotein. p-values and β-estimates from linear regression analysis of *PRRX1* mRNA expression levels with phenotype residuals are shown.

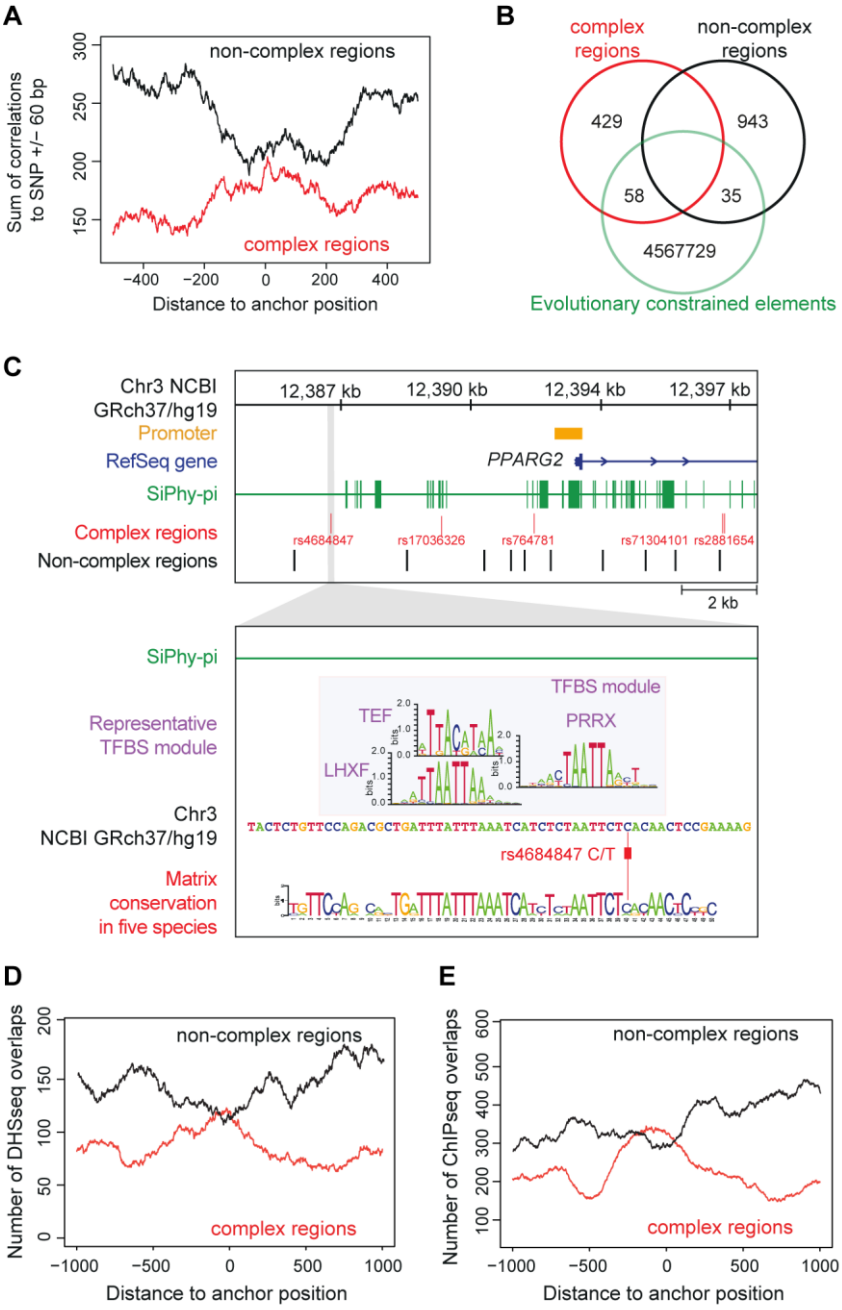
Figures

Figure 1



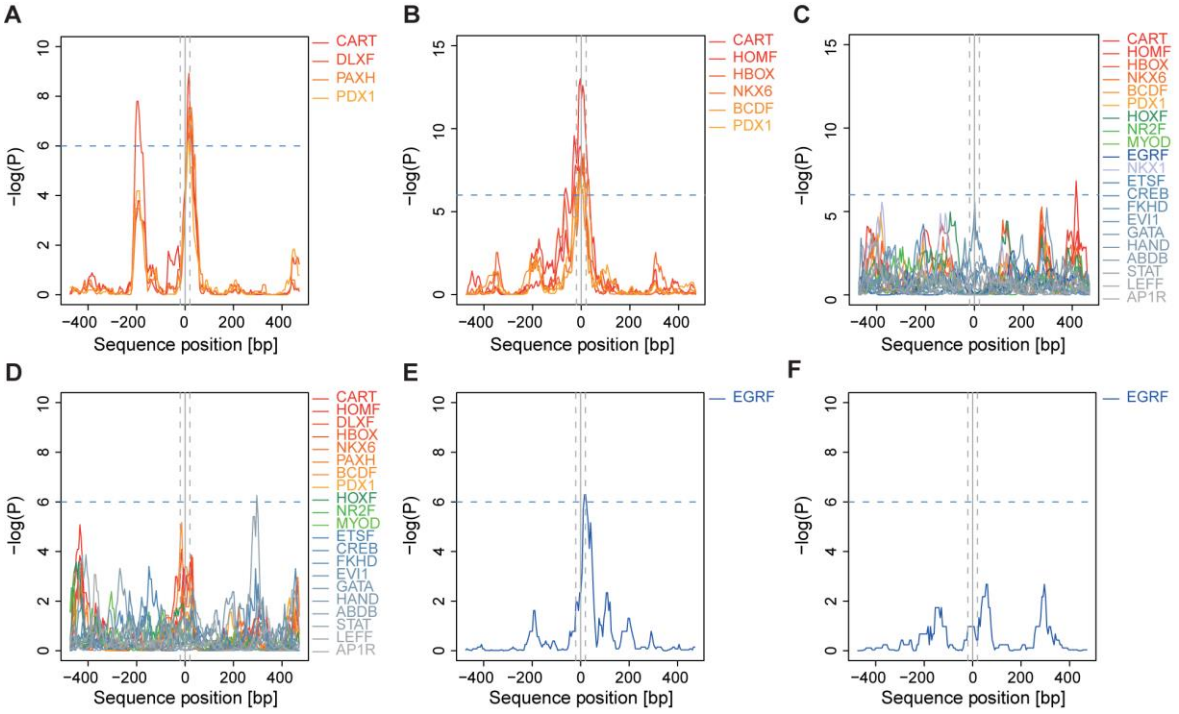
Claussnitzer et al, Figure 1

Figure 2



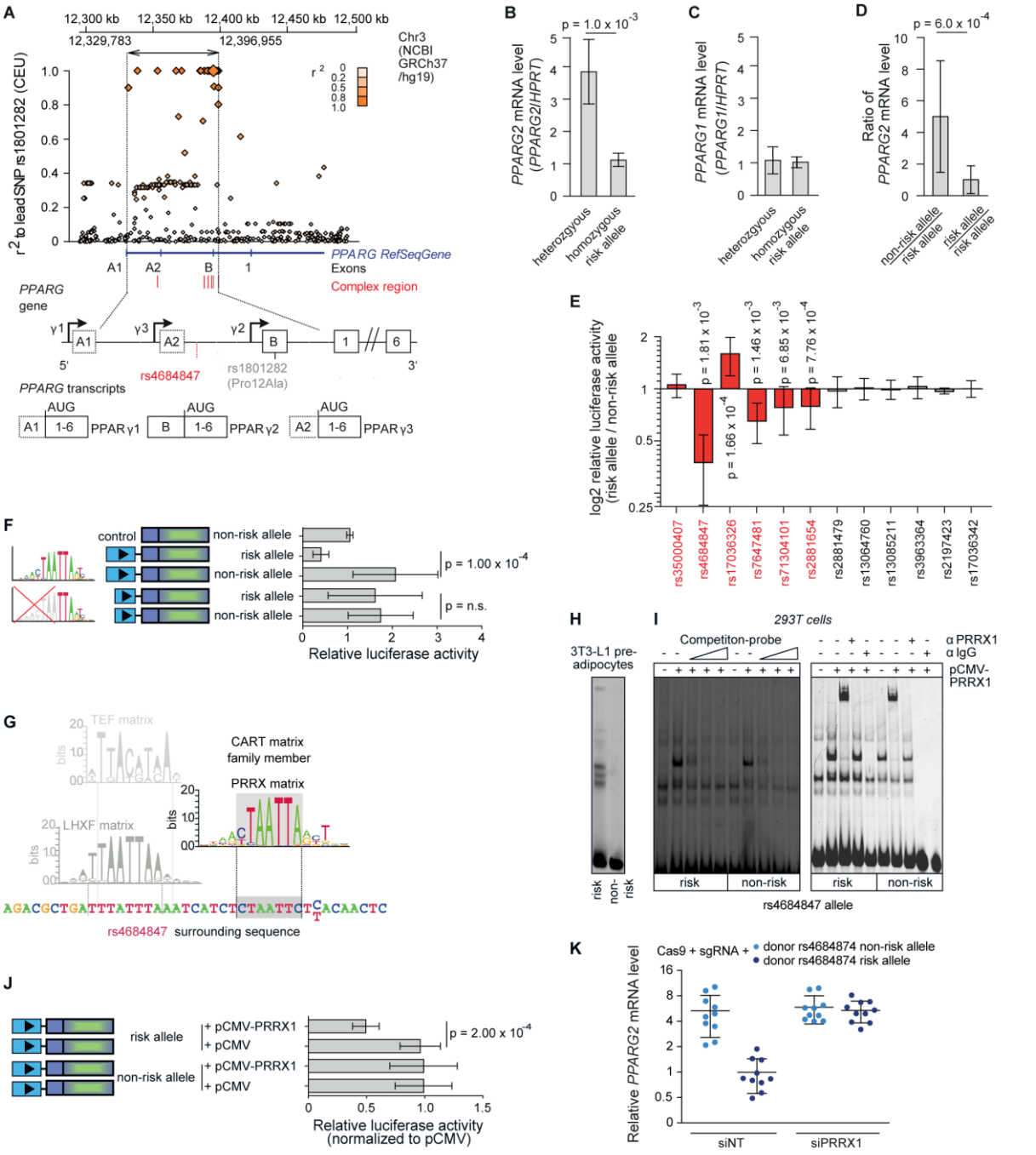
Claussnitzer et al, Figure 2

Figure 3



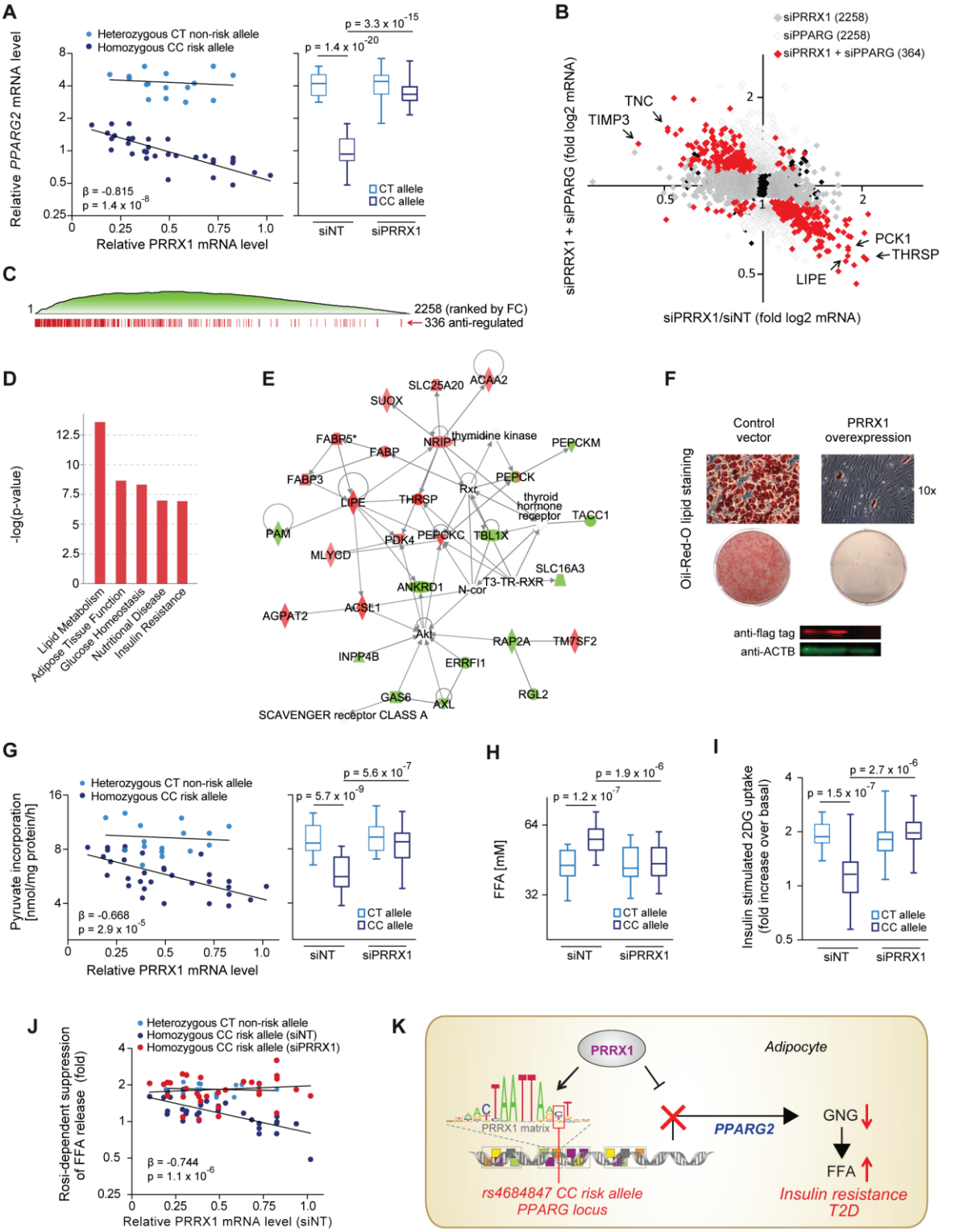
Claussnitzer et al, Figure 3

Figure 4



Claussnitzer et al, Figure 4

Figure 5



Claussnitzer et al, Figure 5

References

(2011). A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol *9*, e1001046.

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Alberobello, A., Congedo, V., Liu, H., Cochran, C., Skarulis, M., Forrest, D., and Celi, F. (2011). An intronic SNP in the thyroid hormone receptor beta gene is associated with pituitary cell-specific over-expression of a mutant thyroid hormone receptor beta2 (R338W) in the index case of pituitary-selective resistance to thyroid hormone. Journal of Translational Medicine *9*, 144.

Arnone, M.I., and Davidson, E.H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development *124*, 1851-1864.

Ballard, F.J., Hanson, R.W., and Leveille, G.A. (1967). Phosphoenolpyruvate carboxykinase and the synthesis of glyceride-glycerol from pyruvate in adipose tissue. J Biol Chem *242*, 2746-2750.

Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., and Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat Genet *42*, 806-810.

Brissova, M., Shiota, M., Nicholson, W.E., Gannon, M., Knobel, S.M., Piston, D.W., Wright, C.V.E., and Powers, A.C. (2002). Reduction in pancreatic transcription factor PDX-1 impairs glucose-stimulated insulin secretion. J Biol Chem *277*, 11225-11232.

Cadoudal, T., Blouin, J.M., Collinet, M., Fouque, F., Tan, G.D., Loizon, E., Beale, E.G., Frayn, K.N., Karpe, F., and Vidal, H., et al. (2007). Acute and selective regulation of glyceroneogenesis and cytosolic phosphoenolpyruvate carboxykinase in adipose tissue by thiazolidinediones in type 2 diabetes. Diabetologia *50*, 666-675.

Califano, A., Butte, A.J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet *44*, 841-847.

Chinwalla, A.T., Cook, L.L., Delehaunty, K.D., Fewell, G.A., Fulton, L.A., Fulton, R.S., Graves, T.A., Hillier, L.W., Mardis, E.R., and McPherson, J.D., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Cho, S.J., Kang, M.J., Homer, R.J., Kang, H.R., Zhang, X., Lee, P.J., Elias, J.A., and Lee, C.G. (2006). Role of Early Growth Response-1 (Egr-1) in Interleukin-13-induced Inflammation and Remodeling. Journal of Biological Chemistry *281*, 8161-8168.

Deeb, S.S., Fajas, L., Nemoto, M., Pihlajamaki, J., Mykkanen, L., Kuusisto, J., Laakso, M., Fujimoto, W., and Auwerx, J. (1998). A Pro12Ala substitution in PPAR[gamma]2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. Nat Genet *20*, 284-287.

Doria, A., Patti, M.-E., and Kahn, C.R. (2008). The Emerging Genetic Architecture of Type 2 Diabetes. Cell Metabolism *8*, 186-200.

Du, B., Cawthorn, W.P., Su, A., Doucette, C.R., Yao, Y., Hemati, N., Kampert, S., McCoin, C., Broome, D.T., and Rosen, C.J., et al. (2013). The Transcription Factor Paired-Related Homeobox 1 (Prrx1) Inhibits Adipogenesis by Activating Transforming Growth Factor-β (TGFβ) Signaling. Journal of Biological Chemistry *288*, 3036-3047.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A.U., Wheeler, E., Glazer, N.L., Bouatia-Naji, N., and Gloyn, A.L., et al. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet *42*, 105-116.

Fajans, S.S., Bell, G.I., and Polonsky, K.S. (2001). Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. N Engl J Med *345*, 971-980.

Fajas, L., Fruchart, J.C., and Auwerx, J. (1998). PPARgamma3 mRNA: a distinct PPARgamma mRNA subtype transcribed from an independent promoter. FEBS Lett *438*, 55-60.

FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A., and Vinson, C. (2004). Clustering of DNA Sequences in Human Promoters. Genome Research *14*, 1562-1574.

Harrison, K.A., Thaler, J., Pfaff, S.L., Gu, H., and Kehrl, J.H. (1999). Pancreas dorsal lobe agenesis and abnormal islets of Langerhans in Hlxb9-deficient mice. Nat Genet *23*, 71-75.

Heikkinen, S., Argmann, C., Feige, J.N., Koutnikova, H., Champy, M.-F., Dali-Youcef, N., Schadt, E.E., Laakso, M., and Auwerx, J. (2009). The Pro12Ala PPARγ2 Variant Determines Metabolism at the Gene-Environment Interface. Cell Metabolism *9*, 88-98.

Hindorff LA, M.J.W.A.J.H.H.P.K.A.a.M.T. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [date of access].

Jonsson, J., Carlsson, L., Edlund, T., and Edlund, H. (1994). Insulin-promoter-factor 1 is required for pancreas development in mice. Nature *371*, 606-609.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.**?**., Birney, E., and Furlong, E.M. (2012). A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History. Cell *148*, 473-486.

Kang, E.S., Park, S.Y., Kim, H.J., Kim, C.S., Ahn, C.W., Cha, B.S., Lim, S.K., Nam, C.M., and Lee, H.C. (2005). Effects of Pro12Ala polymorphism of peroxisome proliferator-activated receptor gamma2 gene on rosiglitazone response in type 2 diabetes. Clin Pharmacol Ther *78*, 202-208.

Laumen, H., Saningong, A.D., Heid, I.M., Hess, J., Herder, C., Claussnitzer, M., Baumert, J., Lamina, C., Rathmann, W., and Sedlmeier, E.-M., et al. (2009). Functional Characterization of Promoter Variants of the Adiponectin Gene Complemented by Epidemiological Data. Diabetes *58*, 984-991.

Lehmann, J.M., Moore, L.B., Smith-Oliver, T.A., Wilkison, W.O., Willson, T.M., and Kliewer, S.A. (1995). An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma). J Biol Chem *270*, 12953-12956.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., and Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476-482.

Maller, J.B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J.M.M., Auton, A., Myers, S., and Morris, A., et al. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. Nat Genet *44*, 1294-1301.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., and Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190-1195.

Mikkelsen, T.S., Xu, Z., Zhang, X., Wang, L., Gimble, J.M., Lander, E.S., and Rosen, E.D. (2010). Comparative Epigenomic Analysis of Murine and Human Adipogenesis. Cell *143*, 156-169.

Millward, C.A., DeSantis, D., Hsieh, C.W., Heaney, J.D., Pisano, S., Olswang, Y., Reshef, L., Beidelschies, M., Puchowicz, M., and Croniger, C.M. (2010). Phosphoenolpyruvate carboxykinase (Pck1) helps regulate the triglyceride/fatty acid cycle and development of insulin resistance in mice. The Journal of Lipid Research *51*, 1452-1463.

Moffatt, M.F., Gut, I.G., Demenais, F., Strachan, D.P., Bouzigon, E., Heath, S., Mutius, E. von, Farrall, M., Lathrop, M., and Cookson, W.O.C.M. (2010). A Large-Scale, Consortium-Based Genomewide Association Study of Asthma. New England Journal of Medicine. N Engl J Med *363*, 1211-1221.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., and Ruda, V.M., et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus *466*, 714-719.

Nekrep, N., Wang, J., Miyatsuka, T., and German, M.S. (2008). Signals from the neural crest regulate beta-cell mass in the pancreas. Development *135*, 2151-2160.

Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations. PLoS Genet *6*, e1000895

EP -.

Ozaki, K., Sato, H., Iida, A., Mizuno, H., Nakamura, T., Miyamoto, Y., Takahashi, A., Tsunoda, T., Ikegawa, S., and Kamatani, N., et al. (2006). A functional SNP in PSMA6 confers risk of myocardial infarction in the Japanese population. Nat Genet *38*, 921-925.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., and Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499-502.

Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., and Davis, M., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet *41*, 882-884.

Post, S.M., Quintás-Cardama, A., Pant, V., Iwakuma, T., Hamir, A., Jackson, J.G., Maccio, D.R., Bond, G.L., Johnson, D.G., and Levine, A.J., et al. (2010). A High-Frequency Regulatory Polymorphism in the p53 Pathway Accelerates Tumor Development. Cancer Cell *18*, 220-230.

Rosen, E.D., Sarraf, P., Troy, A.E., Bradwin, G., Moore, K., Milstone, D.S., Spiegelman, B.M., and Mortensen, R.M. (1999). PPARγ Is Required for the Differentiation of Adipose Tissue In Vivo and In Vitro. Molecular Cell *4*, 611-617.

Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. Genome Research *22*, 1748-1759.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., and Hadjadj, S., et al. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature *445*, 881-885.

Stitzel, M.L., Sethupathy, P., Pearson, D.S., Chines, P.S., Song, L., Erdos, M.R., Welch, R., Parker, S.C., Boyle, A.P., and Scott, L.J., et al. (2010). Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci. Cell Metab *12*, 443-455.

The ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Tontonoz, P., Hu, E., and Spiegelman, B.M. (1994). Stimulation of adipogenesis in fibroblasts by PPARγ2, a lipid-activated transcription factor. Cell *79*, 1147-1156.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., and Chen, F., et al. (2009a). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854-858.

Visel, A., Rubin, E.M., and Pennacchio, L.A. (2009b). Genomic views of distant-acting enhancers. Nature *461*, 199-205.

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R.V., McKinsey, G.L., Pattabiraman, K., Silberberg, S.N., and Blow, M.J., et al. (2013). A High-Resolution Enhancer Atlas of the Developing Telencephalon. Cell *152*, 895-908.

Voight, B.F., Scott, L.J., Steinthorsdottir, V., Morris, A.P., Dina, C., Welch, R.P., Zeggini, E., Huth, C., Aulchenko, Y.S., and Thorleifsson, G., et al. (2010). Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet *42*, 579-589.

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. (2013). One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. Cell *153*, 910-918.

Ward, L.D., and Kellis, M. (2012a). Evidence of Abundant Purifying Selection in Humans for Recently Acquired Regulatory Functions. Science *337*, 1675-1678.

Ward, L.D., and Kellis, M. (2012b). Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotech *30*, 1095-1106.

Zhang, J., Fu, M., Cui, T., Xiong, C., Xu, K., Zhong, W., Xiao, Y., Floyd, D., Liang, J., and Li, E., et al. (2004). Selective disruption of PPARγ2 impairs the development of adipose tissue and insulin sensitivity. Proceedings of the National Academy of Sciences of the United States of America *101*, 10703-10708.

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E.M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature *462*, 65-70.