

Overview of Supplemental Information

SUPPLEMENTAL FIGURES

Supplemental Figure 1. Linkage disequilibrium (LD) block structure at the T2D susceptibility loci included into the proof of concept analysis and cell-type specific cis-regulatory effects of complex SNPs regions at T2D loci. Related to Figure 1.

Supplemental Figure 2. Performance of PMCA for candidate SNPs at T2D, Asthma and Crohn's disease susceptibility loci. Correlations of cis-regulatory predictions from PMCA at Crohn's disease susceptibility loci with evolutionary constraint elements and functionally annotated genomic regions. Frequency distribution and distance to transcriptional start site of predicted cis-regulatory, complex SNP regions. Related to Figure 2-3.

Supplemental Figure 3. Positional bias analysis of TFBS matrices at complex and non-complex SNP regions AND RNAseq AND siRNA data in INS-1 cells. Related to Figure 3.

Supplemental Figure 4 Computational predicted cis-regulatory variants at the PPARG T2D risk locus and the homeobox factor PRRX1 as regulator of endogenous PPAR γ 2 expression. Related to Figure 4.

SUPPLEMENTAL TABLES (see excel datasheets)

Supplemental Table 1. PMCA measures for candidate SNPs at eight T2D susceptibility loci included into the proof of concept analysis comprising 200 SNPs. Related to Figure 1.

Supplemental Table 2. PMCA measures for candidate SNPs at eight T2D susceptibility loci included into the proof of concept analysis (upper 25% of complex SNP region ranking). Related to Figure 1.

Supplemental Table 3. PMCA measures and experimental validation of *cis*-regulatory predictions for candidate SNP regions. Related to Figure 1.

Supplemental Table 4. Association of tag SNPs and predicted *cis*-regulatory SNPs with T2D in DIAGRAM v2 data (A) and glycaemic traits in MAGIC consortium meta-analysis data (B). Related to Figure 1.

Supplemental Table 5. PMCA measures for known *cis*-regulatory SNPs, associated to different traits. Related to Figure 1.

Supplemental Table 6. Central positional bias of distinct TFBS matrix families in complex SNP regions and non-complex SNP regions. Related to Figure 2.

Supplemental Table 7. PMCA measures for candidate SNPs at 47 autosomal T2D susceptibility loci comprising 1,465 SNPs. Related to Figure 2.

Supplemental Table 8. PMCA measures for candidate SNPs at asthma susceptibility loci (A) and for candidate SNPs at Crohn's disease susceptibility loci (B). Related to Figure 2.

Supplemental Table 9. Overlap of complex SNP regions and non-complex SNP regions inferred from the analyzed set of 47 T2D susceptibility loci with evolutionary constraint elements and localization to next TSS. Related to Figure 3.

Supplemental Table 10. Enrichment of functional annotation from DNase-seq and ChIP-seq peaks overlaps to complex SNP regions. Related to Figure 3.

Supplemental Table 11. Association of *cis*-regulatory predictions at PMCA selected complex regions with functional Regulome annotations. Related to Figure 3.

Supplemental Table 12. Experimental validation of PMCA predicted *cis*-regulatory variants at the *PPARG* T2D risk locus, 3p25.3. Related to Figure 4.

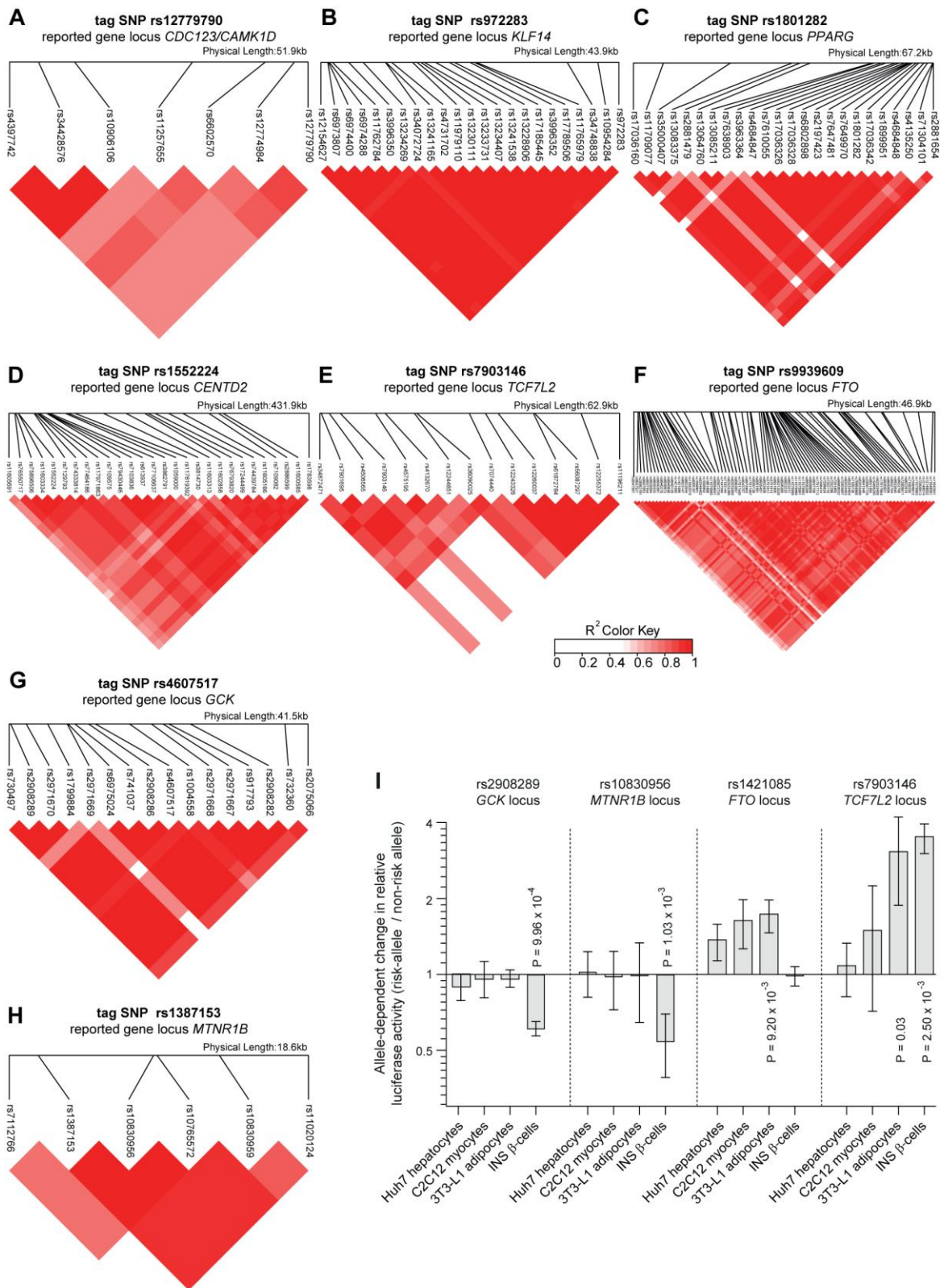
EXTENDED EXPERIMENTAL PROCEDURES

1. Definition of LD blocks
2. Search for orthologous regions
3. PMCA-Procedures: Description of the PMCA method
 - 3.1 Motivation
 - 3.2 General design of the PMCA method
 - 3.3 Detailed description of the PMCA algorithm (pseudo-code)
 - 3.4 Step-by-step example for running PMCA manually using the graphical user interface
4. Positional Bias Analysis: Calculation of the TFBS positional bias
5. Correlation of SNP regions with evolutionary constraint regions.
6. Correlation of SNP regions to DNase-seq regions and ChIP-seq regions
7. Enrichment of complex SNPs in diseases loci
8. Assessment of SNP to TSS distance annotations

8. Culture of cell lines
9. Isolation, culture and differentiation of primary human adipose stromal cells (hASC)
10. Electrophoretic mobility shift assay (EMSA)
11. DNA-Protein affinity chromatography, LC-MS/MS and label free quantification.
12. Gene knock-down by siRNA
13. Quantitative RT-PCR and allele-specific primer extension analysis
14. Glyceroneogenesis and Glucose-uptake measurement in primary hASC
15. Luciferase expression constructs
16. Luciferase expression assays
17. Genome editing of SGBS preadipocytes
18. Correlation of PRRX1 mRNA levels in primary human adipose tissue with measures of lipid-metabolism and insulin-sensitivity
19. Analysis of RNAseq Data from primary human islets
20. Genome wide expression analysis in primary human hASC
21. eQTL analysis

SUPPLEMENTAL NOTES Authors from DIAGRAM+

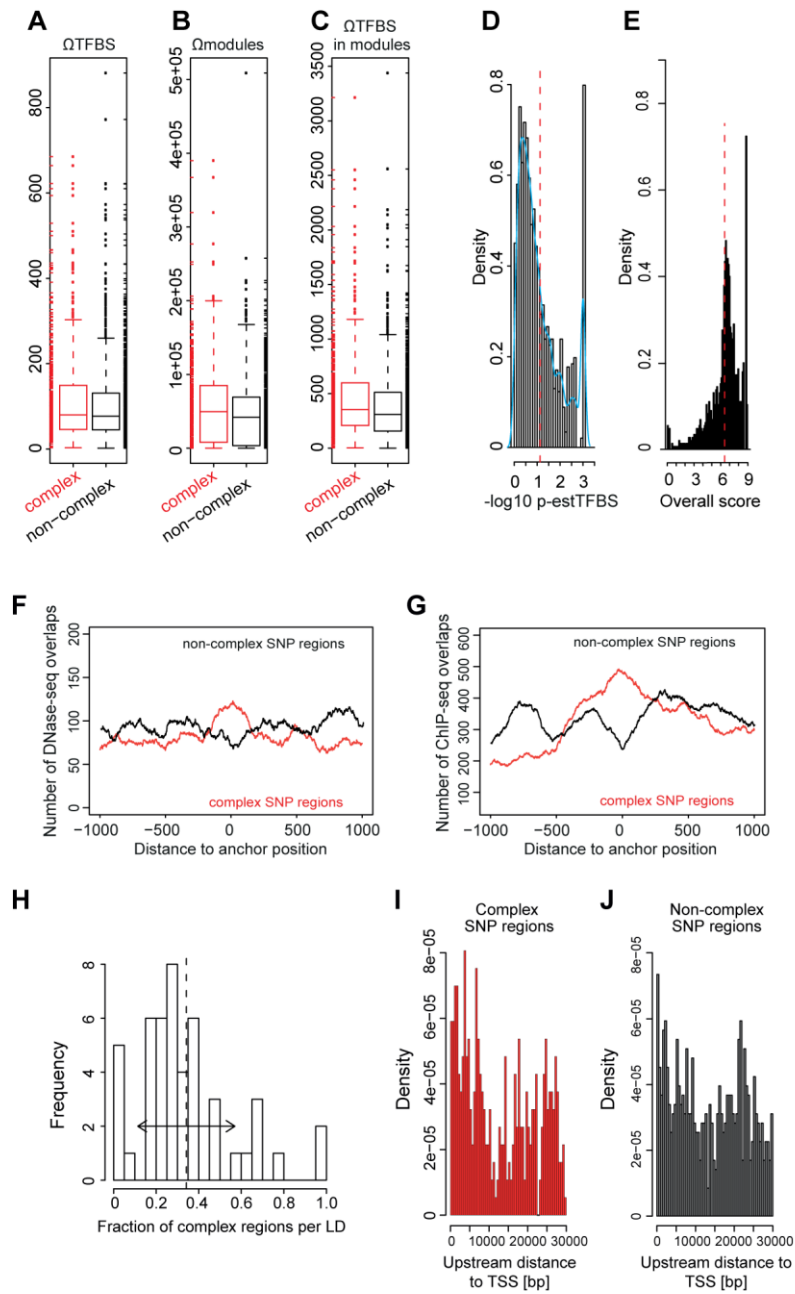
SUPPLEMENTAL REFERENCES



Supplemental Figure 1. (A-H) Linkage disequilibrium (LD) block structure at eight T2D susceptibility loci included into the proof-of-concept analysis. (I) Cell-type specific cis-regulatory effects of complex regions at T2D loci. Related to Figure 1.

(A-H) LD blocks derived from eight tag SNPs that were included in the primary PMCA analysis are shown. Pairwise LD, measured as R^2 , was calculated from 1000G Pilot 1 data CEU (1000 Genomes Project Consortium, 2010) using the SNAP viewer Tool (Johnson et al., 2008), Broad Institute. R^2 is displayed in a range of plain white ($R^2 = 0$) to red ($R^2 = 1.0$). Plots were drawn using the LDheatmap package in R version 2.15. Detailed information on the presented LD blocks is summarized in Table S1.

(I) Cell type-specific *cis*-regulatory effects of complex regions. Luciferase constructs of the respective complex regions were transfected into INS1 pancreatic β -cells (insulin secretory cell line), and differentiated 3T3-L1 adipocytes, C2C12 myocytes, and Huh7 cells (insulin responsive cell lines), respectively. The allele-dependent fold change in relative luciferase activity comparing the risk and non-risk alleles is shown for each SNP, representing an activating or repressing effect of the risk allele on transcriptional activity. Data are represented as mean \pm SD (n=9), ***p < 0.001, p-values from paired t-test.



Supplemental Figure 2. (A-E) Performance of PMCA for candidate SNPs at 47 T2D susceptibility loci; (F,G) Correlations of cis-regulatory predictions from PMCA at Crohn's disease susceptibility loci with evolutionary constraint elements and functionally annotated genomic regions; (H) Frequency distribution of complex regions; and (I,J) Distance to transcriptional start site of predicted cis-regulatory SNPs. Related to Figure 2.

(A-E) PMCA results are shown for 47 T2D susceptibility loci comprising 1,465 candidate SNP-surrounding regions ($R^2 \geq 0.7$, Table S7).

(A-C) Box-whisker plots of the numbers obtained for each classification strategy in the analysis based on sequence number constraint. Plots show the distributions for Ω_{TFBS} (A), $\Omega_{modules}$ (B) and $\Omega_{TFBS_in_modules}$ (C), including the median (horizontal bars), the interquartile region (IQR) representing the middle 50% range (boxes), extreme values (whiskers) and outliers (dots). Data points covered by the IQR and the whisker values were explicitly added as rug at the sides of the plot. The median for complex regions (highlighted in red) was higher than for the non-complex regions for each classification.

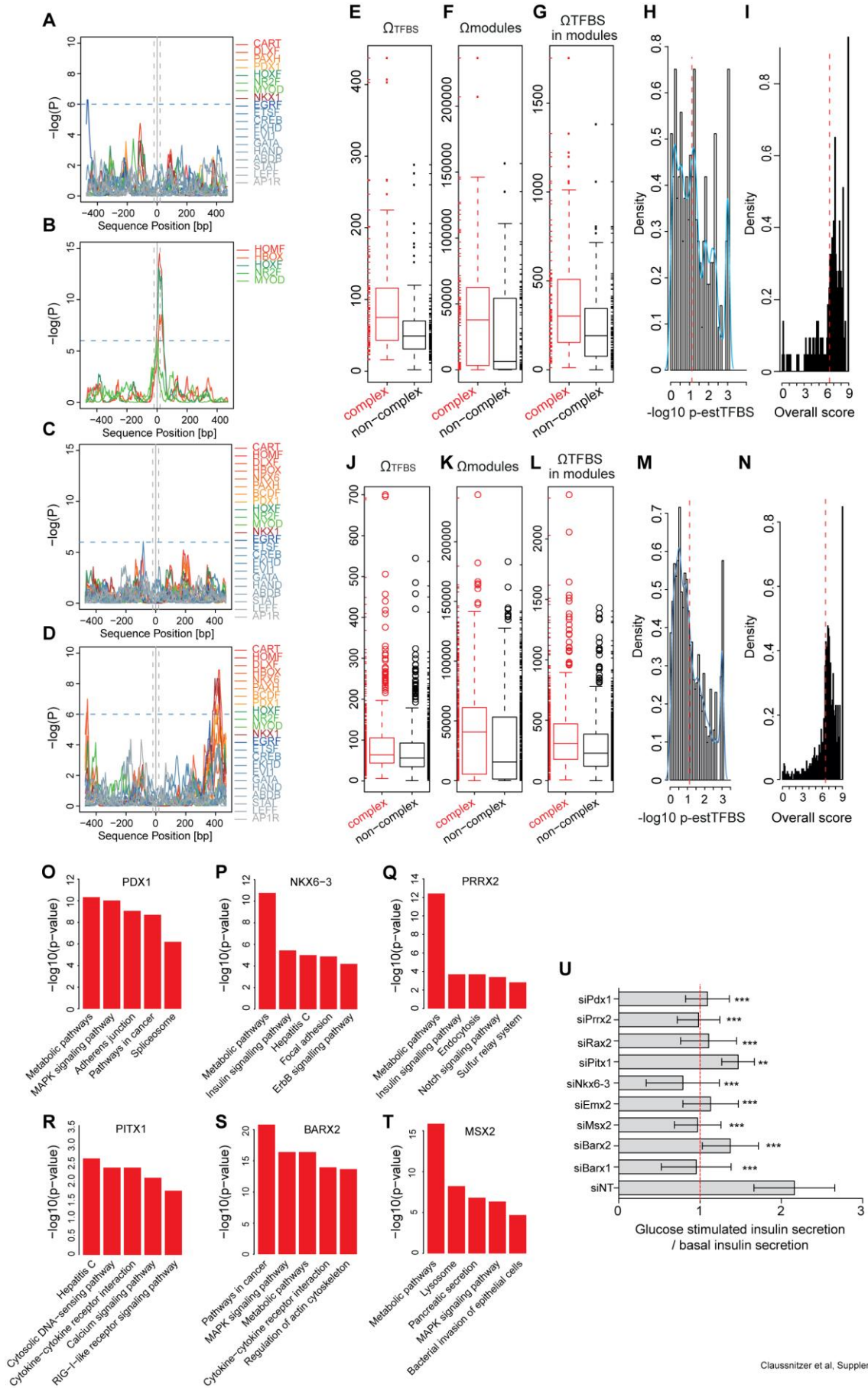
(D,E) Histograms showing the distribution of $-\log_{10}$ of the estimated probability p-est to randomly observe an equal or higher Ω_{TFBS} (D) and the distribution for an equal or higher *overall score* from all three criteria (E), as calculated from observations in the random set derived from 1,000 shuffled sequences per ortholog set. The blue curve illustrates the empirical density function of the histogram data. The red vertical dashed line indicates the cut-off scores separating complex from non-complex regions (SNP regions with a value to the left of this were defined as non-complex). The isolated peak at the right (low p-est / high overall score data) refers to data points that hit the lower limit of p-est calculations.

(F,G) Correlations of PMCA results with DNase-seq (F) and ChIP-seq (G) data for 1,218 SNPs associated with Crohn's diseases. For PMCA classification of SNP-adjacent genomic regions in complex and non-complex regions see Table S8. The occurrences of DNase-seq and ChIP-seq DNA peaks in vicinity of complex and non-complex T2D-associated SNP regions are shown (each position ± 500 bp from the SNP position of complex and non-complex regions was scanned for overlaps with DNase- or ChIP-seq peaks, see Extended Experimental Procedures). The number of complex and non-complex regions that directly overlap DNase-seq and ChIP-seq regions was determined by a comparison of their genomic

positions. Complex regions were significantly enriched for overlaps with DNase-seq and ChIP-seq regions in the set of Crohn's disease associated SNPs ($P = 4.17 \times 10^{-13}$ and $P = 3.06 \times 10^{-6}$, respectively, Fisher's exact test, see also Table S10).

(H) Frequency distribution for fractions of complex regions obtained for 47 analyzed T2D LD blocks. PMCA separates the SNPs at susceptibility loci into complex and non-complex regions. The frequency histogram (bin of LD block sizes = 0.05) displays the fractions of complex regions in the 47 analyzed T2D susceptibility LD blocks (Table S7). The frequency distribution illustrates that the number of complex regions identified per LD block spreads over a large range (median = 29 %, average = 34.2 % (vertical dashed line), SD = 22.6 (horizontal arrow)).

(I,J) Distance to transcriptional start sites (TSS) for complex and non-complex regions obtained for 47 analyzed T2D LD blocks. Density histograms show all distances (bin size 500 bp) between SNPs and transcription start sites (TSSs) (TSS annotated within 30,000 bp downstream of SNP position). The distance distribution is shown for 487 complex regions (N) and 978 non-complex regions (O) identified by PMCA within the set of 47 T2D loci (for detailed information see Table S9). The histogram shapes of (N) and (O) illustrate the equal positioning of PMCA categories (complex and non-complex regions) relative to downstream TSSs.



Clausnitzer et al, Supplemental Figure 3

Supplemental Figure 3. (A-D) Positional bias analysis of TFBS matrices at complex and non-complex regions; (E-H) Performance of PMCA for candidate SNPs at asthma and Crohn's disease susceptibility loci; and (O-T) Combinatorial framework analysis of PMCA, bias analysis and RNA-seq-identified homeobox TFs associated with metabolic processes and impaired glucose stimulated insulin secretion.

(A,C,D) No apparent positional bias of TFBS matrices at non-complex regions. Distribution of TFBS matrices relative to SNP position (denoted by grey lines) within non-complex regions at eight T2D loci (A), eight asthma loci (C), and a set of Crohn's disease susceptibility variants (D), assessed by positional bias analysis (Table S6). Positional bias was calculated from TFBS match occurrence over 1,000bp SNP regions for 192 TFBS matrix families (Genomatix Matrix Library version 8.4) within sliding 50bp windows under a binomial distribution model (detailed in Extended Experimental Procedures). Positional bias profiles are presented for a subset of analyzed TFBS matrix families including the matrix families that matched the selection criteria of central SNP position and $-\log_{10}(P) > 6$ in the complex regions (Figure 3).

(B) Positional bias of TFBS matrices at complex regions identified in a set 1,218 candidate SNPs at Crohn's diseases susceptibility loci. Distribution of TFBS matrices relative to SNP position (denoted by grey lines) within complex regions at the set of Crohn's disease variants (Table S6D), assessed by positional bias analysis. Positional bias was calculated from TFBS match occurrence over 1,000bp SNP regions for 192 TFBS matrix families (Genomatix Matrix Library version 8.4) within sliding 50bp windows under a binomial distribution model (detailed in Extended Experimental Procedures). Positional bias profiles are presented for a subset of analyzed TFBS matrix families including the matrix families which matched the selection criteria of central SNP position and $-\log_{10}(P) > 6$ in the complex regions. The positional bias analysis within complex regions reveals specific clustering at SNP position ± 20 bp (denoted by grey dashed lines) of the TFBS matrix families NR2F, MYOD, HOXF (green) and HOMF and HBOX (red, see also bias at the size matched set of T2D loci, Figure 2B).

(E-N) PMCA results are shown for asthma (E-I) and Crohn's disease (J-N) susceptibility loci ($R^2 \geq 0.7$, Table S7).

(E-G,J-L) Box-whisker plots of the numbers obtained for each classification strategy in the analysis based on sequence number constraint. Plots show the distributions for Ω_{TFBS} (A), $\Omega_{modules}$ (B) and $\Omega_{TFBS_in_modules}$ (C), including the median (horizontal bars), the interquartile region (IQR) representing the middle 50% range (boxes), extreme values (whiskers) and outliers (dots). Data points covered by the IQR and the whisker values were explicitly added as rug at the sides of the plot. The median for complex regions (highlighted in red) was higher than for the non-complex regions for each classification.

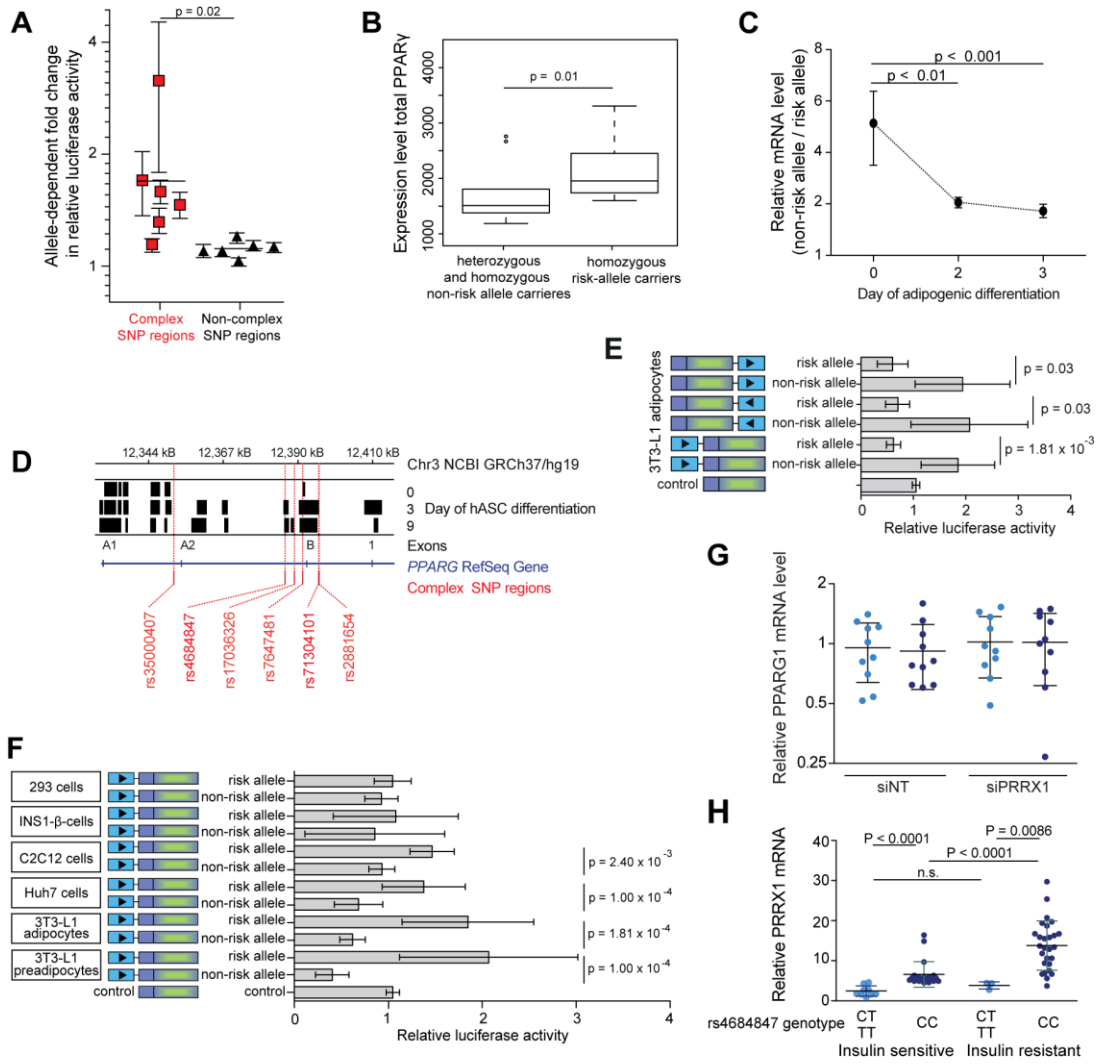
(H,I,M,N) Histograms showing the distribution of $-\log_{10}$ of the estimated probability p-est to randomly observe an equal or higher Ω_{TFBS} (D) and the distribution for an equal or higher *overall score* from all three criteria (E), as calculated from observations in the random set derived from 1,000 shuffled sequences per ortholog set. The blue curve illustrates the empirical density function of the histogram data. The red vertical dashed line indicates the cut-off scores separating complex from non-complex regions, SNP regions with a value to the left of this were defined as non-complex). The isolated peak at the right (low p-est / high overall score data) refers to data points that hit the lower limit of p-est calculations.

(O-U) Homeobox TFs, inferred from a combined analysis using PMCA, positional bias analysis and RNA-seq co-expression correlation data from primary human islets, are associated with metabolic pathways (O-T) and impaired glucose stimulated insulin-secretion (U). (1) PMCA identified 487 complex regions out of 1,465 SNPs at 47 T2D loci (Table S4), (2) bias analysis identified five TFBS matrix families (Figure 3B), comprising a set of TFBS matrices for 63 homeobox TFs (Table S11), and (3) subsequent analysis of mRNA levels in RNA-seq data from primary human islets, comparing donors with and without T2D, implicated the homeobox TFs RAX, PRRX2, BARX1, PITX1, EMX2, NKX6-3, BARX2, MSX2 and PDX1 (Table S14) as novel candidate TFs in T2D pathophysiology.

(O-T) Pathway analysis for gene sets co-expressed with the identified homeobox TFs in pancreatic islets. Gene sets are from co-expression analysis in islets from 51 donors without T2D, correlating the expression levels of all transcripts identified by RNA-seq with the expression levels of the identified genes encoding homeobox TFs PRRX2, PITX1, NKX6-3, BARX2, MSX2 and PDX1 (for which significantly co-expressed genes with FDR 5% were identified, see Table S15). The top five significantly enriched pathways (hypergeometric test, FDR 5%) inferred from WEBGESTALT analysis using the KEGG database are presented;

including the T2D-related categories *metabolic pathways*, *MAPK signaling*, *Notch signaling*, *calcium signaling* and *pancreatic secretion*.

(U) Glucose-stimulated insulin secretion in rat INS-1 β -cells transfected with non-targeting (NT) control siRNA and siRNAs targeting expression of the homeodomain TFs Barx1, Barx2, Msx2, Emx, Nkx6-3, Pitx1, Rax2, Prrx2 or Pdx1 that were differentially regulated in the human islets of subjects with T2D (Table S14). Insulin levels in the medium after 1 h stimulation with high glucose were measured by ELISA (Extended Experimental Procedures). The ratio of (glucose-stimulated insulin levels in siNT transfected cell) / (glucose-stimulated insulin levels in siRNA-homeobox TF) was calculated for siNT control and for each homeobox TF siRNA. p-values from paired t-test, n = 5. The experiments were performed in triplicate.



Supplemental Figure 4. Computational predicted cis-regulatory variants at the *PPARG* T2D risk locus and the homeobox factor *PRRX1* as regulator of endogenous *PPAR γ 2* expression. Related to Figure 4 and Table 2.

(A) Validation of *cis*-regulatory predictions at the *PPARG* T2D risk locus. *Cis*-regulatory predictions for complex regions (red dots) were validated at the level of transcriptional activity. Non-complex regions were included as a control (black dots). Reporter assays were performed with luciferase promoter constructs matching the risk and non-risk alleles of the respective SNP-surrounding regions, reflecting the allele-specific change in transcriptional activity. The quantified change in luciferase expression comparing the risk/non-risk or non-risk/risk allele (change ≥ 1) is shown for each SNP as mean of repeated measures \pm SD (n=3-13), p-value from linear mixed-effects model. Details on the analyzed SNPs are given in Table S12. Complex regions significantly differed from non-complex regions at the transcriptional level.

(B) Genotype-dependent increase in mRNA expression of total *PPAR γ* in human subcutaneous adipose tissue (n = 36). Box plots of the total *PPAR γ* expression level is shown for risk- and non-risk haplotype carriers of rs7638903, Pro12Ala and rs4684847 (*cis*-regulatory variant). Risk-haplotype (GG + CC + CC) versus non-risk haplotype (GA/AA + CG/GG + CT/TT). The three SNPs are in perfect LD in the 1000G Pilot 1 data set (1000 Genomes Project Consortium, 2010) ($R^2 = 1.0$). mRNA was measured by microarrays and analyzed by Wilcoxon rank-sum test.

(C) Allelic imbalance of *PPAR γ 2* mRNA expression levels during early stages of adipocyte differentiation measured in samples of primary hASC (human adipose stromal cells) heterozygous for the risk allele (genotyped for Pro12Ala and rs4684847, $R^2 = 1.0$) at different time points after induction of differentiation. Allele-specific primer extension analysis of RNA (n = 6), calculated as ratio of the non-risk allele to risk allele. Data are presented as mean \pm SD, p-values from Dunn's Multiple Comparison post-test after Kruskal-Wallis Oneway ANOVA ($p < 0.0001$).

(D) Mapping of experimentally verified complex regions to H3K27ac regions at the *PPARG* locus. H3K27ac regions in undifferentiated primary hASC, hASC three days and hASC nine days after induction of adipogenic differentiation were extracted from (Mikkelsen et al., 2010) (data accessible at NCBI GEO database Edgar et al., 2002, accession GSE20752). H3K27ac chromatin state across the *PPARG* locus is shown as region plot, the localizations of SNPs at complex regions and the *PPARG* exons A1, A2, the *PPAR γ 2* specific exon B and the first exon of *PPAR γ 1* and *PPAR γ 2* at the *PPARG* locus are indicated. rs4684847 and rs71304101 reveal cell stage-dependent H3K27ac marks. rs4684847 is distinguished from rs71304101 by H3K4me1, H3K4me2 and H3K36me3.

(E,F) All reporter assays were performed with luciferase promoter constructs matching the risk and non-risk alleles of the respective SNP-surrounding regions, reflecting the allele-specific changes in transcriptional activity. The data are presented as mean \pm SD, p-values from paired t-tests.

(E) Reporter assays with constructs harboring the rs4684847-surrounding region in 5', 3', forward and reverse orientation (arrows) transfected in 3T3-L1 adipocytes (n = 9).

(F) Allele-dependent repression of reporter gene activity in 3T3-L1 adipocytes, Huh7 hepatocytes, C2C12 myocytes, INS1- β -cells and 293 cells. Luciferase assays in 3T3-L1 adipocytes, Huh7 hepatoma cells, C2C12 muscle cells, INS1 pancreatic β -cells and 293T cells reveal cell type-specific *cis*-regulatory activity of the complex region SNP rs4684847.

(G) Regulation of *PPARG1* mRNA expression in SGBS adipocytes with homozygous risk or non-risk allele introduced by the CRISPR/Cas9 genome editing approach. siPRRX1 and siNT were transfected concurrent with induction of differentiation. *PPARG2* mRNA was assessed by qRT-PCR, standardized to *HPRT* mRNA. The data are presented as mean \pm SD, n = 12, p-values from paired t-test.

(H) Genotype-dependent expression of *PRRX1* mRNA levels in insulin resistant and insulin sensitive subjects matched for BMI, body fat, age and sex. *PRRX1* mRNA in abdominal subcutaneous and omental adipose tissue was measured by qRT-PCR, standardized to *HPRT* mRNA. Insulin sensitivity was measured by euglycemic hyperinsulinemic clamp. Data are presented as mean \pm SD (n = 30 per group), p-values from unpaired t-test.

Extended Experimental Procedures

1. Definition of LD blocks

Tag SNPs were derived from reported GWAS loci (corresponding references are listed in Tables S1, S7 and S8A). For each tag SNP, LD blocks were defined based on 1000G Pilot 1 CEU data {1000 Genomes Project Consortium 2010 #31} ($r^2 \geq 0.7$, NCBI GRCh37/hg19) using the SNAP viewer tool {Johnson 2008 #81}, Broad Institute. For Crohn's diseases susceptibility loci, a previously published SNP set {Schaub 2012 #214} was chosen for PMCA analysis of candidate SNPs at (Tables S8B).

2. Search for orthologous regions

For each SNP the 120 bp sequence with the SNP at central position (SNP region) was extracted from the human genome (NCBI GRCh37/hg19). Moreover, orthologous sequences for each of the 120 bp SNP-surrounding region of the human reference sequence were searched in 15 closely and distantly related vertebrate species, using the RegionMiner tool (Genomatix, Munich). First, loci homologous to the human SNP region were searched across the target organisms. In case no homologous loci could be identified, the flanking genes (up to 20 gene loci in both directions) were considered in order to identify a syntenic region in the target species. To be assigned as a syntenic region, two homologous genes in the target organism need to be on the same contig and must show the same relative strand orientation as the genes in the source organism. Second, the input sequence (SNP region) was aligned to the syntenic region using a Smith-Waterman alignment. The syntenic regions had to fulfill the following alignment criteria: the alignment contained a highly conserved 50 bp stretch; the

alignment had to be shorter than 1.5-fold the length of the input SNP region, and a sufficient overall alignment quality had to be reached.

Reference genome: Human (*Homo sapiens*)

Aligned genomes: Rhesus macaque (*Macaca mulatta*)

Common chimpanzee (*Pan troglodytes*)

Mouse (*Mus musculus*)

Rat (*Rattus norvegicus*)

Rabbit (*Oryctolagus cuniculus*)

Horse (*Equus caballus*)

Dog (*Canis lupus familiaris*)

Cow (*Bos Taurus*)

Pig (*Sus scrofa*)

Opossum (*Monodelphis domestica*)

Platypus (*Ornithorhynchus anatinus*)

Zebrafish (*Danio rerio*)

Chicken (*Gallus gallus*)

Western clawed frog (*Xenopus tropicalis*)

Zebra fish (*Taeniopygia guttata*)

3. PMCA-Procedures: Description of the PMCA method

This chapter describes the PMCA method at different degrees of detail. After the description of the general **motivation** for the choice of the method we describe the **general design of the PMCA method** that is intended for a general readership. We then provide a **detailed description of the PMCA algorithm** in the form of a pseudo-code that an experienced bioinformatician can use to implement the steps described in the method in an automated manner. Finally, we provide a **step-by-step example for running PMCA manually using the graphical user interface**

3.1 Motivation

Bioinformatics approaches that reliably assess the regulatory role of specific genetic variants would be highly desirable. However, rapid evolutionary turnover results in many lineage-specific regulatory regions that are functionally conserved, have low phylogenetic conservation, challenging the use of phylogenetic conservation of genomic sequences as a sole denominator in the search for non-coding regulatory regions. Nucleotide-level evolutionary conservation alone has proven to be a poor predictor.

Gene regulatory regions in eukaryotes tend to be organized into *cis*-regulatory modules (CRMs), comprising complex patterns of co-occurring TFBSs for the combinatorial binding of TFs. CRMs integrate a variety of upstream signals to regulate the expression of coordinated sets of genes, making them an obvious target to achieve broad phenotypic changes as a result of adaptive evolution.

Here we hypothesize that the presence of patterns of evolutionarily conserved TFBSs in a CRM (TFBS modularity), within genomic regions surrounding a candidate variant are

predictive of its *cis*-regulatory functionality, regardless of the cross-species conservation of the complete sequence on the nucleotide-level. In order to test this hypothesis we need a bioinformatics method that is able to detect and classify genetic regions that contain evolutionarily conserved TFBS modules. In the following, we describe such a method, called phylogenetic module complexity analysis (PMCA).

3.2 General design of the PMCA method

The starting point of the PMCA method is a genetic variant that has been reported in a genome-wide association study as a tag SNP for the risk of a given disease or a phenotype. In this analysis we individually test all non-coding SNPs that are in linkage disequilibrium (LD, $r^2 < 0.7$) with the tag SNP (see *Definition of LD blocks* for the analysis performed in this manuscript. Note that any set of variants may be analyzed by PMCA). For each non-coding SNP the PMCA method shall eventually provide a classification of the region surrounding the non-coding SNP as being either complex or non-complex. Complex regions are defined as being significantly enriched in phylogenetically conserved TFBS modules according to the scoring scheme we developed for this purpose. In non-complex regions, in contrast, the number of phylogenetically conserved TFBS modules does not exceed what is expected by chance. We estimate this significance using randomized sequences.

The following procedure is executed for each non-coding SNP. We use the commercially available Genomatix software suite (Genomatix Co., Munich) for these tasks, i.e. the *RegionMiner* for extraction of orthologous regions and the *FrameWorker*, which extracts TFBS modules from a set of DNA sequences. Briefly, the *FrameWorker* tool returns the most complex TFBS modules that are common to the input sequences, satisfying the user parameters. TFBS modules are defined as all TFBS that occur in the same order and in a certain distance range in all (or a subset of) the input sequences. However, in principle any

equivalent method can be applied. A more detailed description of the individual computing steps in terms of pseudo-code is given further down.

1. The flanking region (+/-60nt) of the non-coding SNP is extracted from the human genome;
2. Ortholog regions are searched in the genomes of 15 fully sequenced vertebrate species and extracted if a region with a high degree of similarity is found;
3. TFBS are identified in the set of ortholog sequences using position weight matrices from the Genomatix library;
4. TFBS modules are identified in each ortholog sequence; TFBS modules are specifically defined as all two or more TFBSs that occur in the same order and in a certain distance range in all or a subset of the input sequences.
5. Phylogenetically conserved TFBS (Ω_{TFBS}), TFBS modules (Ω_{modules}), and occurrence of TFBSs in TFBS modules ($\Omega_{\text{TFBS_in_modules}}$) are counted.
6. Repeated counting weighs the degree of cross species conservation and the number of TFBS in the modules. This counting scheme alone would overestimate genetic regions that only have orthologs in a subset of closely related vertebrate species (e.g. mammal-lineage specific TFBSs). To account for this possibility, we also determine phylogenetically conserved TFBS with more restricted parameters ($\Omega_{\text{restr-TFBS}}$, details see below).
7. Steps 3-5 are repeated one thousand times using randomized input sequences to estimate the probability of observing a given Ω_{TFBS} , $\Omega_{\text{restr-TFBS}}$, Ω_{modules} , and $\Omega_{\text{TFBS_in_modules}}$. Randomization of the sequences is done using local shuffling in order to conserve local nucleotide frequency distributions. The randomization accounts for the issue that certain TFBSs might be favored merely due to the sequences nucleotide composition, *i.e.* high

GC content may predict additional matches for matrices of the SP1 transcription factor; which might provoke overestimation of the variant-surrounding sequence; and that different ortholog set sizes for candidate variants might result in an artificial bias, i.e. a set of only three sequences allows only two combinations of sequences that contain the reference sequence and fulfill the 50% quorum in contrast to larger sets. Contrary, a region with only primate sequences as orthologous shows a much higher, probably overestimated score.

8. Based on the four weighed counts Ω_{TFBS} , $\Omega_{\text{restr-TFBS}}$, Ω_{modules} , and $\Omega_{\text{TFBS_in_modules}}$ and the estimated background probability of observing these counts by chance, we determine an overall classification criterion.
9. The overall classification criterion labels the input region as *complex* or *non-complex*.

The basic assumption of the PMCA methods is that a genetic variant in a complex region has a measurable functional effect. For classification of a genomic regions as complex or non-complex we determined scoring criteria on the weighed counts (described in detail below) based on the experimental validation of *cis*-regulatory functionality for 21 sequence variants (whether this variant was functional or not in one of two assays: DNA binding activity or reporter gene activity), including the *cis*-regulatory SNPs in Table S2. The gold standard for the test of a classification method is replication in an independent data set that has been measured after the method was fully established. In order to provide such as test we conducted experiments on DNA binding activity or reporter gene activity for a set of 62 SNPs that were selected from a representative set of potential candidate SNPs at genomic regions with different levels of GC content and different intronic or intergenic localization. The PMCA method with the parameters set as described below (and fixed before the experiments on the 62 SNPs were conducted) results in 57 correct classifications, only 3 SNPs were

misclassified as false positives and 2 SNPs as false negatives. We thus expect the PMCA method to have over 90% selectivity and sensitivity.

3.3 Detailed description of the PMCA algorithm (pseudo-code)

Here we describe in detail the steps that need to be taken when using the PCMA method with the Genomatix software in the format of a pseudo-code. In order to get a better feeling of these steps, and how complex regions differ from non-complex regions for a region of interest, we provide a step-by-step tutorial that can be followed manually using the interactive version of the Genomatix software (see provided screenshots). In order to process a large number of SNPs, and to compute the randomized background distributions, we recommend use of the command-line version and scripting of the processing and counting of the output (XML format). While we believe that the RegionMiner and FrameWorker tools (Genomatix Co., Munich) presently represent the state-of-the-art, all steps in our method can be replaced by open-access tools and databases, such as AlignACE {Roth 1998 #381} for the identification of homologous regions, TRANSFAC {Matys 2006 #380} as TFBS databases, and custom-made TFBS module identification schemes.

Pseudo-code for the PMCA algorithm

For a given tagSNP select all non-coding SNPs in the LD region.

For each non-coding SNP do the following:

1. Prerequisites

1.1 Generate a BED-file with

- start position = SNP position – 60 bp

- end position = SNP position + 60 bp

1.2 Search for orthologous regions:

Input the BED-file from step 1.1 input to RegionMiner subtask ‘Search for orthologous regions in other species’

1.3 Download all sequences found in step 1.2

2. Assessment of ‘modular complexity’

2.1 From 1.3 obtain a set of sequence files (S) where each file contains the human sequence surrounding the SNP according to the BED-file contents from 1.1 and up to 15 orthologous sequences from other species as found in 1.2. (Called ‘ortholog sets’).

$$\Omega_{TFBS} = 0$$

$$\Omega_{modules} = 0$$

$$\Omega_{TFBS_in_modules} = 0$$

2.2 For each sequence set S do the following:

N_S = number of sequences in S

For ($i = 2$ to N_S) do the following:

Call FrameWorker using these parameters:

$\zeta = i / \text{number}$ (ζ is the 'quorum')

number of elements in Module: 2 to 10

maximal distance variance: 10

distance between elements: 5 to 200

Parse the output file and determine the following numbers by parsing the XML output:

ω_{TFBS} = number of TFBS in at least $\zeta * N_s$ sequences of S

$\Omega_{TFBS} = \Omega_{TFBS} + \omega_{TFBS}$

For $\gamma = 2$ to 10 do the following

γ is the number of TFBSs that are required to occur

in a module to be counted

$\omega_{\gamma\text{-modules}}$ = number modules with γ TFBS in at least $\zeta * N_s$ sequences of S

$\omega_{TFBS_in_ \gamma\text{-modules}}$ = number of TFBS modules with γ TFBS in at least $\zeta * N_s$ sequences of S

$\Omega_{modules} = \Omega_{modules} + \omega_{\gamma\text{-modules}}$

$\Omega_{TFBS_in_modules} = \Omega_{TFBS_in_modules} + \omega_{\gamma\text{-modules}}$

2.3 Repeat the calculations in step 2.2 but limited to parameter settings of $\zeta \geq 0.5$ sequence set to compute $\Omega_{restr-TFBS}$ (note that $\Omega_{modules}$ and $\Omega_{TFBS_in_modules}$ are not used in the scoring later and thus need not be computed on the restricted set)

2.4 Repeat the following 1,000 times

Randomly shuffle the sequence set S ; use a sliding window of 10 bp and permute the bases in each window, thus leaving the local nucleotide distribution mainly unchanged. This generates randomized sequence sets that are similar in their local nucleotide distribution to S .

Repeat steps 2.2 and 2.3 to obtain a random distribution of Ω_{TFBS}^{rnd} , $\Omega_{restr-TFBS}^{rnd}$, $\Omega_{modules}^{rnd}$, and $\Omega_{TFBS_in_modules}^{rnd}$.

3. Scoring and classification

3.1 Estimate the probability $p-est_i = f(\Omega_i^{rnd} > \Omega_i)$ of observing a given number Ω_i (where i stands for $TFBS$, $rest-TFBS$, $modules$, or $TFBS_in_modules$) as the fraction of randomly observed values of Ω_i^{rnd} that are greater or equal than the Ω_i observed on the true sequences. For numeric stability reasons $p-est_i$ is set to $1/1001$ if this never occurs:

$$p-est_{TFBS} = f(\Omega_{TFBS}^{rnd} > \Omega_{TFBS})$$

$$p-est_{restr-TFBS} = f(\Omega_{restr-TFBS}^{rnd} > \Omega_{restr-TFBS})$$

$$p\text{-est}_{\text{modules}} = f(\Omega_{\text{modules}}^{\text{md}} > \Omega_{\text{modules}})$$

$$p\text{-est}_{\text{TFBS_in_modules}} = f(\Omega_{\text{TFBS_in_modules}}^{\text{md}} > \Omega_{\text{TFBS_in_modules}})$$

3.2 Compute an Overall-score $S_{\text{all}} = -\log(p\text{-est}_{\text{TFBS}} * p\text{-est}_{\text{modules}} * p\text{-est}_{\text{TFBS_in_modules}})$

3.3 Classify a non-coding SNP as being located in a complex region if and only if:

$$(S_{\text{all}} > 6.5) \text{ and } (p\text{-est}_{\text{restr-TFBS}} < 0.15) \text{ and } (p\text{-est}_{\text{TFBS}} < 0.075)$$

(Scoring criteria for classification)

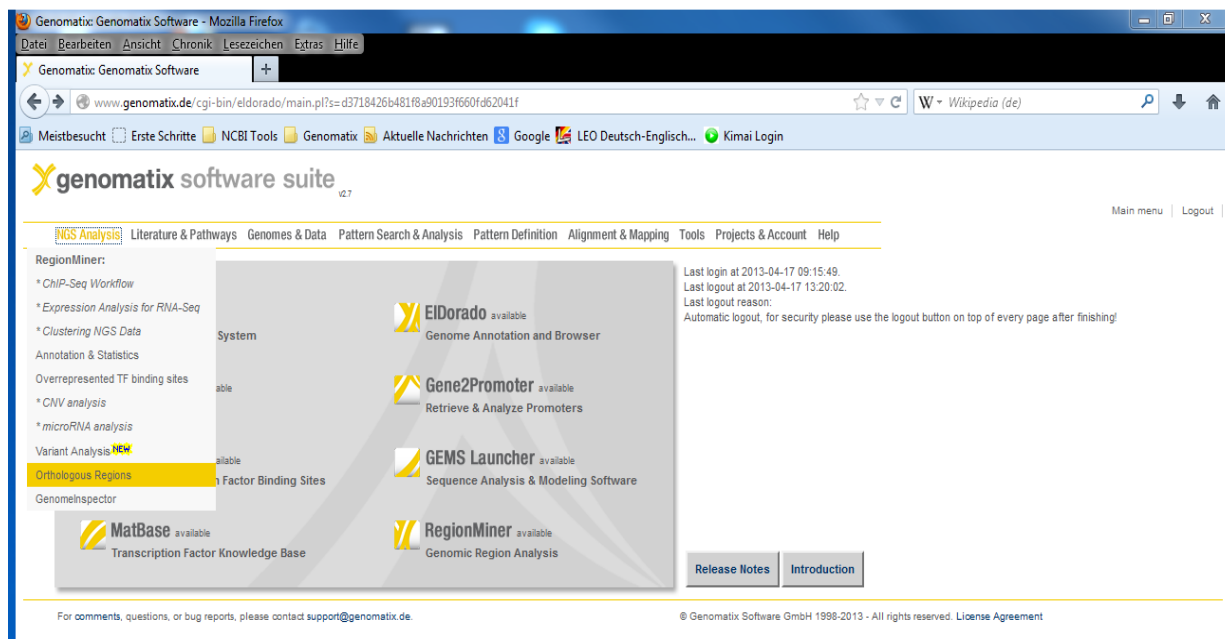
3.4 Step-by-step example for running PMCA manually using the graphical user interface

→Generate a BED-file describing the regions +/- 60 bp around the SNPs. A bed file can be created with any text editor and should contain a single line containing the chromosome, genomic start and end position of the 120 nucleotide region and the SNP identifier.

Below is an example for such a BED-file:

| | | | |
|-------------|-----------------|-----------------|-----------------|
| chr3 | 12386277 | 12386397 | rs468484 |
|-------------|-----------------|-----------------|-----------------|

→Upload the bed file to the Genomatix genome analyzer (GGA) software.



➔ Search for orthologous regions by clicking on ‘Orthologous regions’

Genomatix software suite v2.7

RegionMiner

NGS Analysis Literature & Pathways Genomes & Data Pattern Search & Analysis Pattern Definition Alignment & Mapping Tools Projects & Account Help

Current project: MyProject Current genome version: ElDorado 12-2012

RegionMiner: Search for orthologous regions in other species

Identifies orthologous regions in selected target species.
Sets of orthologous sequences can be saved and analyzed for common TFBS patterns to identify phylogenetically conserved regulatory structures. See help for more.
Limit: max. 300000 regions with at most 250000 bp each

Input

Your files: No BED files in this project yet. [Add BED files](#)

Exclude short sequences

Target

Try to get orthologous sequences from

Vertebrate section select / deselect all vertebrates

- Homo sapiens
- Macaca mulatta
- Pan troglodytes
- Mus musculus
- Oryctolagus cuniculus
- Rattus norvegicus
- Equus caballus
- Canis lupus familiaris
- Bos taurus
- Sus scrofa
- Monodelphis domestica
- Ornithorhynchus anatinus
- Xenopus tropicalis
- Danio rerio
- Gallus gallus
- Taeniopygia guttata

Insect section select / deselect all insects

Plant section select / deselect all plants

Output

Result: Result name:
(special characters except --, _ are not allowed and will be replaced by _)

Your email address:
Use the email option for long-running jobs, to avoid server-timeout messages

For comments, questions, or bug reports, please contact support@genomatix.de © Genomatix Software GmbH 1998-2013 - All rights reserved. License Agreement

➔ Extract the sequences

Genomatix software suite v2.7

RegionMiner Ortholog Search

0.0% regions have an ortholog in Sus scrofa (pig)

| Number of input regions | Percentage input regions | Description |
|-------------------------|--------------------------|---|
| 0 | 0.0% | regions have an ortholog in Monodelphis domestica (opossum) |
| 0 | 0.0% | regions have an ortholog in Ornithorhynchus anatinus (platypus) |
| 0 | 0.0% | regions have an ortholog in Xenopus tropicalis (frog) |
| 0 | 0.0% | regions have an ortholog in Danio rerio (zebrafish) |
| 0 | 0.0% | regions have an ortholog in Gallus gallus (chicken) |
| 1 | 100.0% | regions have an ortholog in Taeniopygia guttata (zebra finch) |

100.0% regions have orthologous seq. in 7 target species

Orthologous Regions

Orthologous Sequences

| Input | Select | Orthologous sequences | Comparative Analysis |
|---|-------------------------------------|--|-------------------------------------|
| Region_1 Id:rs4684847 chr3 12386278-12386397 (120bp) GenomeBrowser | <input checked="" type="checkbox"/> | 7 orthologous sequences found <ul style="list-style-type: none"> Mus musculus (54.5%) Rattus norvegicus (57.4%) Taeniopygia guttata (45.9%) Macaca mulatta (92.5%) Pan troglodytes (99.2%) Equus caballus (51.5%) Oryctolagus cuniculus (57.1%) | Start FrameWorker Start DAlignTF |

Select regions with 7 orthologous sequences

Available tasks for selected regions

Extract selected regions from (Use shift/ctrl-keys to select combinations)

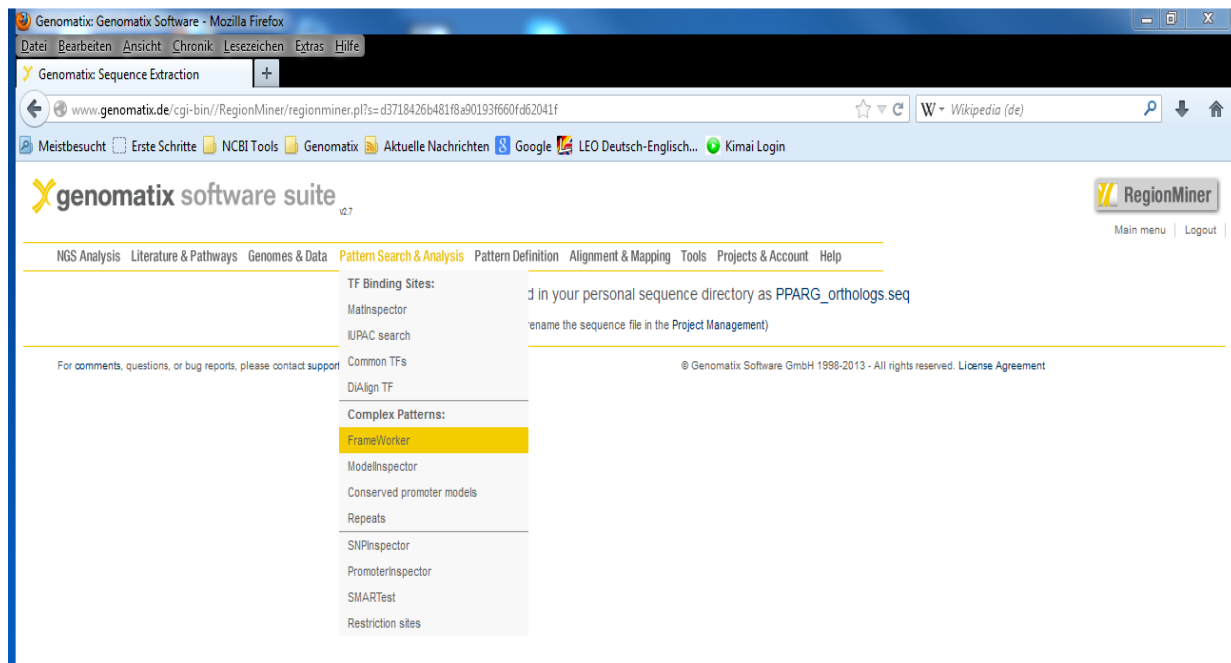
Save selected regions

in GenBank format
 in BED file format

Name for extracted file:

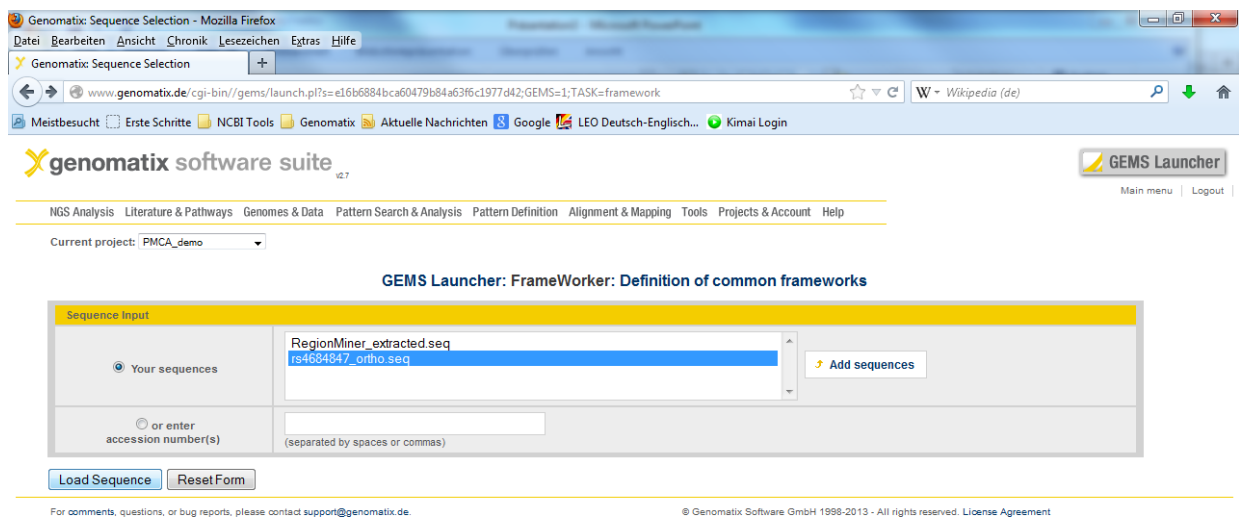
For comments, questions, or bug reports, please contact support@genomatix.de © Genomatix Software GmbH 1998-2013 - All rights reserved. License Agreement

→ Start FrameWorker



The screenshot shows the Genomatix software suite website in a Mozilla Firefox browser. The address bar displays the URL: `www.genomatix.de/cgi-bin/RegionMiner/regionminer.pl?s=d3718426b481f8a90193f660f462041f`. The website header includes the Genomatix logo and the text "genomatix software suite v2.7". A navigation menu is visible, with "Pattern Search & Analysis" selected. A dropdown menu is open under "Pattern Search & Analysis", listing various tools. "FrameWorker" is highlighted in yellow. Other tools listed include TF Binding Sites, MatInspector, IUPAC search, Common TFs, DiAlign TF, ModelInspector, Conserved promoter models, Repeats, SNPInspector, PromoterInspector, SMARTest, and Restriction sites. The footer contains the copyright notice: "© Genomatix Software GmbH 1998-2013 - All rights reserved. License Agreement".

→ Load the ortholog set that has been extracted in the previous step



The screenshot shows the GEMS Launcher interface for FrameWorker. The browser address bar displays the URL: `www.genomatix.de/cgi-bin/gems/launch.pl?s=e16b6884bca60479b84a63f6c1977d42;GEMS=1;TASK=framework`. The website header includes the Genomatix logo and the text "genomatix software suite v2.7". A navigation menu is visible, with "Pattern Search & Analysis" selected. The "GEMS Launcher" button is highlighted. The current project is set to "PMCA_demo". The main section is titled "GEMS Launcher: FrameWorker: Definition of common frameworks". The "Sequence Input" section is active, showing a text box with the file name "RegionMiner_extracted.seq" and the accession number "rs4684847_ortho.seq". An "Add sequences" button is visible. Below the text box, there is a radio button for "Your sequences" and a radio button for "or enter accession number(s)". The "Load Sequence" and "Reset Form" buttons are visible. The footer contains the copyright notice: "© Genomatix Software GmbH 1998-2013 - All rights reserved. License Agreement".

→ Select parameter as described below and click on ‘Start Frameworker’

(Note that in the Genomatix software *TFBS modules* are designated as *models*).

GEMS Launcher: FrameWorker: Definition of common frameworks
working on rs4684847_orf10a.seq (7 sequences, 953 bp)

Quorum constraint: 7 of 7 (100%)

Mandatory sequence. This sequence must contain the models: rs4684847_Human

Distance variance: 10

Distance constraint: 5

Elements in modules: OSINRE, OSMTEN, OSTFEP, OSTF2B, OSTF2D

How many TFBS should be in the found models?

StartFrameWorker ResetForm

For comments, questions, or bug reports, please contact support@genomatix.de © Genomatix Software GmbH 1998-2013 - All rights reserved. License Agreement

→ Count TFBS in the graphical output according to counting scheme described in the algorithm

6 common elements:

| Element | Strand | Matrix sim. | Common to |
|---------|--------|-------------|-----------------------------|
| VSOCT1 | + | Optimized | 7 matches in 7 seq. (100%) |
| VSPARF | - | Optimized | 7 matches in 7 seq. (100%) |
| VSCREB | + | Optimized | 8 matches in 7 seq. (100%) |
| VSHOIF | + | Optimized | 8 matches in 7 seq. (100%) |
| VSPARF | + | Optimized | 8 matches in 7 seq. (100%) |
| VSLIXF | - | Optimized | 10 matches in 7 seq. (100%) |

Graphical view of sites in all found models:

No models were found here.

Ω_{TFBS} : count the common sites in the Human founder sequence = 6

100 bp

→ Repeat the step with the next quorum setting

| FrameWorker Parameters | |
|--|---|
| Quorum constraint for framework | Minimum number of input sequences to contain a framework: 6 of 7 (85%) of input sequences |
| Sequence constraints <small>NEW</small> | Mandatory sequences (sequences that must contain framework, max. 10): rs4684847_Human rs4684847_Rhesus rs4684847_Chimp rs4684847_Mouse rs4684847_Rabbit |
| Distance constraints for framework | Maximum distance VARIANCE between two elements: <input type="text" value="10"/> (max: 100) Distance between two elements: min. <input type="text" value="5"/> max. <input type="text" value="200"/> (max: 500) |
| Element constraints | Number of elements in models: min. <input type="text" value="2"/> max. <input type="text" value="10"/> <input type="checkbox"/> Show intermediate models (else only the longest models are shown) Mandatory elements for models (max. 5): OSINRE OSMTEN OSPFBP OSTF2B OSTF2D Combination of mandatory elements: <input checked="" type="radio"/> ALL selected elements must be present in model <input type="radio"/> ONE of the selected elements must be present in model |
| Options | <input type="checkbox"/> Determine p-values of models |
| Your email address | <input checked="" type="radio"/> Show result directly in browser window <input type="radio"/> Send the URL of the result to <input type="text" value="klocke@genomatx.de"/> <small>Use the email option for long-running jobs, to avoid server-timeout messages</small> |
| Result | |
| Result name (optional) | <input type="text"/> <small>(special characters like # \$ % & + ; , ; < > ? @ not allowed)</small> |

Repeat with different Quorum settings

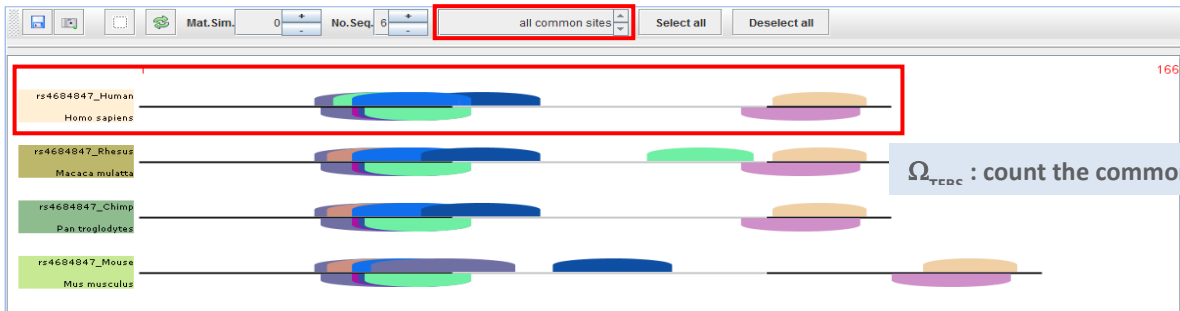
→ Count again

Overview: Models common to at least 6 sequences (85%) and in mandatory sequences

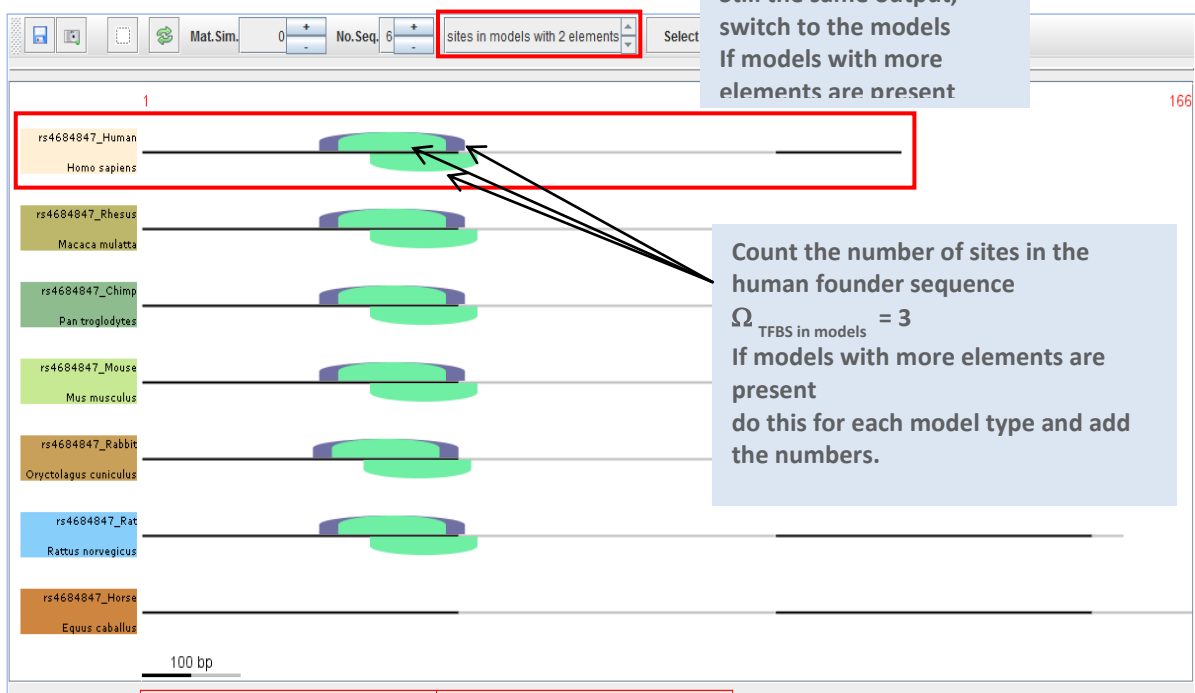
| Models consisting of | | # of different models | |
|----------------------------|------------|--------------------------|-----------------------------|
| single element | 2 elements | 12 common elements found | 2 models found |
| 12 common elements: | | | |
| Element | Strand | Matrix sim. | Common |
| OSPTBP | - | Optimized | 6 matches in 6 seq. (85%) |
| VSHOXC | - | Optimized | 6 matches in 6 seq. (85%) |
| VSOCT1 | + | Optimized | 7 matches in 7 seq. (100%) |
| VSPARF | - | Optimized | 7 matches in 7 seq. (100%) |
| VSSTEM | - | Optimized | 6 matches in 6 seq. (85%) |
| VSCREB | + | Optimized | 8 matches in 7 seq. (100%) |
| VSHOMF | + | Optimized | 8 matches in 7 seq. (100%) |
| VSPARF | + | Optimized | 8 matches in 7 seq. (100%) |
| VSHBOX | - | Optimized | 8 matches in 6 seq. (85%) |
| VSLHXF | + | Optimized | 8 matches in 6 seq. (85%) |
| VSLHXF | + | Optimized | 10 matches in 7 seq. (100%) |
| OSVTBP | + | Optimized | 10 matches in 6 seq. (85%) |

Two models of 2 TFBS were found

Graphical view of sites in all found models:



Graphical view of sites in all found models:



Still the same output, switch to the models If models with more elements are present

Count the number of sites in the human founder sequence
 Ω_{TFBS} in models = 3
 If models with more elements are present do this for each model type and add the numbers.

The counting is cumulative over the Quorum constraint steps, i.e. at this point we have:

$$\Omega_{TFBS} = 6 + 12 = 18$$

$$\Omega_{modules} = 2$$

$$\Omega_{TFBS \text{ in modules}} = 3$$

Keep a second counting for $\Omega_{restr-TFBS}$ which shall only be counted for Quorum constraints of $\geq 50\%$.

Finally for the ortholog set of rs4684847 we obtain four count values:

1. Ω_{TFBS} over all Quorum constraints
2. $\Omega_{modules}$ over all Quorum constraints
3. $\Omega_{TFBS \text{ in modules}}$ over all Quorum constraints
4. $\Omega_{restr-TFBS}$ over Quorum restricted to $\zeta \geq 50\%$

The cumulative counting over all TFBS modules and Quorum constraints gives more weight on sets that yield TFBS modules with higher numbers of TFBS.

Generate a large number of randomized sequence sets and repeat the same steps while keeping track of the count values as above. In order to get robust statistics this step should be performed a thousand times using the command line version of Frameworker tool.

The Genomatix Genome Analyzer (GGA) provides a Unix command line interface (Bioinformatics Workbench) to access the programs through scripting. Frameworker generates XML output files that can be parsed to obtain the Ω_{TFBS} , $\Omega_{restr-TFBS}$, $\Omega_{modules}$, $\Omega_{TFBS \text{ in modules}}$ counts as shown in the manual examples.

Finally each of the counts Ω_{TFBS} , $\Omega_{restr-TFBS}$, $\Omega_{modules}$, $\Omega_{TFBS \text{ in modules}}$ from the 1,000 random sets is compared to the numbers from the original set. These values are then used to estimate the random occurrence of these counts and to derive the final overall-score as described above.

3.5 Key Definitions of PMCA methodology

Key definitions

SNP region. Candidate SNP +/- 60 bp.

TFBS module. At least 2 TFBSs (transcription factor binding sites) that co-occur in the same orientation and a defined distance range across several species.

TFBS modularity. Patterns of conserved TFBSs, reflecting functional conservation of the human candidate SNP region. TFBS modularity is assessed by three criteria (1) TFBS, (2) TFBS modules and (3) TFBS forming TFBS modules that are found across 15 vertebrate species (orthologous regions).

Ortholog set. The orthologous regions that are found in 16 vertebrate species for the human reference SNP region.

Random set. The set of randomly shuffled sequences obtained from an ortholog set by shuffling each orthologous region (simulations were performed with 1,000 random sets). Random sets were used to estimate the probability for random occurrence of a PMCA measure (p-estimates; *p-est.*).

Complex region. A human candidate SNP region meeting the threshold or above for PMCA scores assessed for (1) TFBS, (2) TFBS modules and (3) TFBS in modules within the ortholog set and the shuffled random sets. PMCA identifies SNP regions as complex, based on the conserved complex TFBS modularity, to predict those SNPs that affect gene expression (i.e., *cis*-regulatory SNP).

4. Positional Bias Analysis: Calculation of the TFBS positional bias

The positional bias of a TFBS matrix was calculated as outlined for the assessment of *de novo* detected motifs {Hughes 2000 #78}. For positional bias analysis, the 120 bp sequences analyzed with PMCA were extended to 1,000 bp sequences, serving as a background to check for a significant clustering of certain TFBS at SNP position. The 1,000 bp sequences with the respective SNP at central position were extracted from the human genome build (NCBI GRCh37/hg19) for all complex SNP regions and non-complex SNP regions. The sequences were scanned by MatInspector {Cartharius #47}{Quandt 1995 #121}(Genomatix, Munich, Germany) for the presence of TFBS matrix family matches with respect to SNP position (192

TFBS matrix families; 182 vertebrate families plus 10 other general families, Genomatix Matrix Library version 8.4). Matrix is used in the sense of positional weight matrix (PWM). This is a concept describing TFBSs by the information content of the nucleotide distribution of the positions within a binding site. Hence the scale in the most popular visualization of TFBS matrices (PWMs), the so called LOGO is in bits. What we refer to is weight matrix matches as indicators of putative binding sites. Individual weight matrices describing highly similar binding sites are placed into matrix families {Cartharius #47}. Searching with families eliminates redundant output by giving only the best match within a family. Match positions on the sequences were scanned using overlapping 50 bp sliding windows in steps of 10 bp. The total number of matches for a given TFBS matrix family is regarded as independent individual trials that may match anywhere in the sequence. The positional bias for a scan window under this model becomes the cumulative binomial probability to obtain the exact number of matches found there up to the total number of matches in the sequence. The probability for the occurrence of a single match within a scan window, independent of any sequence constraint, is given as the ratio of the window size to the sequence length. The positional bias (P) was calculated for each matrix family and each window (Table S6). For graphical visualization, $-\log_{10}(P)$ was plotted over the mid-positions of the scan windows. The evaluation of the positional bias was done by parsing the output of MatInspector with a perl script that tabulates for each TF-family the total number of matches, the scan windows, number of matches in the scan windows and binomial P values. For graphical output these tables were input to R and used for plotting.

5. Correlation of SNP regions with evolutionary constraint regions.

Genomic regions surrounding a candidate SNP were classified as complex and non-complex and were correlated to evolutionary constrained regions according to the method and data

from Lindblad-Toh et al {Lindblad-Toh 2011 #27}. We used the *RegionMiner-GenomeInspector* tool (Genomatix, Munich) for this task. From the mid position (anchor position; 0 on the x axis of the plot) of each constrained region (determined by Siphy- π -method Lindblad-Toh et al., 2011) 500 bp in up and downstream direction were scanned for the positions overlapping with the 120 bp of analyzed SNP regions. For each position relative to the anchor the overlaps are counted (correlations) and these correlations *versus* position relative to the anchor are plotted. A preferred distance of complex or non-complex SNPs to constrained elements would be visible as enrichment at defined positions relative to the anchor position. We used the 120 bp extended SNP regions in this analysis since we used the same regions to determining the TFBS module complexity. The use of 120 bp regions has the effect of smoothing the correlation graph, which in case of using exact SNP positions would more adopt the shape of a bar graph since accumulation of overlaps for extended regions is more likely than for single positions. The use of the midpoint of constrained regions as an anchor was chosen since constraint regions do not have the same size. The results are presented in Figure 3, Table S9.

6. Correlation of SNP regions to DNase-seq regions and ChIP-seq regions

Genomic regions surrounding a candidate SNP were classified as complex and non-complex and were correlated to DNase hypersensitive regions (referred to as DNase-seq peaks; summary of Encode data wgEncodeRegDnaseClustered.bed from UCSC regulation super-track) and regions of Transcription Factor binding (referred to as ChIP-seq peaks; summary of Encode ChIP seq data wgEncodeRegTfbsClusteredV2.bed from UCSC regulation super-track). We used the *RegionMiner-GenomeInspector* tool (Genomatix, Munich) for this task. From the mid position (anchor position; 0 on the x axis of the plot) of each SNP region 500 bp in up and downstream direction were scanned for the positions overlapping with the 120

bp of analyzed SNP regions. For each position relative to the anchor the overlaps were counted (correlations) and these correlations *versus* position relative to the anchor were plotted. Enrichment in the vicinity of SNPs would become visible as a peak around the anchor position (0). We used the 120 bp extended SNP regions in this analysis since PCMA used the same regions in determining the TFBS module complexity. The use of 120 bp regions further has the effect of smoothing the correlation graph, which in case of using exact SNP positions would more adopt the shape of a bar graph since accumulation of overlaps for extended regions is more likely than for single positions. The results are presented in Figure 3, Figure S5 and Table S10.

7. Enrichment of complex SNPs in diseases loci

Matched random variants were drawn from the 1000G data. Matching was done as follows: Minor allele frequencies (MAF) of the disease associated SNPs were group into 10 bins. Then for each disease-associated SNP a random equivalent was drawn with a MAF score in the same bin, with the same genomic context (either intergenic, intronic, or exonic) according to Genomatix EIDorado 2012 annotation (Genomatix, Munich, Germany), and for the distance to the nearest TSS within $\pm 10\%$ of the disease-associated SNP. The process of random drawing was done using a Pearl Script.

8. Assessment of SNP to TSS distance annotations

We analyzed SNPs by the *Annotation and Statistics* task of *RegionMiner* tool (Genomatix, Munich) with the option next neighbor analysis. This results in the transcript start sites (TSS) which are next to each SNP upstream and downstream and on either strand of the DNA (Table S9). For visualization we used all distances where a TSS was annotated within 30,000

bp downstream of a SNP. To directly compare these distances for complex and non-complex SNPs we used density histograms with a bin size of 500 bp (Figure S6B).

9. Culture of cell lines

The rat insulinoma cell line INS-1 was cultured in RPMI medium (supplemented with 10 % FBS (fetal bovine serum), 100 mM sodium pyruvate, penicillin/streptomycin and 50 μ M 2-mercaptoethanol). Human Huh7 hepatoma, mouse C2C12 myoblast and mouse 3T3-L1 preadipocyte cell lines were cultured in DMEM medium (supplemented with penicillin/streptomycin and 10 % FBS). The human preadipocyte SGBS (Simpson–Golabi–Behmel Syndrome) cell line was cultured as previously described {Fischer-Posovszky 2008 #178} in DMEM/Ham's F12 (1:1) medium (supplemented with 10% FCS, 17 μ M biotin, 33 μ M pantothenic acid and 1% penicillin/streptomycin). All cells were maintained at 37°C and 5% CO₂. To promote adipose differentiation of the mouse preadipocyte cell line 3T3-L1, cells were grown to confluence with 10% FCS and medium was then additionally supplemented with 250 nM dexamethasone and 0.5 mM isobutyl-methylxanthine for the first three days and 10% FCS and 66 nM insulin throughout the entire differentiation period. C2C12 myoblasts were cultured in DMEM medium containing 10% horse serum to induce differentiation. The SGBS preadipocyte cell strain was grown to confluence. For induction of adipocyte differentiation cells were cultured in serum free MCDB-131/DMEM/Ham's F12 (1:2) medium supplemented with 11 μ M biotin, 22 μ M pantothenic acid, 1% penicillin/streptomycin, 10 μ g/ml human transferrin, 66 nM insulin, 100 nM cortisol, 1 nM triiodothyronine, 20 nM dexamethasone, 500 μ M 3-isobutyl-1-methyl-xanthine (Serva, Germany) and 2 μ M rosiglitazone (Alexis, Germany). 72 hours after induction of differentiation the cells were harvested in TRIzol reagent (Invitrogen, Germany). Unless

other suppliers are mentioned, all cell culture materials were obtained from Invitrogen (Germany) and all chemicals from Sigma-Aldrich (Germany).

10. Luciferase expression constructs

To characterize the SNP-adjacent regions for allele-specific transcriptional activity, genomic sequences surrounding the respective SNPs were cloned into a basal pGL4.22 promoter vector. For the promoter construct, a 752 bp thymidine kinase (TK) promoter was cloned upstream of the firefly luciferase gene into the EcoRV and BglII sites of the pGL4.22 firefly luciferase reporter vector (Promega, Germany). SNP regions were extracted from human genome build (NCBI GRCh37/hg19). SNP regions were commercially synthesized as plasmid vectors (Mr. Gene, Germany) and as double-stranded oligonucleotides (MWG, Germany). Complementary oligonucleotides were annealed and purified on a 12% polyacrylamide gel. SNP regions were subcloned either upstream of the TK promoter into the KpnI and SacI sites of the pGL4.22-TK vector or downstream of the luciferase gene into the BamHI site of the pGL4.22-TK vector. To further test for enhancer activity, SNP-adjacent regions were subcloned downstream of the luciferase gene in both 5'-to-3' and 3'-to-5' orientations into the BamHI site. The QuickChange Site-Directed Mutagenesis Kit (Stratagene, Germany) was used to alter single nucleotides (for the respective SNP, NCBI dbSNP). The orientation and integrity of each luciferase vector was confirmed by sequencing (MWG, Germany).

11. Luciferase expression assays

Huh7 cells (96-well plate, 1.1×10^4 / well) were transfected one day after plating with approximately 90% confluence, INS-1 cells (12-well plate, 8×10^4 / well) were transfected three days after plating with approximately 70% confluence, 3T3-L1 cells (12-well plate, 8×10^4 / well) were transfected at day eight after the induction of differentiation with

approximately 80% confluence and C2C12 cells (12-well plate, 2×10^5 / well) were transfected at day four after induction of differentiation with approximately 90% confluence. Huh7 were transfected with 0.5 μg of the respective firefly luciferase reporter vector and 1 μl Lipofectamine 2000 transfection reagent (Invitrogen, Germany), differentiated C2C12 myocytes were transfected with 1 μg of the respective pGL4.22-TK construct and 2 μl Lipofectamine reagent, and both INS-1 β -cells and differentiated 3T3-L1 adipocytes were transfected with 2 μg of the respective pGL4.22-TK construct and 2 μl Lipofectamine reagent. The firefly luciferase constructs were co-transfected with the ubiquitin promoter-driven renilla luciferase reporter vector pRL-Ubi {Laumen 2009 #249} to normalize the transfection efficiency. Twenty-four hours after transfection, the cells were washed with PBS and lysed in 1x passive lysis buffer (Promega, Germany) on a rocking platform for 30 min at room temperature. Firefly and renilla luciferase activity were measured (substrates D-luciferin and Coelenterazine from PJK, Germany) using a Luminoscan Ascent microplate luminometer (Thermo, Germany) and a Sirius tube luminometer (Berthold, Germany), respectively. The ratios of firefly luciferase expression to renilla luciferase expression were calculated and normalized to the TK promoter control vector. P-values comparing luciferase expression from risk and non-risk alleles or from overexpression experiments was calculated using paired t-test.

For validation of PMCA-driven *cis*-regulatory predictions, and for the comprehensive analysis of the *PPARG* gene locus, allele-dependent change in reporter gene activity was calculated from 3-14 independent experiments for each analyzed SNP (ratio of the respective allelic activities). The quantified change in luciferase activity comparing risk / non-risk or non-risk / risk alleles (change ≥ 1) was calculated for each SNP as mean and standard deviation. P-values were derived from linear mixed-effects model comparing the binary

logarithm of the quantified ratios in allelic luciferase activity between SNPs in complex regions *versus* SNPs in non-complex regions.

12. Electrophoretic mobility shift assay (EMSA)

EMSA was performed with Cy5-labelled oligonucleotide probes. Respective SNP-adjacent region oligonucleotides were commercially synthesized containing either the major or the minor variant (MWG, Germany). Cy5-labelled forward strands were annealed with non-labeled reverse strands, and the double-stranded probes were separated from single-stranded oligonucleotides on a 12% polyacrylamide gel. Complete separation was visualized by DNA shading. The efficiency of the labeling was tested by a dot plot, which confirmed that all of the primers were labeled similarly. For analysis of overexpressed PRRX1 protein in EMSA, a PRRX1 expression vector (pCMV-PRRX1-flag, provided by M. Kern) and the empty expression vector as control were transiently transfected into 293T cells using Lipofectamine 2000 (Invitrogen, Germany). 24 h after transfection, the transfected cells were harvested as total native protein. Nuclear protein extracts from each analyzed cell line were prepared with adapted protocols based on the method described by Schreiber et al (Schreiber et al., 1989). The supernatant was recovered and stored at -80°C. DNA-protein binding reactions were conducted in 50 mM Tris-HCl, 250 mM NaCl, 5 mM MgCl₂, 2.5 mM EDTA, 2.5 mM DTT, 20% v/v glycerol and the appropriate concentrations of poly(dI-dC). For DNA-protein interactions, 3-5 µg of nuclear protein extract from the respective cell line was incubated for 10 min on ice, and Cy-5-labelled genotype-specific DNA probe was added for another 20 min. For competition experiments 11-, 33- and 100-fold molar excess of unlabeled probe as competitor was included with the reaction prior to addition of Cy5-labeled DNA probes. Binding reactions were incubated for 20 min at 4°C. For supershift experiments, cell extracts were pre-incubated with 1 µl of antibody αPRRX1, provided by M. Kern) or 0.4 µg of control

IgG (Santa Cruz Biotechnology, USA) for 20 min at 4 °C. The DNA-protein complexes were resolved on a non-denaturation 5.3% polyacrylamide gel in 0.5x Tris/borate/EDTA buffer. All EMSAs were performed in triplicate or more, and fluorescence was visualized with a Typhoon TRIO+ imager (GE Healthcare, Germany). For comparison of genotype-specific DNA-binding activity in EMSA, competition EMSA and supershift experiments, the intensity of the DNA-protein complexes was quantified for both the major and minor allelic DNA-protein interactions using ImageJ Software (<http://rsbweb.nih.gov/ij/>). Quantification was related to the fluorescence intensity of the whole lane. Quantification was performed in quintuplicate for each single EMSA, and the change in quantified allele-dependent fluorescence intensity was calculated (ratio of the respective allelic activity). For validation of PMCA-driven predictions on allele-specific DNA-binding activity, the quantified change in fluorescence comparing risk / non-risk or non-risk / risk alleles (change ≥ 1) is calculated for each SNP as mean and standard deviation. 3-4 independent EMSA experiments were conducted per SNP and p-values are derived from linear mixed-effects model comparing the decadic logarithm of the quantified change in fluorescence between SNPs in complex regions *versus* SNPs in non-complex regions.

13. DNA-Protein affinity chromatography, LC-MS/MS and label free quantification

To identify DNA-binding proteins interacting with the *cis*-regulatory SNP rs4684847 at the *PPARG* gene locus, we performed DNA-Protein affinity chromatography, LC-MS/MS and label free quantification. *Affinity chromatography*. Streptavidin magnetic beads (Dynabeads M-280, Invitrogen) were coupled with allele-specific biotinylated DNA-probes (the risk and non-risk allele, respectively, of rs4684847 at central position in a 42 bp sequence probe) overnight, washed, equilibrated with 1 x binding buffer (10 mM Tris-HCl, 1 mM MgCl₂, 0.5

mM EDTA, 0.5 mM DTT, 4% v/v glycerol) and incubated with nuclear extracts (binding buffer with 50 mM NaCl and 0.01% CHAPS) and poly (dI-dC) was added. Supernatant was recovered and beads were washed in binding buffer without CHAPS followed by stepwise elution of bound protein from the magnetic beads using increasing concentrations of NaCl. All steps were performed at 4°C. Input protein, wash supernatants and eluates were assayed in EMSA to confirm the binding activity. *Mass Spectrometry.* Eluates revealing allele-specific DNA-protein binding activity were subjected to tryptic digest and mass spectrometry was performed as described before (Hauck et al., 2010; Merl et al., 2012). Briefly, eluted samples were precipitated and protein pellets were resolved in ammoniumbicarbonate followed by tryptic digestion. LC-MS/MS analysis was performed on an Ultimate3000 nano HPLC system (Dionex, USA) online coupled to a LTQ OrbitrapXL mass spectrometer (Thermo Fisher Scientific, Germany) by a nano spray ion source. Peptides were automatically injected and loaded onto the trap column in 5% buffer B (98% ACN/0.1% formic acid in HPLC-grade water) and 95% buffer A (2% ACN/0.1% FA in HPLC-grade water). The peptides were eluted from the trap column and separated on the analytical column by gradient from 5 to 31 % of buffer B followed by a gradient from 31 to 95 % buffer. From the MS prescan, the 10 most abundant peptide ions were selected for fragmentation in the linear ion trap if they exceeded an intensity of at least 200 counts and if they were at least doubly charged. During fragment analysis a high-resolution (60,000 full-width half maximum) MS spectrum was acquired in the Orbitrap with a mass range from 200 to 1500 Da. *Label-free quantification.* The mass spectrometry data were analyzed and quantified using the Progenesis LC-MS software (version 2.5, Nonlinear) as described {Hauck 2010 #166}. Proteins were identified by searching MS and MS/MS data of peptides against the Ensembl mouse protein database (Version NCBI m37; 56410 sequences; 26202967 residues). Averaged LF quantification (LFQ) intensity values were used to calculate protein risk versus non-risk allele ratios. At the

end, the analysis revealed an allele-specific 2.3-fold increased binding of the homeobox TF PRRX1 at the risk-allele of the rs4684847-adjacent region ($p = 0.034$ from one-way Anova comparing the allelic difference of three independent experiments).

14. Genome editing of SGBS preadipocytes

To change the rs4684847 risk allele in SGBS preadipocytes to the non-risk allele we applied an adopted CRISPR/Cas homology directed repair (HDR) genome editing approach {Wang 2013 #353}{Ding 2013 #356}. The CRISPR/Cas expression vector and the sgRNA-expression vector were kindly provided by Dr. Ralf Kühn (Helmholtz Zentrum München, München-Neuherberg). For cloning of the NGG PAM sequence located 203 bp upstream of the rs4684847 variant we annealed the primers 5' *CACCGAAACTCACAACAATGCTGGG*-3' and 5' *AAACCCAGCATTGTTGTGAGTTTC*-3' (the sgRNA target sequence (underlined) and nucleotides for cloning (italics) are indicated) and cloned the resulting double-stranded DNA into a BbsI cloning site of the sgRNA expression vector in front of the U6 promoter, resulting in the sgRNA-rs4684847 vector. The sgRNA target sequence was predicted as *high quality guide sequence* for the low numbers of off-target sites using the algorithms published by {Hsu 2013 #352} (the online tool *Optimized CRISPR Design* at <http://www.genome-engineering.org/> predicted 220 potential off-target sites). To generate a genomic DNA targeting-vector providing the rs4684847 risk and non-risk allele (C- and T-allele, respectively) for HDR-mediated genome editing we amplified the genomic region surrounding the rs4684847 variant (-600 bp and +1,200bp from chr3:12386337, NCBI 37.1/hg19) from SGBS genomic DNA using the Q5 Hot Start High-Fidelity DNA Polymerase (New England Biolabs) and the primers 5'-GGCTTCCCAAAGTCCTGGGATTA-3' and 5'-CTTCCTTTTCTGCCAGCTTCAAA-3'. The PCR-product was cloned into the pJET1.2 vector using the CloneJET PCR Cloning kit (Fermentas). Next, the endogenous homozygous

rs4684847 C-allele was changed to the T-allele (underlined) using the primers 5'-CATCTCTAATTCTTTACA ACTCCGAAAAGATAAGAAAACAGAG-3' and 5'-CTCTGTTTTCTTATCTTTTTCGGAGTTGTAAGAATTAGAGATG-3'. Additionally in both targeting-vectors the NGG-PAM sequence was mutated from AGG→ACG (underlined) using the primers 5'-GCTTTGAATAACGTCCCAGCATTGT-3' and 5'-ACAATGCTGGGACGTTATTCAAAGC-3' to avoid that targeting-vector DNA which was successfully integrated into SGBS genomic DNA will be recognized by the sgRNA-rs4684847. The site directed mutagenesis was performed by overlap-extension PCR (Ho 1989) and both, orientation and integrity of each vector was confirmed by sequencing (MWG, Germany). Next, the sgRNA-rs4684847 vector, the CRISPR/Cas expression vector, the rs4684847 allele-specific targeting-vectors and a GFP-expression vector (to assess transfection efficiency) were co-transfected into the SGBS-preadipocyte cell line using the Amaxa-Nucleofector device (program U-033) and the basis nucleofector kit for primary mammalian fibroblasts (Lonza). Additionally, a truncated CD4 expression vector – lacking all intracellular domains – was co-transfected to enable sorting of transfected cells after transfection, by magnetic bead selection using the MACSelect™ Transfected Cell Selection Kit (Miltenyi Biotec). The sorted cells were grown to confluence (transfection efficiency reached >95%, visually assessed by determining GFP-positive cells) and induced for adipogenic differentiation as described in the Extended Experimental Procedure chapter *Culture of cell lines*, PPAR γ 1 and PPAR γ 2 expression levels were determined as described in the chapter *Quantitative RT-PCR and allele-specific primer extension analysis*. We assessed the genotype of the rs4684847 variant after HDR-mediated genome editing by sequencing 200bp surrounding the SNP and confirmed homozygous C-allele and T-allele in the cells transfected with the respective genomic DNA targeting-vector.

15. Analysis of human tissue samples

Informed consent was obtained from all patients who donated biological samples. The study was approved by the local ethics committee of the Faculty of Medicine of the Technical University of Munich, Germany or the local ethics committee of Karolinska University Hospital, Stockholm, Sweden.

PPRX1 mRNA was measured by qRT-PCR (see section 21) in subcutaneous and omental adipose tissue samples obtained from severely obese subjects matched for BMI ($45 \pm 1.3 \text{ kg/m}^2$), body fat, age and sex, as described previously {Klötting 2010 #285}. Linear regression analyses were performed for free fatty acids (FFA) and glucose infusion rate (GIR) during euglycemic hyperinsulinemic clamps, for risk-allele and non-risk-allele carriers, respectively. Subjects in both the high and low range of GIR were included to enable comparison of different levels of insulin sensitivity.

To determine correlation with insulin sensitivity and circulating lipids (HOMA-IR and TG/HDL ratio), *PPRX1* mRNA was also measured in another cohort comprising 30 obese (BMI > 30 kg/m²) otherwise healthy and 26 non-obese (BMI < 30 kg/m²) healthy women {Arner 2012 #382}. All were pre-menopausal and free of continuous medication. They were investigated in the morning after an overnight fast. A venous blood sample was obtained for measurements of glucose, insulin, and lipids, and for preparation of DNA. HOMA-IR was calculated by the formula $\text{fP-Glucose (mmol/L)} \times (\text{fS-Insulin (microU/ml)} / 22.5)$ {Bonora 2000 #392} After the blood sampling an abdominal subcutaneous adipose tissue biopsy was obtained by needle aspiration. Adipose microarray analysis was performed exactly as described {Arner 2012 #382} using the Affymetrix GeneChip miRNA Array protocol with 1 μg of total adipose RNA from each subject. Gene and miRNA expression have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO; <http://ncbi.nlm.nih.gov/geo>) and are accessible using GEO series accession number

GSE25402. Linear regression analyses were performed to assess correlation of *PRRX1* mRNA with HOMA-IR and TG/HDL in a genotype-dependent and BMI- and age-independent manner for 20 risk-allele and 18 non-risk allele carriers with available phenotype data.

16. Analysis of RNAseq data from primary human islets

Informed consent was obtained from all patients who donated biological samples. The study was approved by the local ethics committee of Lund University, Sweden.

RNA-seq libraries of total RNA from 59 human pancreatic islet donors were made using the standard Illumina mRNA-Seq protocol. Sequencing was done in an Illumina HiSeq 2000 machine. Paired-end 101bp length output reads were aligned to the human reference genome (hg19) with TopHat {Trapnell 2009 #389}. Gene expression was measured as the normalized sum of expression of all its exons. The dexseq_count python script {Anders 2012 #390} was used by counting uniquely mapped reads in each exon. Gene expression normalization was done with the TMM method {Robinson 2010 #391}. Further normalization was applied by adjusting the expression to gene length.

Differential gene expression between normoglycaemic (n=51) and T2D donors (n=8) was assessed with the edgeR Bioconductor package {Robinson 2009 #393}, and significance was defined as FDR < 1%. Further, Pearson's correlation and linear regression analyses were run using the R statistical computing environment for the 9 TFs, i.e. BARX1, BARX2, EMX2, MSX2, NKX6-3, PDX1, PITX1, PRRX2 and RAX, separately for the normoglycaemic and T2D subject groups against 18,567 genes with available gene expression data. The linear regression analysis was performed adjusting for sex, age and BMI. The obtained p-values of correlation/regression were FDR-corrected and a 5% significance threshold was used to select significantly co-expressed genes. Interestingly, expression levels

for RAX from the group with HbA1c < 6 was found to be equal to 0 for all individuals, therefore no genes were co-expressed with RAX for HbA1c < 6. Similarly, after FDR-correction BARX1 did not have any significantly co-expressed genes for HbA1c < 6.

Using the lists of significantly co-expressed genes (FDR 5%) for each of the 9 TFs, pathway analysis was performed by WEBGESTALT {Wang 2013 #394} with KEGG {Kanehisa 2011 #395} and Disease Association Analysis databases. The pathway enrichment analysis was based on the hypergeometric test, and an FDR threshold of 5% was used for selecting pathways significantly associated with the lists of significantly co-expressed genes for each TF. No pathway analysis was possible for RAX and BARX1. EMX2 had only 6 significantly co-expressed genes which could not be unified to any pathway. In summary, 6 out of the 9 TFs had pathway analysis information for HbA1c < 6.

Using the 47 T2D tagSNPs, a list of genes located \pm 500kb away from the tagSNPs was created which included 380 genes in total. For each TF we determined how many of these 380 “T2D genes” were among the FDR 5% significantly co-expressed genes and compared this number with the total amount of significantly co-expressed genes for this TF, i.e. the T2D gene enrichment analysis was performed based on the Fisher’s exact test for each TF for individuals with HbA1c < 6. P-values and odds ratios (ORs) of significant enrichment of the T2D genes were calculated using R statistical computing environment.

17. eQTL analysis

Informed consent was obtained from all patients who donated biological samples. The study was approved by the local ethics committee of Lund University, Sweden. Total PPAR γ expression levels of carriers and non-carriers of the protective allele of the rs7638903 variant (perfect LD ($r^2 = 1.0$) to the tag SNP Pro12Ala and the rs4684847 *cis*-regulatory variant; 1000G Pilot 1 (1000 Genomes Project Consortium, 2010)) were compared using Wilcoxon

signed rank test. RNA was extracted from subcutaneous adipose tissue biopsies from 31 males from Malmö, Sweden, recruited for an exercise intervention {Elgzyri 2012 #259}. Only baseline (before exercise) examination data have been used here. Microarray analysis was performed using the GeneChip® Human Gene 1.0 ST whole transcript based array (Affymetrix, Santa Clara, CA, USA) following the Affymetrix standard protocol. Basic Affymetrix chip and experimental quality analyses were performed using the Expression Console Software, and the robust multi-array average (RMA) method was used for background correction, data normalization and probe summarization. Genotyping was performed using the Illumina Omni express following the Illumina standard protocol.

18. Isolation, culture and differentiation of primary human adipose stromal cells (hASC)

Primary human adipocyte progenitor cells for allele-specific primer extension analysis were obtained by lipoaspiration or surgical excision of subcutaneous adipose tissue, and were isolated and cultured as previously described {Hauner 2001 #177} with some modification. Briefly, after expansion and freezing, the cells were cultured in 6-well plates DMEM/F12 (1:1) medium (supplemented with 10% FCS and 1% penicillin/streptomycin) for 18 h, followed by expansion in DMEM/F12 medium (supplemented with 2.5% FCS, 1% penicillin/streptomycin, 17 μ M biotin, 33 μ M pantothenic acid), 132nM insulin (Sigma, Germany), 10ng/ml EGF (R&D, Germany), and 1ng/ml FGF (R&D, Germany)) until confluence. Adipogenic differentiation was then induced by additionally adding 50 μ L insulin (10mg/ml), 100 μ L cortisol (0.1mM), 1ml transferrin (1mg/ml), 50 μ L T3 (1nM/L), 50 μ L rosiglitazone (2mM), 100 μ L dexamethasone (25 μ M) and 1.25ml IBMX (20mM). The cells were harvested in TRIzol reagent (Invitrogen, Germany) (qRT-PCR) or buffer RLT (Qiagen, Germany) (microarrays, see section 22).

19. Genotyping

Primary hASCs and adipose tissue samples were genotyped for rs1801282 and rs4684847 with a concordance rate of > 99.5% using the MassARRAY system with iPLEX™ chemistry (Sequenom, USA), as previously described {Holzapfel 2008 #77}. Genotypes in primary hASC were additionally confirmed by Sanger sequencing. For rs1801282 the following primers were used: F, 5'-GATGTCTTGACTCATGGGTG-3' and R, 5'-CTGGAGTGTACACATGATAGT-3' (PCR primers) and 5'-GACTCATGGGTGTATTCACA-3' (sequencing primer). For rs4684847 the following primers were used: F, 5'-CCTGAAGCGTATTTATGTAGCTCC-3' and R, 5'-CATTCAAGCCTTGTCACATCTCTG-3' (PCR primers) and 5'-CCTGAAGCGTATTTATGTAGCTCC-3' (sequencing primer). The PCR reaction was performed with around 50ng of input genomic DNA in a Professional Thermocycler (Biometra, Jena, Germany) as follows: 12 min at 95°C, 50 cycles of 20 sec at 95°C, 40 sec at 56°C and 90 sec at 72°C, and finally 2 min at 72°C before cooling.

20. Gene knock-down by siRNA

SGBS cells grown to confluence in 6-well plates (day 0) were treated to induce adipocyte differentiation (section 9) and simultaneously transfected using the same protocol and siRNA as for primary hASCs. 72 hours after induction of differentiation, the cells were harvested in TRIzol reagent (Invitrogen, Germany) and frozen at -80°C.

The rat insulinoma cell line INS-1 was cultured as described above. Cells were treated with 25nM non-targeting (NT) control or siRNA targeting the homeodomain transcription factors Barx1, Barx2, Msx2, Emx, Nkx6-3, Pitx1, Rax2, Prrx2 or Pdx1 (ON-TARGETplus human siRNA SMARTpool (Dharmacon, USA)) using HiPerFect (Qiagen, Germany)

according to the manufacturer's protocol. After 72 hours, the medium was changed to low glucose concentration (5 mM) for 24 h. On the next day the medium was changed to low glucose (5mM) or high glucose medium (25mM) for 1 hour to induce glucose-stimulated insulin-secretion. The medium supernatant was collected and insulin-concentrations were measured using a commercially available insulin-ELISA (Merckodia, Sweden). The cells were harvested in buffer RLT (Qiagen, Germany) and frozen at -80°C for extraction of RNA and determination of knockdown efficiency.

21. Quantitative RT-PCR and allele-specific primer extension analysis

RNA from SGBS cells, adipose tissue biopsies and primary hASCs was isolated by TRIzol reagent (Invitrogen, Germany) followed by the NucleoSpin Kit (Macherey-Nagel, Germany). The high capacity cDNA Reverse Transcription kit (Applied Biosystems, Germany) was used for transcription of 1µg total RNA into cDNA. qPCR analysis of PRRX1, the human PPAR γ 1 and PPAR γ 2 isoform transcripts (NCBI Accession: NM_138712, NM_015869), and other genes (Table 1, primers are shown in table below) was performed using a qPCR SYBR-Green ROX Mix (ABgene, Germany) and using the Mastercycler Realplex system (Eppendorf, Germany) with an initial activation of 15 min at 95°C followed by 40 cycles of 15 sec at 95°C, 30sec at 60°C and 30 sec at 72°C. Amplification of specific transcripts was confirmed by melting curve profiles (cooling the sample to 68°C and heating slowly to 95°C with measurement of fluorescence) at the end of each PCR. Mean target mRNA level was calculated by the $\Delta\Delta$ CT method relative to the level of hypoxanthin phosphoribosyltransferase (*HPRT*) (human) or *Gapdh* (rat) based on technical duplicates.

For allele-specific primer extension analysis of the human PPAR γ 2 isoform transcript in primary hASCs (heterozygous for rs1801282 and rs4684847) mRNA was reverse transcribed into cDNA using random hexamers. Next, the region surrounding the SNP

rs1801282 was amplified using the cDNA forward and reverse primers. Genomic DNA regions surrounding the SNP rs1801282 was amplified using the genomic DNA primers. Annealing temperatures for genomic DNA PCR and RT-PCR were 59°C and 60°C respectively. PCR products were analyzed on an agarose gel and purified by gel extraction using the Wizard VS Gel and PCR Clean-Up System (Promega, Germany). Molarity of purified amplicons were calculated and primer extension assays were performed with Snapshot forward (51°C annealing temperature) and Snapshot reverse (54°C annealing temperature) primers using the ABI Prism SNaPshot Kit. cDNA synthesis and primer extension assays were performed with kits from Applied Biosystems (Germany). For amplification of genomic DNA the GoTaq DNA Polymerase Kit (Promega, Germany) was used. The reaction products were analyzed by gel capillary electrophoresis on ABI 3100 DNA Analyzer and the electropherograms were analyzed with the Gene Mapper 4.0 software. The peak area values from RNA (or cDNA) primer extension products were normalized to the corresponding peak area values from genomic DNA primer extensions products in each experiment for both, the risk allele and the non-risk allele. To normalize for the mean expression level from the risk allele, the (RNA/genomic DNA) ratios for both, risk and non-risk allele, were divided by the mean of all risk-allele ratios (Figure 4D). To assess allelic imbalance of PPAR γ 2 mRNA expression during adipogenic differentiation the ratio of RNA levels (normalized to genomic DNA levels) from non-risk to risk allele were calculated (Figure S7C).

Isoform specific primers for PPARG mRNA (MWG, Germany) were designed using the NCBI Primer Blast software (<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>) and optimized for secondary structures using the Net Primer analysis software (<http://www.premierbiosoft.com/netprimer/>).

Primers and probes used for qPCR

| Target gene | Forward primer | Reverse primer |
|--------------------|---------------------------|---------------------------|
| Human | | |
| <i>PCK1/PEPCKC</i> | GCTCTGAGGAGGAGAATGG | TGCTCTGGGTGACGATAAC |
| <i>PDK4</i> | TGCCAATTTCTCGTCTGTATG | AAAAACAGATGGAAAAGTGGG |
| <i>LIPE</i> | AGAAGATGTCGGAGCCATA | GGTCAGGTTCTTGAGGGAATC |
| <i>BBOX1</i> | TTCCAAGCAGGCCAGAG | CTGAACCCCAGGTGGATG |
| <i>ADIPOQ</i> | CATGACCAGGAAACCACGACT | TGAATGCTGAGCGGTAT |
| <i>OPG</i> | TTATGAGCATCTGGGACGGTGCTGT | AAGGAAGGTACAGTTGGTCCAGGGT |
| <i>GLUT4</i> | CTGTGCCATCCTGATGACTG | CCAGGGCCAATCTCAAAA |
| <i>TIMP3</i> | CTGACAGGTCGCGTCTATGA | AGTCACAAAGCAAGGCAGGT |
| <i>THRSP</i> | CGAGAAAGCCCAGGAGGTGA | AGCATCCCGGAGAAGTGGAGC |
| <i>PPARG1</i> | CGTGGCCGCAGATTTGA | AGTGGGAGTGGTCTTCCATTAC |
| <i>PPARG2</i> | GAAAGCGATTCCTTCACTGAT | TCAAAGGAGTGGGAGTGGTC |
| <i>PRRX1</i> | GTGGAGCAGCCATCGTA | TGGGAGGGACGAGGATCT |
| <i>HPRT</i> | TGAAAAGGACCCACGAAG | AAGCAGATGGCCACAGAAGT |
| Rat | | |
| <i>Pdx1</i> | TCCCGAATGGAACCGAGA | GTCAAGTTGAGCATCACTGCC |
| <i>Barx2</i> | AGTACCTCTTACCCAGACAG | CGTCTTACCTGTAAGTGGCT |
| <i>Pitx1</i> | ACTCAGCCAGCGAGTCATCC | TTCTTCTTGGCTGGGTCTTCC |
| <i>Rax2</i> | AGCGGGACCTCAGTTTGG | CTTGGTCTTCGTGCCGTACTC |
| <i>Msx2</i> | AAGGCAAAAAGACTGCAGGA | GGATGGGAAGCACAGGTCTA |
| <i>Emx2</i> | GTCCCAGCTTTTAAGGCTAGA | CTTTTGCCTTTTGAATTTTCGTT |
| <i>Nkx6-3</i> | ATGCAGCAACACCCAGCA | CCAGTGAATAAGCCAGCCTC |
| <i>Prrx2</i> | ACTTCTCGGTGAGCCACCT | GCTGCTTCTTCTCCGTTTG |
| <i>Barx1</i> | CCTAGCCGTGGTTCGCAT | GCCAGTGGGAAGTGAACA |
| <i>Gapdh</i> | TGGGAAGCTGGTCATCAAC | GCATCACCCCATTTGATGTT |

Allele-specific primer extension analysis PPAR γ 2

| | Forward primer | Reverse primer |
|-------------|-------------------------|--------------------------|
| genomic DNA | TCCATGCTGTTATGGGTGAA | GGAGCCATGCACAGAGATAA |
| cDNA | TCCATGCTGTTATGGGTGAA | GATGCAGGCTCCCATTTGAT |
| Snapshot | CTCTGGGAGATTCTCCTATTGAC | TATCAGTGAAGGAATCGCTTTCTG |

22. Genome-wide expression analysis in primary human hASC

Subcutaneous stromal vascular cells were obtained from liposuction aspirate of ten healthy rs4684847 risk-allele carriers, with written informed consent from each subject. The study was approved by the Regional Committee for Medical Research Ethics (REK) of Haukeland University Hospital, Bergen, Norway. The tissue was prepared as described previously {Veum 2011 #179}. Briefly, tissue was digested for 2 hours at 37°C using a 1:1 ratio of tissue and KRP buffer containing ~55 Wunch/liter collagenase with thermolysin (Liberase Blendzyme TM 10X, Roche) and 0.1% BSA. The digested tissue was filtered through a 210µm nylon mesh into a cup, adipocytes were allowed to float, and the other cells in solution underneath were collected and centrifuged at 200g for 10 min. The floating fraction was washed two times with 15ml PBS to release more cells. Red blood cells were lysed using a buffer with 155mM ammonium chloride, 5.7mM dipotassium phosphate and 0.1mM EDTA, followed by filtration through a 70µm nylon mesh cell strainer (BD Falcon).

The stromal vascular cells were cultured in 12-well plates in DMEM GlutaMax (Gibco) supplemented with 10% FCS and 1% penicillin/streptomycin, and induced to differentiate the day after (“day 1”) by adding cortisol (100nM/L), insulin (66nM/L), transferrin (10µg/ml), biotin (33µM), pantothenate (17µM/L), T3 (1nM/L) and rosiglitazone (10µM). On day 1 or 2, cells were additionally transfected with 35nM non-targeting (NT) control or siRNA targeting PRRX1 or PRRX1 and PPARG simultaneously (25 and 10nM siRNA, respectively) (ON-TARGETplus human siRNA SMARTpool, Dharmacon, USA), using HiPerFect (Qiagen, Germany) according to the manufacturer’s protocol. After 72 hours, the cells were harvested in buffer RLT (Qiagen, Germany) and frozen at -80°C.

RNA was extracted from siRNA-transfected lysates using the RNeasy Lipid Tissue Mini Kit (Qiagen, Germany), and quality controlled by the Agilent 2100 Bioanalyzer (RIN > 9). About 240ng of total RNA from each sample was biotin-labelled using the Illumina TotalPrep RNA Amplification Kit. 750 ng cRNA amplified from each sample with T7 RNA Polymerase was then hybridised at 58°C for 17 hours, according to the Whole-Genome Gene Expression Direct Hybridization Assay Guide from Illumina. Global gene expression was measured with Illumina Bead Array Technology (HumanHT-12 v4 Expression Bead Chip, including 47,323 probes covering more than 28,000 annotated coding transcripts). The raw data are available in the MIAME compliant public repository ArrayExpress (accession number to be included upon publication).

Data were quantile normalized and log₂-transformed, and differential expression was determined by paired Significance Analysis of Microarray (SAM) using the J-Express software ({Dysvik 2001 #396}). A total of 2,258 transcripts were defined as differentially regulated by *PRRX1* knock-down (q-value < 0.2), thereof 1,072 up-regulated transcripts. We selected a matching number of transcripts regulated by simultaneous *PPARG* knock-down (q<0.428), of which 1,125 were up-regulated, and identified 364 *PRRX1*-regulated transcripts that were also regulated by *PPAR*γ₂, 336 for which siPPARG reversed the effect of siPRRX1 (anti-regulation). Because the *PPARG* siRNA targeted total *PPARG* mRNA, we assume that these anti-regulated transcripts were regulated via *PPAR*γ₂ and not *PPAR*γ₁, since *PRRX1* specifically regulates *PPARG2* mRNA expression (verified by qPCR, data not shown; see also Table 1 and Figure S4G).

Gene Set Enrichment Analysis (GSEA) {Subramanian 2005 #397} was performed for the 336 transcripts regulated by siPRRX1 and reversed by siPPARG, to evaluate to what extent the effect of *PRRX1* on global gene expression was mediated via *PPAR*γ₂. Ranking all 2,258 *PRRX1*-regulated transcripts by fold change, an accumulated score for the 336 anti-

regulated genes was calculated by starting at the top of the FC-ranked list, giving a positive value 1 for each transcript in the 336 list, while a negative value 1 was subtracted for each transcript not in the list. All genes at the top of the list within a positive accumulated score comprise the “leading edge”, which was used to obtain the enrichment p-value relative to the full set of 2,258 transcripts. Finally, for the 336 genes that were inversely regulated by PRRX1 and PPAR γ 2, Ingenuity Pathway Analysis (IPA, www.ingenuity.com) (Qiagen, Germany) was performed to describe the best scoring molecular and cellular functional categories and molecular networks. Standard settings for IPA were used. The top-scoring network (Figure 5E) is displayed with color overlay for each gene corresponding to the sum of fold change after PRRX1 knock-down and PRRX1+PPARG knock-down (darker red color indicates up-regulation by PRRX1 knock-down/down-regulation by PRRX1/PPARG knock-down, and green vice versa).

23. Assessment of lipid accumulation after PRRX1 overexpression

To assess an inhibitory effect of PRRX1 on lipid accumulation in adipose cells, we stably overexpressed PRRX1 using lentiviral transduction in SGBS cells. Cells were differentiated into mature adipocytes as described above (section 9). Eight days after induction, medium was removed, cells were washed twice with PBS, followed by fixing in 3.7% formaldehyde for 5 min. The fixation solution was then changed and the cells were incubated for an additional 1.5 hours at room temperature, followed by two washes with PBS and incubation with 60% isopropanol for 5 minutes. The isopropanol was removed and replaced by Oil-Red-O stain solution (0.3% Oil-Red-O in 60/40 isopropanol/H₂O, filtered through a 0.2 μ m mesh) for 60 min, before carefully washing twice with PBS, adding 1ml PBS, and photography under a Nikon TE2000 microscope.

24. Glyceroneogenesis and 2-deoxyglucose uptake measurements in primary hASC

For metabolic studies, primary hASCs were treated with NT control or PRRX1 siRNA as described above and treated or not with 10 μ M rosiglitazone. After 72 hours, cells were fasted for 3 hours in serum-deprived, glucose-free DMEM containing 0.3% (w/v) fatty acid-free BSA. Then, cells were transferred in a Krebs Ringer Bicarbonate buffer containing 0.3% BSA, 5 mM pyruvate and 20 μ M [1-¹⁴C]-pyruvate (0.5 μ Ci) as precursor of glycerol-3-phosphate. 2 hours later, cells were rinsed in PBS and scraped in 10 mmol/l Tris-Cl, pH 7.4, containing 0.25 mol/l sucrose, 0.1 mmol/l EDTA, 0.1 mmol/l dithiothreitol, and 0.1% Triton and frozen in liquid nitrogen before lipid extraction according the simplified method of Bligh and Dyer {Bligh 1959 #384}. The subsequent [1-¹⁴C]-pyruvate incorporation was estimated by counting the radioactivity associated with the lipid fraction. The incubation medium (2 h) was stored at -20 C for further NEFA (Free Fatty Acids Half Micro Test, Roche Diagnostics) determinations.

Insulin-stimulated 2-deoxyglucose (2DG) uptake studies were performed as previously described {Richling 2011 #385}. Briefly, hASCs were transferred to glucose-free Krebs-Ringer-Hepes buffer containing 2.5 mM pyruvate, and 0.5% BSA 2.5 hours prior to the experiment. Cells were stimulated or not with 1 μ M insulin for 30 sec. Basal and insulin-stimulated 2-DG uptake was initiated by the addition of KRH buffer containing 0.5% BSA, 2.5 mM pyruvate, 50 μ M 2-DG and [³H]-2-DG [2 μ Ci/ml]. Uptake was terminated by addition of ice-cold KRH containing 150 μ M phloretin and 15 μ M cytochalasin B. Cells were lysed in 0.1 M NaOH and radioactivity was measured using liquid scintillation counting. Quenching of radioactivity was considered applying an external standard. 2-DG transport values were corrected for protein content determined by the bicinchoninic acid method (BCA Protein Assay Reagent, PIERCE, Rockford, USA).

25. Statistical analysis

A $P < 0.05$ was considered statistically significant. P-values in luciferase assays were calculated by unpaired t-test. In experiments assessing allelic imbalance of PPAR γ 2 mRNA expression during adipogenesis, p-values were calculated using Kruskal-Wallis Oneway ANOVA followed by Dunn's Multiple Comparison post-test. For qRT-PCR analysis of siRNA experiments, p-values were calculated using the Wilcoxon rank-sum test (INS1 cells, n=9) or paired t-test (hASC, n=16/32). Unpaired t-test was used for qRT-PCR experiments assessing genotype-dependent effects on mRNA expression (hASC, n=16/32), and Mann Whitney U test was used for allele-specific primer extension analysis. Correlations of *PPRX1* mRNA with *PPARG2* mRNA, pyruvate incorporation and free fatty acid release were calculated by Pearson's correlation. For correlation analysis of adipose tissue *PPRX1* mRNA expression with FFA levels and GIR (glucose infusion rate) in the BMI-matched study sample (n=67), we performed linear regression with log transformed values. For correlations with HOMA-IR, BMI and TG/HDL levels (n=38) we performed linear regression with log-transformed residuals (adjusted for age, sex and BMI). Statistical analyses were done using the Graph Pad Prism software version 5.02 or the Statistical Software R, version 2.14.2.

SUPPLEMENTAL NOTES – Authors from DIAGRAM+:

Benjamin F Voight^{1,2,3}, Laura J Scott⁴, Valgerdur Steinthorsdottir⁵, Andrew P Morris⁶, Christian Dina^{7,8}, Ryan P Welch⁹, Eleftheria Zeggini^{6,10}, Cornelia Huth^{11,12}, Yurii S Aulchenko¹³, Gudmar Thorleifsson⁵, Laura J McCulloch¹⁴, Teresa Ferreira⁶, Harald Grallert^{11,12}, Najaf Amin¹³, Guanming Wu¹⁵, Cristen J Willer⁴, Soumya Raychaudhuri^{1,2,16}, Steve A McCarroll^{1,17}, Claudia Langenberg¹⁸, Oliver M Hofmann¹⁹, Josée Dupuis^{20,21}, Lu Qi²²⁻²⁴, Ayellet V Segre^{1,2,17}, Mandy van Hoek²⁵, Pau Navarro²⁶, Kristin Ardlie¹, Beverley Balkau^{27,28}, Rafn Benediktsson^{29,30}, Amanda J Bennett¹⁴, Roza Blagieva³¹, Eric Boerwinkle³², Lori L Bonnycastle³³, Kristina Bengtsson Boström³⁴, Bert Bravenboer³⁵, Suzannah Bumpstead¹⁰, Noël P Burtt¹, Guillaume Charpentier³⁶, Peter S Chines³³, Marilyn Cornelis²⁴, David J Couper³⁷, Gabe Crawford¹, Alex SF Doney^{38,39}, Katherine S Elliott⁶, Amanda L Elliott^{1,17,40}, Michael R Erdos³³, Caroline S Fox^{21,41}, Christopher S Franklin⁴², Martha Ganser⁴, Christian Gieger¹¹, Niels Grarup⁴³, Todd Green^{1,2}, Simon Griffin¹⁸, Christopher J Groves¹⁴, Candace Guiducci¹, Samy Hadjadj⁴⁴, Neelam Hassanali¹⁴, Christian Herder⁴⁵, Bo Isomaa^{46,47}, Anne U Jackson⁴, Paul RV Johnson⁴⁸, Torben Jørgensen^{49,50}, Wen HL Kao^{51,52}, Norman Klopp¹¹, Augustine Kong⁵, Peter Kraft^{22,23}, Johanna Kuusisto⁵³, Torsten Lauritzen⁵⁴, Man Li⁵¹, Aloysius Lieveise⁵⁵, Cecilia M Lindgren⁶, Valeriya Lyssenko⁵⁶, Michel Marre^{57,58}, Thomas Meitinger^{59,60}, Kristian Midthjell⁶¹, Mario A Morken³³, Narisu Narisu³³, Peter Nilsson⁵⁶, Katharine R Owen¹⁴, Felicity Payne¹⁰, John RB Perry^{62,63}, Ann-Kristin Petersen¹¹, Carl Platou⁶¹, Christine Proença⁷, Inga Prokopenko^{6,14}, Wolfgang Rathmann⁶⁴, N William Rayner^{6,14}, Neil R Robertson^{6,14}, Ghislain Rocheleau⁶⁵⁻⁶⁷, Michael Roden^{45,68}, Michael J Sampson⁶⁹, Richa Saxena^{1,2,40}, Beverley M Shields^{62,63}, Peter Shrader^{3,70}, Gunnar Sigurdsson^{29,30}, Thomas Sparsø⁴³, Klaus Strassburger⁶⁴, Heather M Stringham⁴, Qi Sun^{22,23}, Amy J Swift³³, Barbara Thorand¹¹, Jean Tichet⁷¹, Tiinamaija Tuomi^{46,72}, Rob M van Dam²⁴, Timon W van Haeften⁷³, Thijs van Herpt^{25,55}, Jana V van Vliet-Ostapchouk⁷⁴, G Bragi Walters⁵, Michael N Weedon^{62,63}, Cisca Wijmenga⁷⁵, Jacqueline Witteman¹³, Richard N Bergman⁷⁶, Stephane Cauchi⁷, Francis S Collins⁷⁷, Anna L Gloyn¹⁴, Ulf Gyllenstein⁷⁸, Torben Hansen⁷⁸, Winston A Hide¹⁹, Graham A Hitman⁸⁰, Albert Hofman¹³, David J Hunter^{22,23}, Kristian Hveem^{61,81}, Markku Laakso⁵³, Karen L Mohlke⁸², Andrew D Morris^{38,39}, Colin NA Palmer^{38,39}, Peter P Pramstaller⁸³, Igor Rudan^{42,84,85}, Eric Sijbrands²⁵, Lincoln D Stein¹⁵, Jaakko Tuomilehto⁸⁶, Andre Uitterlinden²⁵, Mark Walker⁸⁷, Nicholas J Wareham¹⁸, Richard M Watanabe^{76,88}, Goncalo R Abecasis⁴, Bernhard O Boehm³¹, Harry Campbell⁴², Mark J Daly^{1,2}, Andrew T Hattersley^{62,63}, Frank B Hu²²⁻²⁴, James B Meigs^{3,70}, James S Pankow⁸⁹, Oluf Pedersen^{43,90,91}, H.-Erich Wichmann^{11,12,92}, Inês Barroso¹⁰, Jose C Florez^{1,2,3,93}, Timothy M Frayling^{62,63}, Leif Groop^{56,72}, Rob Sladek⁶⁵⁻⁶⁷, Unnur Thorsteinsdottir^{5,94}, James F Wilson⁴², Thomas Illig¹¹, Philippe Froguel^{7,95}, Cornelia M van Duijn¹³, Kari Stefansson⁹⁴, David Altshuler^{1,2,3,17,40,93}, Michael Boehnke⁴, Mark I McCarthy^{6,14,96}.

1. Broad Institute of Harvard and Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts 02142, USA
2. Center for Human Genetic Research, Massachusetts General Hospital, 185 Cambridge Street, Boston, Massachusetts 02114, USA
3. Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA
4. Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, USA
5. deCODE Genetics, 101 Reykjavik, Iceland
6. Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK
7. CNRS-UMR-8090, Institute of Biology and Lille 2 University, Pasteur Institute, F-59019 Lille, France
8. INSERM UMR915 CNRS ERL3147 F-44007 Nantes, France
9. Bioinformatics Program, University of Michigan, Ann Arbor MI USA 48109
10. Wellcome Trust Sanger Institute, Hinxton, CB10 1HH, UK
11. Institute of Epidemiology, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany
12. Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, 81377 Munich, Germany
13. Department of Epidemiology, Erasmus University Medical Center, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands.
14. Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, OX3 7LJ, UK
15. Ontario Institute for Cancer Research, 101 College Street, Suite 800, Toronto, Ontario M5G 0A3, Canada
16. Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA
17. Department of Molecular Biology, Harvard Medical School, Boston, Massachusetts 02115, USA
18. MRC Epidemiology Unit, Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge CB2 0QQ, UK
19. Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA
20. Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts 02118, USA
21. National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, Massachusetts 01702, USA
22. Department of Nutrition, Harvard School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA
23. Department of Epidemiology, Harvard School of Public Health, 665 Huntington Ave, Boston, MA 02115, USA
24. Channing Laboratory, Dept. of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Ave, Boston, MA 02115, USA
25. Department of Internal Medicine, Erasmus University Medical Centre, PO-Box 2040, 3000 CA Rotterdam, The Netherlands
26. MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, EH4 2XU, UK
27. INSERM U780, F-94807 Villejuif, France
28. University Paris-Sud, F-91405 Orsay, France
29. Landspítali University Hospital, 101 Reykjavik, Iceland
30. Icelandic Heart Association, 201 Kopavogur, Iceland
31. Division of Endocrinology, Diabetes and Metabolism, Ulm University, 89081 Ulm, Germany
32. The Human Genetics Center and Institute of Molecular Medicine, University of Texas Health Science Center, Houston, Texas 77030, USA
33. National Human Genome Research Institute, National Institute of Health, Bethesda, Maryland 20892, USA
34. R&D Centre, Skaraborg Primary Care, 541 30 Skövde, Sweden
35. Department of Internal Medicine, Catharina Hospital, PO-Box 1350, 5602 ZA Eindhoven, The Netherlands
36. Endocrinology-Diabetology Unit, Corbeil-Essonnes Hospital, F-91100 Corbeil-Essonnes, France
37. Department of Biostatistics and Collaborative Studies Coordinating Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, 27599, USA
38. Diabetes Research Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee DD1 9SY, UK
39. Pharmacogenomics Centre, Biomedical Research Institute, University of Dundee, Ninewells Hospital, Dundee DD1 9SY, UK
40. Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA
41. Division of Endocrinology, Diabetes, and Hypertension, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA
42. Centre for Population Health Sciences, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, UK

43. Hagedorn Research Institute, DK-2820 Gentofte, Denmark
44. Centre Hospitalier Universitaire de Poitiers, Endocrinologie Diabetologie, CIC INSERM 0801, INSERM U927, Université de Poitiers, UFR, Médecine Pharmacie, 86021 Poitiers Cedex, France
45. Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
46. Folkhälsan Research Center, FIN-00014 Helsinki, Finland
47. Malmøsk Municipality Health Center and Hospital, 68601 Jakobstad, Finland
48. Diabetes Research and Wellness Foundation Human Islet Isolation Facility and Oxford Islet Transplant Programme, University of Oxford, Old Road, Headington, Oxford, OX3 7LJ, UK
49. Research Centre for Prevention and Health, Glostrup University Hospital, DK-2600 Glostrup, Denmark
50. Faculty of Health Science, University of Copenhagen, 2200 Copenhagen, Denmark
51. Department of Epidemiology, Johns Hopkins University, Baltimore, Maryland 21287, USA
52. Department of Medicine, and Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University, Baltimore, Maryland 21287, USA
53. Department of Medicine, University of Kuopio and Kuopio University Hospital, FIN-70211 Kuopio, Finland
54. Department of General Medical Practice, University of Aarhus, DK-8000 Aarhus, Denmark
55. Department of Internal Medicine, Maxima MC, PO-Box 90052, 5600 PD Eindhoven, The Netherlands
56. Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, 205 02 Malmö, Sweden
57. Department of Endocrinology, Diabetology and Nutrition, Bichat-Claude Bernard University Hospital, Assistance Publique des Hôpitaux de Paris, 75870 Paris Cedex 18, France
58. INSERM U695, Université Paris 7, 75018 Paris, France
59. Institute of Human Genetics, Helmholtz Zentrum Muenchen, 85764 Neuherberg, Germany
60. Institute of Human Genetics, Klinikum rechts der Isar, Technische Universität München, 81675 Muenchen, Germany
61. Nord-Trøndelag Health Study (HUNT) Research Center, Department of Community Medicine and General Practice, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway
62. Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Magdalen Road, Exeter EX1 2LU, UK
63. Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, University of Exeter, Barrack Road, Exeter EX2 5DW, UK
64. Institute of Biometrics and Epidemiology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
65. Department of Human Genetics, McGill University, Montreal H3H 1P3, Canada
66. Department of Medicine, Faculty of Medicine, McGill University, Montreal, H3A 1A4, Canada
67. McGill University and Genome Quebec Innovation Centre, Montreal, H3A 1A4, Canada
68. Department of Metabolic Diseases, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany
69. Department of Endocrinology and Diabetes, Norfolk and Norwich University Hospital NHS Trust, Norwich, NR1 7UY, UK.
70. General Medicine Division, Massachusetts General Hospital, Boston, Massachusetts, USA
71. Institut interrégional pour la Santé (IRSA), F-37521 La Riche, France
72. Department of Medicine, Helsinki University Hospital, University of Helsinki, FIN-00290 Helsinki, Finland
73. Department of Internal Medicine, University Medical Center Utrecht, 3584 CG Utrecht, The Netherlands
74. Molecular Genetics, Medical Biology Section, Department of Pathology and Medical Biology, University Medical Center Groningen and University of Groningen, 9700 RB Groningen, The Netherlands
75. Department of Genetics, University Medical Center Groningen and University of Groningen, 9713 EX Groningen, The Netherlands
76. Department of Physiology and Biophysics, University of Southern California School of Medicine, Los Angeles, California 90033, USA
77. National Institute of Health, Bethesda, Maryland 20892, USA
78. Department of Genetics and Pathology, Rudbeck Laboratory, Uppsala University, S-751 85 Uppsala, Sweden.
79. University of Southern Denmark, DK-5230 Odense, Denmark
80. Centre for Diabetes, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK
81. Department of Medicine, The Hospital of Levanger, N-7600 Levanger, Norway
82. Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599, USA
83. Institute of Genetic Medicine, European Academy Bozen/Bolzano (EURAC), Viale Druso 1, 39100 Bolzano, Italy
84. Croatian Centre for Global Health, Faculty of Medicine, University of Split, Soltanska 2, 21000 Split, Croatia
85. Institute for Clinical Medical Research, University Hospital 'Sestre Milosrdnice', Vinogradska 29, 10000 Zagreb, Croatia
86. Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki FIN-00300, Finland,
87. Diabetes Research Group, Institute of Cellular Medicine, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK
88. Department of Preventive Medicine, Keck Medical School, University of Southern California, Los Angeles, CA, 90089-9001, USA
89. Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota 55454, USA
90. Department of Biomedical Science, Panum, Faculty of Health Science, University of Copenhagen, 2200 Copenhagen, Denmark
91. Faculty of Health Science, University of Aarhus, DK-8000 Aarhus, Denmark
92. Klinikum Grosshadern, 81377 Munich, Germany
93. Diabetes Unit, Massachusetts General Hospital, Boston, Massachusetts 02144, USA
94. Faculty of Medicine, University of Iceland, 101 Reykjavík, Iceland
95. Genomic Medicine, Imperial College London, Hammersmith Hospital, W12 0NN, London, UK
96. Oxford National Institute for Health Research Biomedical Research Centre, Churchill Hospital, Old Road Headington, Oxford, OX3 7LJ, UK

SUPPLEMENTAL REFERENCES

1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061-1073.

Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M., and Werner, T. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21, 2933-2942.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207-210.

Elgzyri, T., Parikh, H., Zhou, Y., Nitert, M.D., Rönn, T., Segerström, Å.B., Ling, C., Franks, P.W., Wollmer, P., and Eriksson, K.F., et al. (2012). First-Degree Relatives of Type 2 Diabetic Patients Have Reduced Expression of Genes Involved in Fatty Acid Metabolism in Skeletal Muscle. *Journal of Clinical Endocrinology & Metabolism* 97, E1332.

Fischer-Posovszky, P., Newell, F.S., Wabitsch, M., and Tornqvist, H.E. (2008). Human SGBS Cells – a Unique Tool for Studies of Human Fat Cell Biology. *Obes Facts* 1, 184-189.

Hauck, S.M., Dietter, J., Kramer, R.L., Hofmaier, F., Zipplies, J.K., Amann, B., Feuchtinger, A., Deeg, C.A., and Ueffing, M. (2010). Deciphering membrane-associated molecular processes in target tissue of autoimmune uveitis by label-free quantitative mass spectrometry. *Molecular & Cellular Proteomics*.

Holzappel, C., Baumert, J., Grallert, H., Müller, A.M., Thorand, B., Khuseyinova, N., Herder, C., Meisinger, C., Hauner, H., and Wichmann, H.E., et al. (2008). Genetic variants in the USF1 gene are associated with low-density lipoprotein cholesterol levels and incident type 2 diabetes mellitus in women: results from the MONICA/KORA Augsburg case-cohort study, 1984–2002. *European Journal of Endocrinology* 159, 407-416.

Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M. (2000). Computational identification of Cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* 296, 1205-1214.

Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., and Bakker, P.I.W. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 2938-2939.

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.

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.

Merl, J., Ueffing, M., Hauck, S.M., and Toerne, C. von (2012). Direct comparison of MS-based label-free and SILAC quantitative proteome profiling strategies in primary retinal Müller cells. *Proteomics* 12, 1902-1911.

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.

Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Research* 23, 4878-4884.

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.

Schreiber, E., Matthias, P., Müller, M.M., and Schaffner, W. (1989). Rapid detection of octamer binding proteins with 'mini extracts', prepared from a small number of cells. *Nucleic Acids Research* 17, 6419.

Veum, V.L., Dankel, S.N., Gjerde, J., Nielsen, H.J., Solsvik, M.H., Haugen, C., Christensen, B.J., Hoang, T., Fadnes, D.J., and Busch, C., et al. (2011). The nuclear receptors NUR77, NURR1 and NOR1 in obesity and during fat loss. *Int J Obes*.