Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Ernährungsmedizin

Leveraging transcription factor binding site patterns:

from diabetes risk loci to disease mechanisms

Melina Christine Claussnitzer

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Dr. Hannelore Daniel

Prüfer der Dissertation:
1. Univ.-Prof. Dr. Johann J. Hauner
2. Univ.-Prof. Dr. Thomas Illig
   Medizinische Hochschule Hannover
3. Prof. Dr. Dr. Gunnar Mellgren
   University of Bergen, Norwegen
   (nur schriftliche Beurteilung)

   Priv.-Doz. Dr. Thomas Skurk
   (nur mündliche Prüfung)

Die Dissertation wurde am 07.02.2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 05.05.2014 angenommen.

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

Dissertation submitted to the

Technical University of Munich

Technische Universität München

Melina Christine Claussnitzer

All my studies presented in this thesis were performed at the

ZIEL

Zentrales Institut für Ernährung und Lebensmittel, Freising-Weihenstephan

contact: melina.claussnitzer@wzw.tum.de or

melinaclaussnitzer@hsl.harvard.edu

phone: + 617-(852) 1948

Die Gedanken sind frei und unterliegen keinen Gesetzen.
In ihnen findet man die Freiheit des Menschen. Sie herrschen strahlend in die Welt […] erschaffen ein neues Paradies, eine neue Stütze, eine neue Quelle der Kraft, aus der neue Künste hervorspringen.

Philippus Theophrastus Paracelsus (“Septem Defensiones”)

Acknowledgements

First, I would like to express my great gratitude to my supervisor Prof. Hans Hauner, who brought me to the exciting area of metabolic diseases, i.e. type 2 diabetes and obesity. Only the trust of you, and also Dr. Helmut Laumen, which you put in my rather unconventional PhD project enabled the realization of this thesis. I am grateful for the confidence and for the constructive discussions with both of you, which made me following up my project even in temporary hard times. I am very grateful to Thomas Illig, who brought me to the GWAS field and made me focus my work on following-up statistical diabetes association signals from population studies to elucidate their underlying disease mechanisms. You noticed from early stages on the potential of the computational PMCA approach for human disease genetics. Without your supportive comments on the project, I would not have had the enthusiasm and persuasion that drove me all the time. I have learned a lot professionally and personally from you. I also want to express my gratitude to Bernward Klocke from the Genomatix Company. I loved working with you on the project during all the years and I learned a lot from our discussions on the project.

Throughout the years of my project, I have been very fortunate to have extraordinary collegues with whom I discussed my project, shared the office, had a lot of fun in the lab during fat prep or EMSA experiments, shared amazing times at conferences, in diverse Bavarian biergärten, at sushi places, at the Isar, in the Alps and during extended coffee breaks at the Institute. I want to express my gratitude to Lissy, Manu, Lydia for their incredible technical assistance and Sylvi for her never failing understanding for “late-breaking Dienstreiseantrag Submissions” and long discussions about everything and nothing in her office. I deeply thank Simon, the best collaboration partner ever, for his devotion for the project, especially during the resubmission phase and especially his strong friendship.

I am enormously grateful to my family, especially my father but also all the rest of my family. All of you, Bernd and my “adopted” Griegel/Seehafer family, Oma Inge, my Strobels, my close friends, Tobi, Kerstin, Vroni, Sophie, Alex and Wendelin inspire me on a non-scientific level, which is so important for me and build the basis for everything included in this thesis.

I dedicate this thesis to my beloved grandmother Elisabeth.

Table of contents

Summary

Genome-wide association studies (GWAS) revealed a plethora of risk loci associated with a diverse range of diseases and traits. However, identification and mechanistic elucidation of the disease-causing variants within association loci remains a major challenge in medical genetics. Divergence in gene expression due to *cis*-regulatory variation is central to disease susceptibility. In this thesis, I developed a computational methodology, called Phylogenetic Module Complexity Analysis (PMCA), to identify the specific *cis*-regulatory functional variants as a prerequisite to elucidate their mechanisms in disease. In general, PMCA exploits phylogenetic conservation in terms of a complexity assessment of co-occurring transcription factor binding sites (TFBS) within *cis*-regulatory modules (CRMs) regardless of cross-species conservation of the complete sequence, to effectively identify *cis*-regulatory variants within GWAS-associated disease risk loci. I draw on the PMCA methodology to study type 2 diabetes (T2D) susceptibility loci where the specific underlying causal variants are poorly characterized and the diverse sets of mechanisms by which they may increase disease risk have not been elucidated. Systematic integrative analysis of the total set of established T2D risk loci, by PMCA, could reveal an unexpected clustering of distinct homeobox TFBS at T2D risk SNP positions. In-depth analysis at the *PPARG* diabetes risk locus unveiled a novel activity of the PRRX1 homeobox transcription factor as a repressor of *PPARG2* expression and showed its dysregulation in primary human adipose cells from rs4684847 risk allele carriers, resulting from SNP-mediated increase of PRRX1 binding affinity. Pursuing the computational inferences further revealed that PRRX1, via enhanced binding at the rs4684847 C risk allele perturbs glyceroneogenesis thereby provoking dysregulation in FFA turnover, lipid metabolism and systemic insulin sensitivity. Overall, identification of the *cis*-regulatory variant rs4684847 at the *PPARG* locus, by PMCA, enabled linking the molecular upstream

factor PRRX1 to aberrant downstream mechanisms of impaired lipid handling and insulin sensitivity, explaining the GWAS association with T2D. Apart from using the PMCA methodology for T2D risk loci, I could also utilize parts of the methodology for the analaysis of sequence variants within the adiponectin level-associated adiponectin promoter region and for the analysis of cross-species conserved TFBS within the *BOB.1/OBF.1* promoter pointing to its general usability for the identification of both, regulatory genomic regions and *cis*-regulatory sequence variants. Together, these results of this thesis demonstrate that cross-species conservation analysis at the level of co-occurring TFBS provides a valuable contribution to the translation of genetic association signals to disease-related molecular mechanisms.

Zusammenfassung

Genomweite Assoziationsstudien (GWAS) haben eine Vielzahl von genomischen Regionen mit verschiedensten phenotypischen Merkmalen und Erkrankungen assoziiert. Die Identifizierung von genetischen Varianten, die eine Erkrankung hervorrufen, und die Aufklärung der zugrundeliegenden Mechanismen ist jedoch nach wie vor eine schwierige Herausforderung im Feld der medizinisch ausgerichteten Genetik. Die Suszeptibilität von Erkrankungen wird vornehmlich durch Variabilität der Genexpression bestimmt. In dieser Arbeit habe ich eine bioinformatische Methode namens „Phylogenetic Module Complexity Analysis“ (PMCA) entwickelt, mit dem Ziel *cis*-regulatorische genetische Varianten zu identifizieren und damit die Grundlage zu schaffen, deren erkrankungsrelevante Mechanismen aufzuklären. PMCA macht sich dabei, anders als die herkömmliche Beurteilung von rein Sequenz-basierter Konserviertheitsanalyse, die phylogenetische Konservierung von gemeinsam auftretenden Transkriptionsfaktor Bindungststellen (TFBS) in regulatorischen Elementen des Genoms zu Nutze und schätzt deren Komplexität ein, um *cis*-regulatorische Varianten in GWAS-assoziierten Regionen effizient zu lokalisieren. Die entwickelte PMCA Methode wurde zur Analyse von genomweit assoziierten Typ 2 Diabetes (T2D) Risikoregionen, deren spezifische kausale Varianten und zugrundeliegenden Mechansimen weitestgehend unbekannt sind, herangezogen. Die systematische integrative PMCA Analyse aller etablierten T2D Riskoregionen konnte eine überraschende signifikante Überrepresentierung von bestimmten Homeobox TFBS exakt an Position von T2D Risikovarianten aufdecken. Eine eingehende Analyse des T2D-assoziierten *PPARG* Lokus bestätigte eine Rolle des Homeobox Transkriptionsfaktors PRRX1 als neuer Repressor von *PPARG2* und darüber hinaus dessen veränderte Bindungsaffinität zu dem rs4684847 Allel in primären humanen Präadipozyten. Die bioinformatischen Rückschlüsse zeigten

darüberhinaus, dass PRRX1 durch erhöhtes Binden an das rs4684847 Risko Allel (C Allel) die Glyceroneogenese in Präadipozyten beeinträchtigt und somit zu einer Störung des freien Fettsäurehaushalts, des Lipidmetabolismus und der systemischen Insulinsensitivität führt. Somit konnte die PMCA-basierte Identifizierung der *cis*-regulatorischen Variante rs4684847 im *PPARG* Lokus den Link zwischen dem Faktor PRRX1 und dem T2D zugrundeliegenden Mechanismus herstellen. Abgesehen von der PMCA-basierten Analyse von T2D-relevanten Suszeptibilitätsloki konnte ich Teile des bioinformatischen Ansatzes zum Einen für die Analyse von Sequenzvarianten innerhalb der mit Adiponektinspiegeln assozierten Adiponektin Promoterregion anwenden und zum Anderen evolutionär konservierte TFBS innerhalb des *BOB.1/OBF.1* Promotors lokalisieren und somit die generelle Nutzbarkeit von PMCA für die Identifizierung von regulatorischen Regionen und *cis*-regulatorischen Varianten validieren. Die Ergebnisse dieser Arbeit demonstrieren in ihrer Gesamtheit, dass die Analyse von konservierten Mustern gemeinsam auftretender TFBS einen wertvollen Beitrag zur Translation von genetischen Assoziationssignalen in molekulare Erkrankungs-Mechansimen leisten kann.

1 Introduction

1.1 Complex traits genetics and complex trait mapping in humans

During the last decade, human disease genetics has experienced a revolution. Since the first draft of the human genome sequence – the blueprint of life - was released to the public by two independent groups (Lander et al., 2001; Venter, 2001), the genome-wide study of heritable and somatic human variability became reality, covering a wide spectrum of diseases and traits. However, the big promise that within the genome we would find the blueprints for human function and dysfunction, i.e. disease, is yet to be fulfilled. One of the main challenges in the post-genomic era is to translate the wealth of genomic data into understanding the genetic basis of human disease.

In 1859, Charles Darwin recognized that natural selection requires variation in traits that is passed along to offspring. Genetic variation describes naturally occurring genetic differences among individuals of the same species. It is estimated that any individual genome harbors approximately 3 million variants which translates into a variation at every 1,000 nucleotides (ENCODE Project Consortium, 2012). The most common form of genetic variation between individuals is called single nucleotide polymorphisms (SNPs). Other types of variation include insertions and deletions (indels), copy number variants (CNVs), microsatellites and structural variations. The term SNP was initially referred to as a nucleotide polymorphism above a 5% frequency threshold in the population, and is currently referred to as a position different to the human reference sequence (Human Genome Sequencing Consortium, 2004), regardless of the frequency in the population.

Generally, genetic disorders are categorized into monogenic Mendelian diseases and complex disorders. Mendelian diseases, in which variation in a single gene is both necessary and sufficient to cause disease, run predictably and consistently in families. Common complex diseases in contrast afflict most of the population and have a multifactorial, polygenic background. Although it seems paradoxical that deleterious alleles reach relatively high frequencies in a population, those disease causal common alleles persist due to evolutionary forces such as random drift or natural selection (McClellan and King, 2010). Examples include late onset of diseases without effects on reproductive fitness over longer periods (e.g. Alzheimer´s disease), pleiotropic effects of variants that result in balancing selection (Wagner and Zhang, 2011), geographic-specific variation, and diseases resulting from recent changes in living conditions such as obesity, heart disease and diabetes. The analysis of common complex disorders is particularly challenging because they do not follow simple Mendelian inheritance patterns and are additionally characterized by multiple gene-gene interactions (multiple loci interact to increase disease susceptibility), gene-environment interactions, allelic heterogeneity (different alleles at the same locus increase disease susceptibility), locus heterogeneity (mutations at different loci increase disease susceptibility), incomplete penetrance (not all individuals inheriting the disease-predisposing allele manifest the phenotype) and pleiotropy (a genetic variant influences more than one phenotype).

For rare monogenic traits, family-based linkage studies and positional cloning have been spectacularly successful for ‘Mendelian’ disorders in annotating functional consequences because genetic risk factors are highly penetrant. Two well-known examples are the *CFTR* gene in cystic fibrosis (Riordan et al., 1989) and the *HTT* gene in Huntington´s disease (Gusella et al., 1983). In contrast to Mendelian disorders, extensive linkage analysis and positional cloning in the 1990s failed to identify the genes underlying specific common diseases, owing to the lack of strong linkage signals for a single gene locus (Risch and

Merikangas, 1996). This is reasonable because common diseases result from the complex interplay of environmental factors and many different individual low relative risk susceptibility genetic variants, where each variants explains only a small proportion of the total population risk (Cardon and Abecasis, 2003). The penetrance (effect size) of individual common causal variants is therefore too small to allow detection via cosegregation within pedigrees. To overcome the fundamental barrier of polygenicity in common disease and trait mapping, Risch and Merikangas, 1996 showed in a landmark theoretical paper that a large-scale population-based association mapping involving one million variants in the genome and a sample of unrelated individuals could be more powerful than performing family-based linkage analysis with a few hundred markers. However, it has become clear that the introduction of the genome-wide study of disease- and trait-related genomic loci would require a paradigm shift which was realized by several parallel critical advances:

1. *Comprehensive catalogue of human genetic variation*

   During the last decade, rapid progress in genome-wide genotyping and sequencing by the HapMap Project (Lander et al., 2001; http://hapmap.ncbi.nlm.nih.gov/) and the 1000 Genomes Project (1000 Genomes Project Consortium, 2012; http://www.1000genomes.org/) has allowed for mapping of human genetic variation within and between populations. The result was a comprehensive catalogue of common (minor allele frequency (MAF) $\geq$ 5%), lower-frequency (0.5-5%) and very rare/low-frequency (< 0.5%) genetic variation, including 38 million SNPs, 1.4 million bi-allelic short indels, and 14,000 larger deletions, and a number of structural variants (The International HapMap Consortium, 2005; 2007; 1000 Genomes Project Consortium, 2010, 2012). The total number of common mapped SNPs identified to date exceeds 10 million most of which likely have no functional significance and persist by chance in the absence of selective pressure.

2. *Haplotype structure of the human genome*

Association studies typically use phenotypes measured in collections of unrelated individuals and dense marker (tagSNP) information to detect statistically significant correlations between a marker genotype and the analyzed trait. This principle relies on linkage disequilibrium (LD) which is defined as the nonrandom association between alleles at different loci, based on the fact that nearby SNPs tend to be strongly correlated with each other across individuals. When a new mutation arises on a genetic background, it is in complete association with all of the variants present on that chromosome. Over time these associations are broken down by homologous recombination (Hartl and Clark, 2007). Generally, loci that are physically close together, which therefore are rarely affected by recombination events, exhibit stronger LD than loci that are farther apart on a chromosome (Hill and Robertson, 1968). Common genetic variations are therefore tightly correlated in structures called haplotypes that have not been broken up by meiotic recombination events and are separated by recombination hotspots that occur every 100-200 kb (Reich et al., 2002) and that are inherited together through generations (Paigen and Petkov, 2010). The haplotype structure of the human genome, reflecting recent expansion from a small founding population (Hartl and Clark, 2007), was investigated by the International HapMap project, which was launched in 2002. The outcome was a limited set of SNPs, around 500,000, that could cover ~90% of the common genetic variation in the population.

3. *High-throughput genotyping technologies*

Taking advantage of the identified correlation structure of the human genome, genome-wide genotyping of SNP marker panels became possible with the introduction

of microarrays (Gunderson et al., 2005). Those high-throughput technologies enabled the unbiased detection of statistically genome-wide significant differences in genotype frequencies between thousands of unrelated individuals who have the analyzed phenotype and the general population. In a case-control genome-wide association study, large population-based sample collections of unrelated individuals are genotyped typically for a set of 100,000-2,000,000 markers to detect common variants associated with complex disease and diverse molecular phenotypes. Their application has facilitated detection of even modest risk alleles, with odds ratios of less than 1.1.

The theoretical concept for the systematic unbiased study of common genetic variants was therefore realized in 2006 – with the introduction of genome-wide association studies (GWAS) - an important turning point for human genetics studies.

1.2 The GWAS revolution and its challenges

In the last seven years, GWAS yielded a plethora of loci associated with diverse traits, such as body height and body mass index (BMI, kg/m^2), as well as with the entire spectrum of complex diseases, including neurological disorders, inflammatory diseases, different types of cancer, cardiovascular diseases, and metabolic disorders (Burton et al., 2007a; Burton et al., 2007b; Hakonarson et al., 2007; Rioux et al., 2007; Scott et al., 2007; Zeggini et al., 2007; Eeles et al., 2008; Zeggini et al., 2008). One of the major advances of the GWAS approach – in contrast to candidate gene studies which depend on prior knowledge of the trait - is that it permits an unbiased scan of the human genome which facilitates the identification of novel trait- and disease-related loci without requiring any a priori knowledge. However, despite the introduction of those novel genomic technologies and analytical strategies, signals emerging

from GWAS have rarely been traced to causal variants and even more rarely to the elucidation of disease-related mechanisms.

One of the confounding factors arising from those association studies is linkage disequilibrium (Figure 1). Complex trait mapping by GWAS is based on the premise that a causal variant is located on a haplotype, and therefore a marker allele in LD with the causal variant should be associated only by proxy with the trait of interest. This means that once a variant association to a phenotype is discovered, the probability of any given associated tag variant being causal is miniscule. In fact, when the authors of The 1000 Genomes Project evaluated GWAS hits in Europeans, they found that each signal is, on average, in high LD with 56 variants (51.5 SNPs and 4.5 indels) (1000 Genomes Project Consortium, 2012). GWAS signals are therefore simply markers for large genomic regions in LD that contain many genes, in which the phenotypically causal variants are hidden (Hindorff LA; The ENCODE Project Consortium, 2012). In that scenario, regions of strong LD can be large, and tag SNPs associated with a phenotype have been found to be in perfect LD with potential causal SNPs several hundred kilobases away.

While GWAS have identified disease-associated genomic regions, they have hardly been able to find the specific causal sequence variants and the underlying disease mechanisms. The identification of causal variants has been largely hampered by the fact that large majority of disease-associated variants map in non-coding regions of the genome (93% of disease-associated variants are non-coding, Maurano et al., 2013), where it remains a major challenge to predict and assay a variant's phenotypic impact (see Section 1.4 `The Role of Non-coding Variation in Human Traits and Diseases`). Initial approaches for pinpointing the few phenotypically causal variants among the many variants present in the genome have been largely limited to predicting the effects of protein-coding sequence changes (often called "the

low hanging fruits") based on constraint-, biochemical- and structural-based prediction tools such as PolyPhen-2 (Ng and Henikoff, 2003; Adzhubei et al., 2010; Ward and Kellis, 2012).

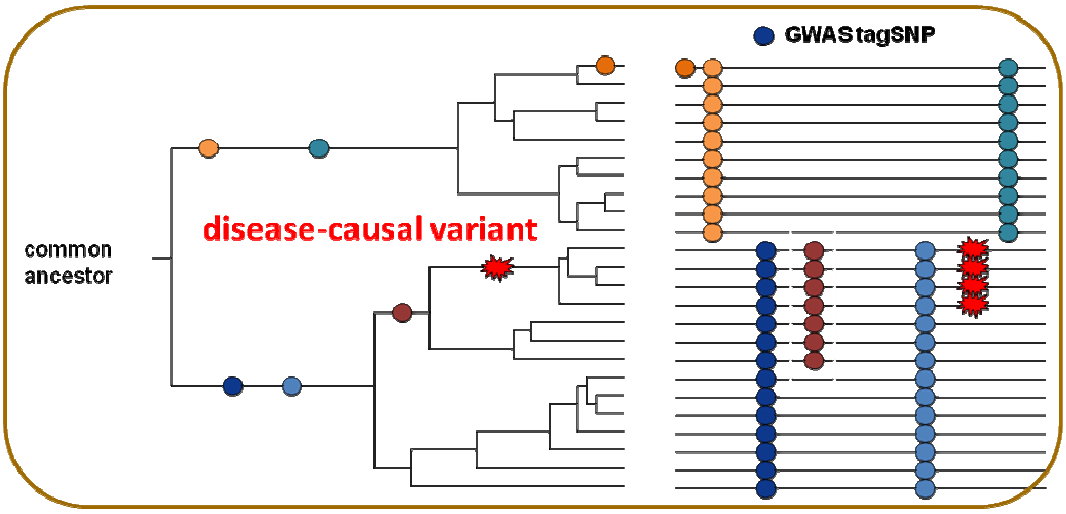


Figure 1: The resolution of results from GWAS and whole genome sequencing studies is mainly limited by the linkage disequilibrium (LD) structure of the human genome. Particular alleles, indicated as coloured circles, at neighboring loci tend to be co-inherited even though recombination limits the extent of the region of association over time. The disease-causing mutation is indicated by a red star. GWAS tag SNP markers that are physically close tend to remain associated with the ancestral mutation.

Overall, the recent genotyping and sequencing technologies have identified a great number of disease- and trait-associated genomic loci as an important step towards understanding disease genetics, but we now face the challenge of identifying the causal variants and a mechanistic understanding for how these variants contribute to disease.

A recent Nature Genetics editorial on post-GWAS analyses began to put this problem into sharper focus and suggested that there should be a more significant investment in functional characterization of risk loci (On beyond GWAS, 2010). In his '2011 vision for the future of genomics research' E. Green from the Human Genome Research Institute at the NIH projected that the current development of genomics will propel personalized medicine approaches (Green and Guyer, 2011). This 'path towards an era of genomic medicine'

strongly depends on establishing 'novel strategies for identifying non-coding variants that influence disease' (page 207). In this thesis, I aimed to develop a systematic means for identifying the specific functional variants within elusive tag SNP-containing LD blocks that have been associated with disease via GWAS. A common complex disease, type 2 diabetes (T2D) was analyzed as proof-of-concept.

1.3 The genetic architecture of type 2 diabetes

The global epidemic of type 2 diabetes (T2D) is a major thread of 21st century and is a leading cause of morbidity and death worldwide, contributing to development of chronic complications such as coronary heart disease, stroke, vascular diseases, renal failure and amputation (Shaw et al., 2010). In 2011, 366 million people suffer from diabetes and the prevalence is expected to rise dramatically worldwide to 552 million by 2030 (http://idf.org/diabetesatlas/5e/the-global-burden). The global health care expenditures on T2D are expected to increase from 338 billion dollars in 2010 to 440 billion dollars in 2030 (Zhang et al., 2010). Mechanistically, T2D arises from impairment in the ability of fat, muscle, and liver to respond to insulin, i.e., insulin resistance, followed by a failure of the pancreatic β-cell to secrete adequate insulin in response to glucose, thereby leading to hyperglycemia (Kahn, 1994). T2D is a genetically heterogeneous disease, and is thought to result from the complex interaction of environmental factors acting on a susceptible genetic background. The heritable component of T2D is obvious from the familial clustering of T2D (40% versus 6% in first-degree relatives of T2D patients versus general population), the higher concordance rate of T2D in monozygotic compared to dizygotic twins, and the high T2D prevalence in specific ethnic groups, *e.g.,* Pima Indians and Mexican Americans (Doria

et al., 2008). The genetics component of T2D risk (heritability) is generally estimated to account for 0.49 (Lander and Schork, 1994). Until 2007, the search for T2D genetic factors was rather unsuccessful with 3 gene loci identified by linkage analysis and candidate gene approaches (*PPARG*, *TCF7L2*, *KCNJ11*). Over the past years, the introduction of GWAS and full-sequencing studies boosted the exploration of T2D genetics. Since 2007, GWAS have reproducibly identified a plethora of metabolic risk loci, including 65 susceptibility loci robustly associated with risk of T2D (Prokopenko et al., 2009; Bonnefond et al., 2010; Dupuis et al., 2010; Voight et al., 2010) (Figure 2). Those genotyping efforts have unequivocally shown that the majority of T2D affecting loci map to non-coding regions in the human genome (Hindorff et al., 2009). The affected genes and transcript isoforms responsible for mediating the effect mostly remain unknown. This gap relies on the fact that moving from the dearth of candidate variants located in those associated regions to the etiological T2D underlying genetic mutation has rarely been achieved, except for the *TCF7L2* (Gaulton et al., 2010) and *WFS1* (Stitzel et al., 2010) loci. Yet, pinpointing diabetes-causal variants, the genes that they affect and the regulatory mechanisms that might converge on T2D loci is a prerequisite to understanding the biological mechanisms underlying T2D pathogenesis, and to eventually develop prognostic and diagnostic means and targeted pharmaceutical interventions.

Figure 2: The Genetic Architecture of type 2 diabetes. All GWAS associations with p-value $\leq 5.0 \times 10^{-8}$, published in the GWAS catalogue (http://www.genome.gov/gwastudies) up to the end of October 2013 (indicated as coloured circles). Genomic loci associated with type 2 diabetes are highlighted in orange.

1.4 The role of non-coding variation in human traits and diseases

Changes in patterns of gene expression are widely believed to underlie many of the phenotypic differences within and between species. Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product, i.e. proteins or functional RNA molecules such as ribosomal RNA (rRNA), transfer RNA (tRNA) or small nuclear RNA (snRNA). This flow of genetic information happens one way as described by

Francis Crick in his early *Central Dogma* of molecular biology referring to the transfer of genetic information from DNA to RNA to proteins (CRICK, 1970). Gene expression in eukaryotes is regulated at different levels including transcription, RNA splicing, translation, and post-translational modification of a protein. Transcriptional regulation is achieved through combinatorial interactions between regulatory elements and a variety of factors that modulate the recruitment and activity of RNA polymerase (Figure 4). Briefly, these transcriptional regulators include (1) transcription factors, which bind specific DNA sequences within proximal promoter elements or enhancer regulatory regions; and (2) chromatin remodellers, which affect the chromatin structure through conformational changes or by covalently modifying histone tails, such as polymerases, helicases, topoisomerases, kinases, chaperones, proteasomes, acetyltransferases, deacetylases and methyltransferases. (3) The preinitiation complex, which binds at the core promoter and recruit RNA polymerase II (Pol II). Overall, gene expression is tightly controlled by gene regulatory regions, so called *cis*-regulatory modules (CRM), comprising clusters of transcription factor binding sites that govern the recruitment of chromatin modifiers and cofactors (see Section 1.5 "Importance of Computational Approaches for *Cis*-regulatory Variant Discovery" for detailed description of the role of TF binding in CRM for gene regulation). In that scenario regulatory elements involve (1) promoters: short stretches of DNA sequence, typically within 200bp of the transcriptional start site (TSS) of a gene, which are composed of the core promoter and nearby proximal regulatory elements; (2) enhancers: long-distance transcriptional regulatory elements, which control gene expression in a highly spatial and temporal manner (Visel et al., 2009). (3) silencers: binding sites for transcription factor that reduce transcription by competitively binding at an activator binding site, blocking the binding of a transcription factor, or recruiting chromatin-modifying factors; (4) Insulators: DNA stretches (typically 0.5-3kb) that prevent inappropriate gene regulation by precluding enhancer – promoter

interactions and spreading of repressive chromatin; (5) Locus Control Regions: multiple cis-regulatory elements cooperatively acting on clusters of genes, thereby detrmining the context-dependent gene expression.

The question thus arises: why is non-coding variation so important for evolution, species development, cell differentiation processes and eventually complex phenotypes? Emile Zuckerkandl, one of the pioneers of molecular evolutionary biology, proposed in 1964 that phenotypes could change by altering the timing or rate of protein synthesis (Zuckerkandl and Pauling, 1965) (Figure 3). It is obvious that despite the phenotypic diversity of species, the DNA sequence diversity is surprisingly limited. Mary-Claire King and Allan Wilson underscored this paradox by comparing the chimpanzee and human genome and the challenge "to explain how species which have such substantially similar genes can differ so substantially in anatomy…" (King and Wilson, 1975). From this landmark paper came the idea that changes in gene regulation, not differences in protein sequences, drives phenotypic divergence in morphology and are responsible for adaptive evolution within and between species (King and Wilson, 1975; Bamshad et al., 2002; Hamblin et al., 2002; Tishkoff et al., 2007). A number of examples in humans have been described that underscore the role of regulatory variation as an important component for evolutionary change (Bamshad et al., 2002; Hamblin et al., 2002; Bersaglieri et al., 2004; Tishkoff et al., 2007). One of the first and well known examples was the lactase persistence phenotype in European populations; Ennattah and colleagues mapped a distal enterocyte-specific *cis*-regulatory variant upstream to *LCT*, a gene encoding the lactase enzyme in the small intestine (Enattah et al., 2002).

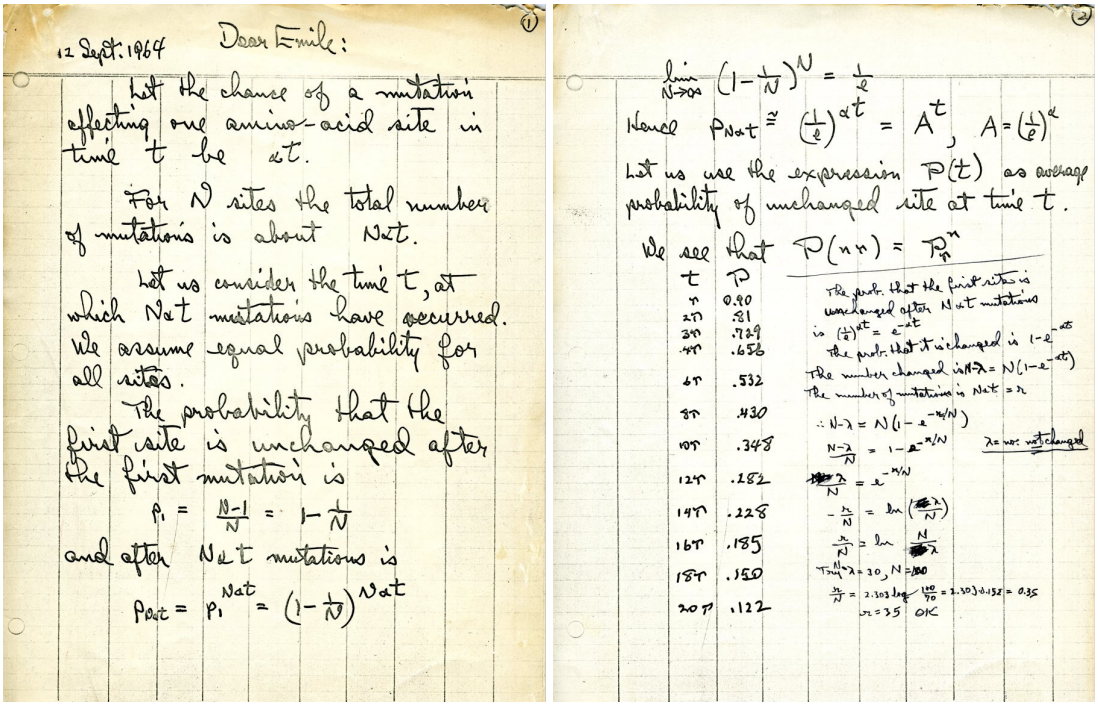


Figure 3: Letter from Linus Pauling to Emile Zuckerkandl. September 12, 1964. In his letter to Zuckerkandl, Pauling details the logical and mathematical foundation underlying his thinking on evolution and molecular disease (adapted from http://osulibrary.oregonstate.edu/specialcollections/coll/pauling/blood/corr/corr465.7-lp-zuckerkandl-19640912-02-large.html).

Regarding complex disease genetics, GWAS have shown that 93% of strongly disease- or trait-associated variants emerging from GWAS localize within the non-genic region of the genome and may affect gene expression rather than changing protein structure (Maurano et al., 2013). Indeed, the variability of gene expression is highly heritable (Dixon et al., 2007; Stranger et al., 2007), and common trait-associated loci in non-coding regions are highly enriched for expression quantitative trait loci (eQTLs) (Nicolae et al., 2010; Nica et al., 2011), suggesting that many common disease variants act by altering transcript levels. Very recently, the large-scale ENCODE project revealed that regulatory variants are pervasive throughout the genome – with 3.85 million and 1.01 million variants overlapping DNase hypersensitive sites (DHS) peaks and DHS footprints, respectively (The ENCODE Project Consortium, 2012). More importantly, disease- and trait-associated variants were shown to be highly

enriched within regulatory DNA marked by DHS peaks, DHS footprints and ChIP-Seq peaks (Maurano et al., 2012a; Schaub et al., 2012; The ENCODE Project Consortium, 2012), strongly indicating that variants modulating gene regulation are major contributors to common disease susceptibility among individuals and their functional understanding is thus critical to interpret GWAS and sequencing data. In this thesis, I thus concentrated on the integration of functional and genomic data to elucidate non-coding variants that may mechanistically mediate genetic predisposition to a disease.

1.5 Importance of computational approaches for *cis*-regulatory variant discovery

G. Cooper and J. Shendure provided a state of the art review on 'approaches to estimate the deleteriousness of single nucleotide variants', pointing out both the power and limitations of current experimental and computational approaches (Cooper and Shendure, 2011). They emphasized the imperative of the 'precise delineation of causal variants' located in the non-coding genome as the 'fundamental goal of human genetics'. Recent chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (Park, 2009) and DNase hypersensitivity analysis (DNase-seq) that map histone modification marks (Boyle et al., 2008), as well as data on accessible chromatin regions, have been used to prioritize candidate functional *cis*-regulatory variants out of a much larger number of candidate variants in GWAS-inferred LD blocks (Dimas et al., 2009; Gaulton et al., 2010; Ernst et al., 2011; Nica et al., 2011; Ward and Kellis, 2011; Maurano et al., 2012b; The ENCODE Project Consortium, 2012). However, experimental approaches based on harnessing functional genomics data have the disadvantage that they require access to appropriate disease-/trait-

relevant human tissue, or to tissue from a particular developmental time stage (which frequently is impossible), and they are further hampered by the spatial, temporal, environmental and epigenetic complexity of gene regulation (Dimas et al., 2009; Nica et al., 2011). Given that it is not possible to assay all candidate variants in all tissue types under all physiological relevant conditions means that potential causal SNPs will likely be missed. These constraints imply a great need for bioinformatics approaches that could reliably assess the regulatory role of specific non-coding variants, while still rather elusive, would be highly desirable.

So far, phylogenetic conservation at the sequence level has been a common denominator in the search for regulatory regions in the non-coding parts of the genome (Pennacchio et al., 2006; Visel et al., 2009; Lindblad-Toh et al., 2011). This is based on the reasonable hypothesis that evolutionary conservation is a strong indicator of molecular functionality (Pennacchio et al., 2006; Visel et al., 2009; Lindblad-Toh et al., 2011). Under the assumption that deleterious mutations within genomic regions with sequence-specific functionality will be removed by purifying selection as opposed to sequences without functionality that will accept mutations at an underlying neutral rate, conservation of sequences that share common ancestry could indicate coding and functional non-coding regions within the genome. Therefore, considering that evolution may be regarded as the ultimate mutagenesis experiment, comparative sequence analysis has been proposed to infer deleteriousness of a genomic mutation (Cooper and Shendure, 2011). In disease genetics, several computational approaches that use evolutionary conservation have therefore been proposed to predict coding (Adzhubei et al., 2010) and non-coding candidate variants for follow-up (*e.g.,* SiPhy π conservation algorithm Lindblad-Toh et al., 2011 and Genomic Evolutionary Rate Profiling GERP, Cooper et al., 2005). Yet, those algorithms based on pure sequence alignment have been only successful in identifying deleterious protein-altering variants from exome studies

and many software tools estimating the functional impact of a specific amino acid substitution in a protein are now publically available. Until very recently, it was widely thought that sequencing more and more vertebrate genomes for comparative analysis, might eventually serve to identify phenotypically causal non-coding mutations. However, although sequences that are critical for organism development, reproduction and survival reveal strong selective constraint (Nobrega et al., 2003; Visel et al., 2009), genome-wide comparative studies have indicated that the fraction of bases under selection corresponds to a minimum of the functional genome: "many functional elements are seemingly unconstrained across mammalian evolution" (The ENCODE Project Consortium, 2012). Indeed, the majority of transcribed and regulatory elements in the genome differ between closely related species (Dermitzakis and Clark, 2002; Kasowski et al., 2010) and between individuals within the same population (Stranger et al., 2007), making their evolutionary constraint-driven detection particularly challenging. Data suggest that the majority of binding sites for specific transcription factors (TF) is not constrained between species, reflecting the lineage-specific use of regulatory elements (Dermitzakis and Clark, 2002; The ENCODE Project Consortium, 2007; Blow et al., 2010; Schmidt et al., 2010; The ENCODE Project Consortium, 2012). Indeed, inter- and cross-species differences in gene expression are often driven by changes in transcription factor binding sites (TFBS) (Kasowski et al., 2010) and their rapid evolutionary turnover results in many regulatory regions that are functionally conserved but have little evidence of conservation at sequence level (Ludwig et al., 2000; Pennacchio et al., 2006; Sosinsky et al., 2007; The ENCODE Project Consortium, 2007; Dimas et al., 2009; Visel et al., 2009; Kasowski et al., 2010; Lindblad-Toh et al., 2011; The ENCODE Project Consortium, 2012). Thus, the predictive power of classical nucleotide-level alignment-score approaches remains limited for causal variant discovery in non-coding regions from GWAS and whole genome sequencing studies (Blow et al., 2010; Maurano et al., 2012b).

The *Central Dogma* of molecular biology refers to the transfer of genetic information from DNA to RNA to proteins (CRICK, 1970). Though the initial sequencing of the human DNA sequence resulted in 3 billion nucleotide "letters" and a surprisingly small number of 22,000 distinct protein-coding loci (making up little more than 1% of the genome), this sequencing effort alone has largely failed to reflect the complexity that make our individuals unique. The widely held view that the human genome is mostly 'junk DNA' was finally debunked by the recent ENCODE project showing that 80% of the genome contains elements linked to biochemical functions – including 2.89 million unique DHSs, 8.4 million distinct DNase I footprints, and 636,336 binding regions (ENCODE. 2012). During development, a single fertilized precursor cell gives rise to a highly complex, multicellular organism comprising a large variety of cell types and tissues. This process of cell differentiation and the determination of cell type morphologies and functions are generally achieved by the extraordinary complex and dynamic regulation of gene expression in response to environmental stimuli and developmental programs.

A major cellular process underlying the central dogma of molecular biology is *cis*-regulation. This process involves the binding of effector molecules to their binding sites in non–coding DNA regions. In eukaryotes, gene expression is typically controlled by multiple *cis*-regulatory genomic elements (Figure 4, *upper panel*), whose spatiotemporal-specific activities additively contribute to target gene expression (Yáñez-Cuna et al., 2013). These regulatory regions tend to be organized into *cis*-regulatory modules (CRMs) comprising clusters of transcription factor binding sites (TFBS) for the coordinated binding of transcription factors (TFs) (Figure 4) (Arnone and Davidson, 1997; Pennacchio et al., 2006; Gerstein et al., 2012).

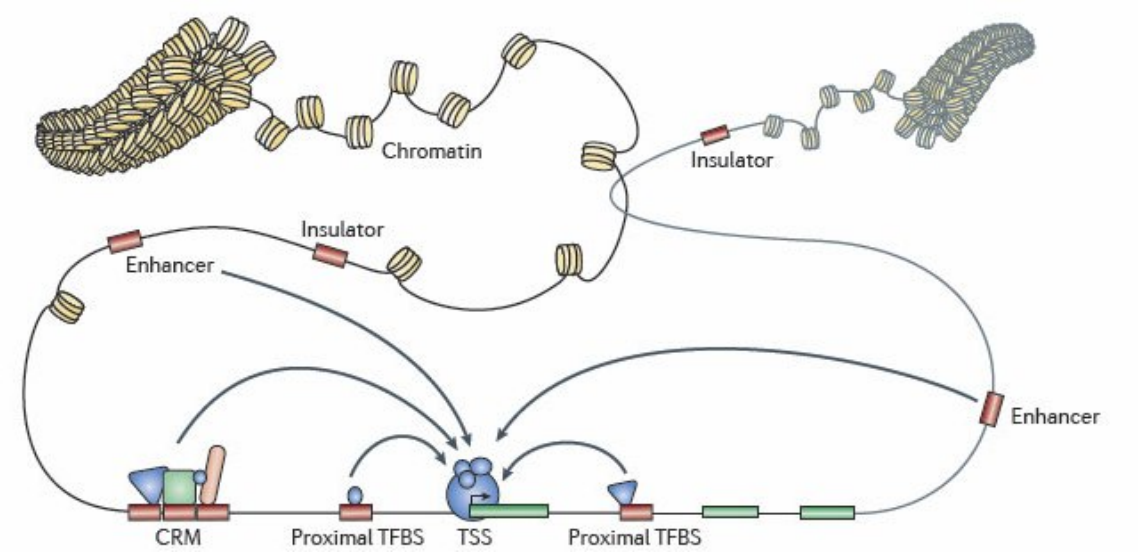


Figure 4: Regulation of gene expression (transcription). Chromatin is composed of DNA, which is wrapped around histones to form nucleosomes. Chromatin exists in a condensed, transcriptionally silent form (heterochromatin) and in a transcriptionally active form (euchromatin). Boundaries between heterochromatin and euchromatin may be marked by insulators. The region around the transcription start site (TSS) is often composed of a core promoter and a proximal promoter upstream of the TSS. Sequence-specific transcription factors bind to specific binding sites (TFBS) that are near to the TSS (proximal elements) or that are far away from it (enhancers), thereby recruiting RNA polymerase II to activate transcription of a gene. TFBS typically occur in clusters within so called *cis*-regulatory modules (CRMs). Adapted from Lenhard et al., 2012.

TFBS, the building blocks of the *cis*-regulatory code, are short DNA sequences - typically 9-12bp (Jolma et al., 2013) – that are required for sequence-specific TF binding. A TFBS motif is most often described as a position weight matrix (PWM): a model for a fixed length sequence that specifies the probability of each nucleotide at each position (Hardison and Taylor, Nature 2012). Hence, the scale in the most popular visualization of TFBS matrices, the so called LOGO, is in bits (information content). Upon binding to their specific TFBS, TFs govern the recruitment of chromatin modifiers and cofactors thereby stabilizing the transcription initiation machinery and eventually regulating the spatiotemporal-specific activation or repression of target gene transcription. Of importance, enhancers placed out of their endogenous genomic context recapitulate TF binding and DNA and histone

modifications, suggesting that the information for gene expression patterns is encoded within the primary DNA sequence of the enhancers (Yanez-Cuna et al., 2012).

CRMs thus integrate a variety of upstream signals – translating this information into the regulated expression of coordinated sets of genes, making them an obvious target to achieve broad phenotypic changes as a result of adaptive evolution (Blow et al., 2010; Schmidt et al., 2010; Taher et al., 2011). The computational CRM discovery is an important challenge in computational biology due to the plasticity of the mammalian regulatory DNA landscape and the high diversity and variability of TFBS. TFBS are variable in length and typically interspersed by gaps of neutral sequence. Moreover, even though TFBS often occur in specific combinations within a regulatory sequence, the order and arrangement of TFBS within CRMs of similar function is extraordinarily flexible. Thus, CRM are often functionally conserved, rather than conserved on a nucleotide level. As delineated above, the complex structure and flexibility of gene regulatory regions makes it particularly difficult to use existing computational approaches that assess sequence conservation or singleton TFBS annotation to detect functional regulatory genomic regions and within regulatory variants. Despite the critical importance of CRMs, no algorithms have so far been developed to harness the potential power of conserved TFBS patterns within CRMs to predict regulatory variants.

1.6 Aims of the thesis

Dissecting disease loci and their underlying molecular mechanisms depends on identifying the phenotypically causal disease variants. GWAS have largely failed to find the specific causal sequence variants that lead to proven disease mechanisms, despite that a plethora of highly significant SNPs have been found in the non-coding part of the genome. Although transcription is understood in broad conceptual terms, building predictive models for the identification of *cis*-regulatory sequence variants that affect transcriptional regulation has

proven challenging. Computational approaches that reliably assess the regulatory role of specific genetic variants would therefore be highly desirable. In this thesis, I concentrated on developing a computational methodology for finding and better defining functionally conserved regulatory regions, with the aim to help pinpointing *cis*-regulatory sequence variants that may explain disease associations by changing gene expression levels (Figure 5).

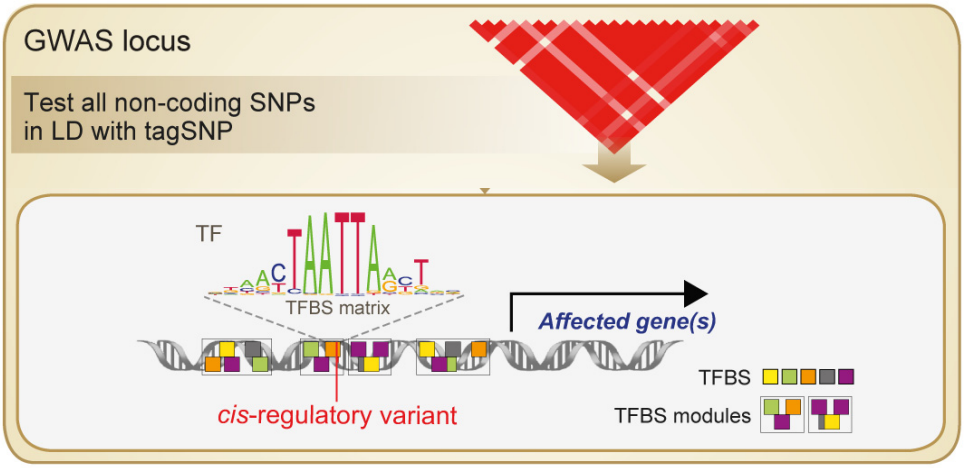


Figure 5: Discovery of *cis*-regulatory sequence variants and their affected disease genes via systematically testing all sequence variants within a GWAS-associated genomic region for the presence of conserved co-occurring TFBS patterns. One representative TFBS matrix is explicitly shown. TFBS: transcription factor binding sites; TFBS modules: two or more TFBS occurring in the same order and in certain distance range in all or a subset of the orthologous sequences.

I hypothesized that the presence of patterns of evolutionarily conserved TFBS in a CRM, within genomic regions surrounding a candidate variant is predictive of its *cis*-regulatory functionality, regardless of the cross-species conservation of the complete sequence on the nucleotide level (Figure 6). In order to test this hypothesis I developed a bioinformatics method, called "Phylogenetic Module Complexity Analysis" (PMCA), that is able to detect and classify genetic regions that contain evolutionary conserved TFBS patterns, thus leading to the identification of *cis*-regulatory genomic regions and the *cis*-regulatory sequence

variants within them that may mechanistically mediate genetic predisposition to a disease. The method relies primarily on regional co-occurrences and conservation of TFBS into clusters of transcriptional activity as a means to finding the functional regulatory variants and the effector molecules. I used primarily T2D as a showcase where novel *cis*-regulatory variants were pinpointed and novel TFBS were determined for T2D.

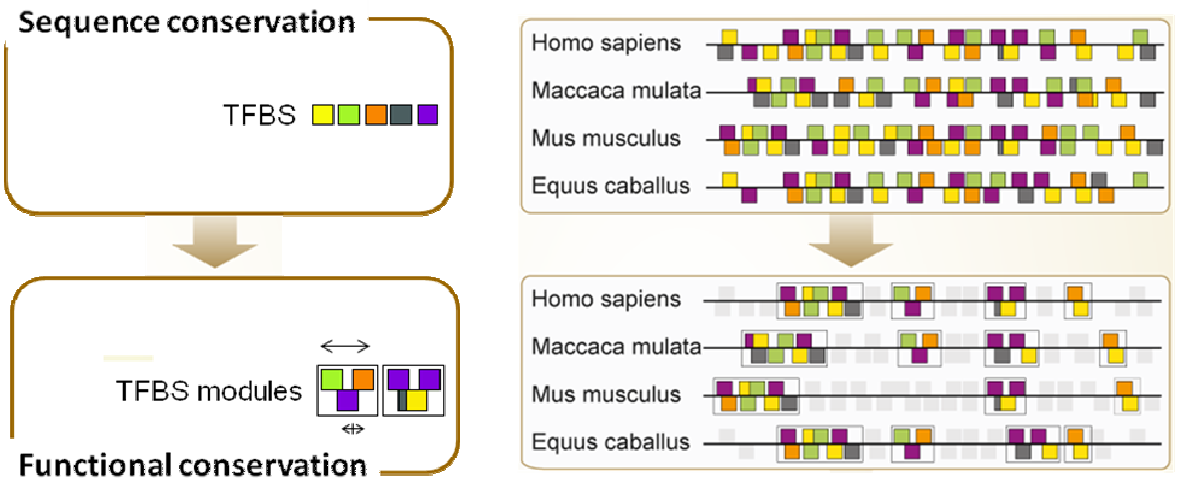


Figure 6: PMCA extends sequence conservation to functional conservation. The genomic location of TFBS is rapidly evolving, challenging the use of sequence alignment algorithms for localizing gene regulatory elements. To get around this issue, PMCA exploits the presence of complex patterns of evolutionarily conserved co-occurring TFBS, regardless of the cross-species conservation of the complete sequence, to effectively identify functional *cis*-regulatory variants with a role in disease.

In general, this thesis includes the computational PMCA methodology and its application on (1) GWAS-associated T2D risk loci; (2) the adiponectin locus associated with adiponectin levels; (3) the *BOB.1/OBF.1* promoter. Those results are either in press or recently published and are described in the following chapters.

2 Results

2.1 Computational Phylogenetic Module Complexity Analysis (PMCA) Methodology

The content of this chapter pertains to the computational "Phylogenetic Module Complexity Analysis" (PMCA) developed in this thesis and describes the general design of the method.

The PMCA methodology is the basis for the manuscript entitled "Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms" (in press at Cell, CELL-D-13-00664). In the course of developing the complexity-based PMCA framework, I used basic tools of the commercially available Genomatix software suite (Genomatix Co., Munich), i.e. the *RegionMiner* for extraction of orthologous regions and the *FrameWorker*, which extracts TFBS modules from a set of DNA input sequences. The processing of a large number of SNPs, and computation of randomized background distributions, needed the implementation of PMCA in a software interface. This was only possible in a collaborative work with Bernward Klocke, PhD (Genomatix, Munich) and with the constructive help of Karsten Suhre (Weill Cornell Medical College in Qatar) for writing the pseudo-code. PMCA can be run by using the command-line version and scripting of the processing and counting of the output (XML format). This chapter describes (1) the general design of the PMCA method and (2) 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 PMCA method in an automated manner. I refer to Appendix "PMCA

Procedures: General design of the PMCA method" for a step-by-step example for running PMCA manually using the Genomatix graphical user interface.

2.1.1 PMCA Procedures: 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 tagSNP for the risk of a given disease or a phenotype. For the analysis in this manuscript, we individually test all non-coding SNPs that are in linkage disequilibrium (LD, $r^2 \geq 0.7$) with the tag SNP (please note that any set of sequence 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. I used 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.

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 TFBS 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 ($\Omega_{modules}$), and occurrence of TFBS in TFBS modules ($\Omega_{TFBS\_in\_modules}$) 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 TFBS). To account for this possibility, we also determine phylogenetically conserved TFBS with more restricted parameters ($\Omega_{restr\text{-}TFBS}$, 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}, $\Omega_{restr\text{-}TFBS}$, $\Omega_{modules}$, and $\Omega_{TFBS\_in\_modules}$. 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 TFBS 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}, $\Omega_{restr\text{-}TFBS}$, $\Omega_{modules}$, and $\Omega_{TFBS\_in\_modules}$ and the estimated background probability of observing these counts by chance, we determine an overall classification criterion S_{all}.
9. The overall classification criterion labels the input region as *complex* or *non-complex*. (*Note***:** steps (1-9) are detailed in the **pseudo code** on page 21-23 of the Supplemental Experimental Procedures, Appendix).
10. To further select the variant with a function in disease, the overall disease-distinct clustering of TFBS at complex regions is assessed using positional bias analysis. (*Note*: the calculation of positional bias in step (10) is detailed in chapter 3 of the Extended Experiment Procedures, Appendix).

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). 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 I conducted experiments on DNA binding activity or reporter gene activity for a set of 62 T2D-associated 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. I thus expected the PMCA method to have over 90% selectivity and sensitivity.

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

Here, I 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, a step-by-step tutorial that can be followed manually using the interactive version of the Genomatix software (see provided screenshots in Appendix). While the RegionMiner and FrameWorker tools (Genomatix Co., Munich) presently represent the state-of-the-art, all steps in this method can be replaced by open-access tools and databases, such as AlignACE (Roth et al., 1998) for the identification of homologous regions, TRANSFAC (Matys et al., 2006) 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:

ζ = *i / 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 γ = *2 to 10 do the following*

 # γ is the number of TFBS 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 $\zeta >= 0.5$ sequence set to compute $\Omega_{restr\text{-}TFBS}$

2.4 Repeat the following 1,000 times

Randomly shuffle the sequence set S; use a sliding window of 10 bp and permutate 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}, $\Omega_{restr\text{-}TFBS}^{rnd}$, $\Omega_{modules}^{rnd}$, and $\Omega_{TFBS\_in\_modules}^{rnd}$.

3. Scoring and classification

3.1 Estimate the probability $p\text{-}est_i = f(\Omega_i^{rnd} > \Omega_i)$ 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 $p\text{-}est_i$ is set to 1/1001 if this never occurs:

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

3.2 Compute an Overall-score $S_{all} = -log(p\text{-}est_{TFBS} * p\text{-}est_{modules} * p\text{-}est_{TFBS\_in\_modules})$

3.3 Classify a non-coding SNP as being located in a complex region if and only if:
$(S_{all} > 6.5)$ and $(p\text{-}est_{restr\text{-}TFBS} < 0.15)$ and $(p\text{-}est_{TFBS} < 0.075)$
(Scoring criteria for classification)

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

The result of this project that was conducted during this thesis is in press at the journal Cell (CELL-D-13-00664). The content of the manuscript includes the established computational PMCA methodology and its application on recently GWAS-associated T2D susceptibility loci. The main text of the manuscript is included in this chapter and follows a general summary of the study`s highlights.

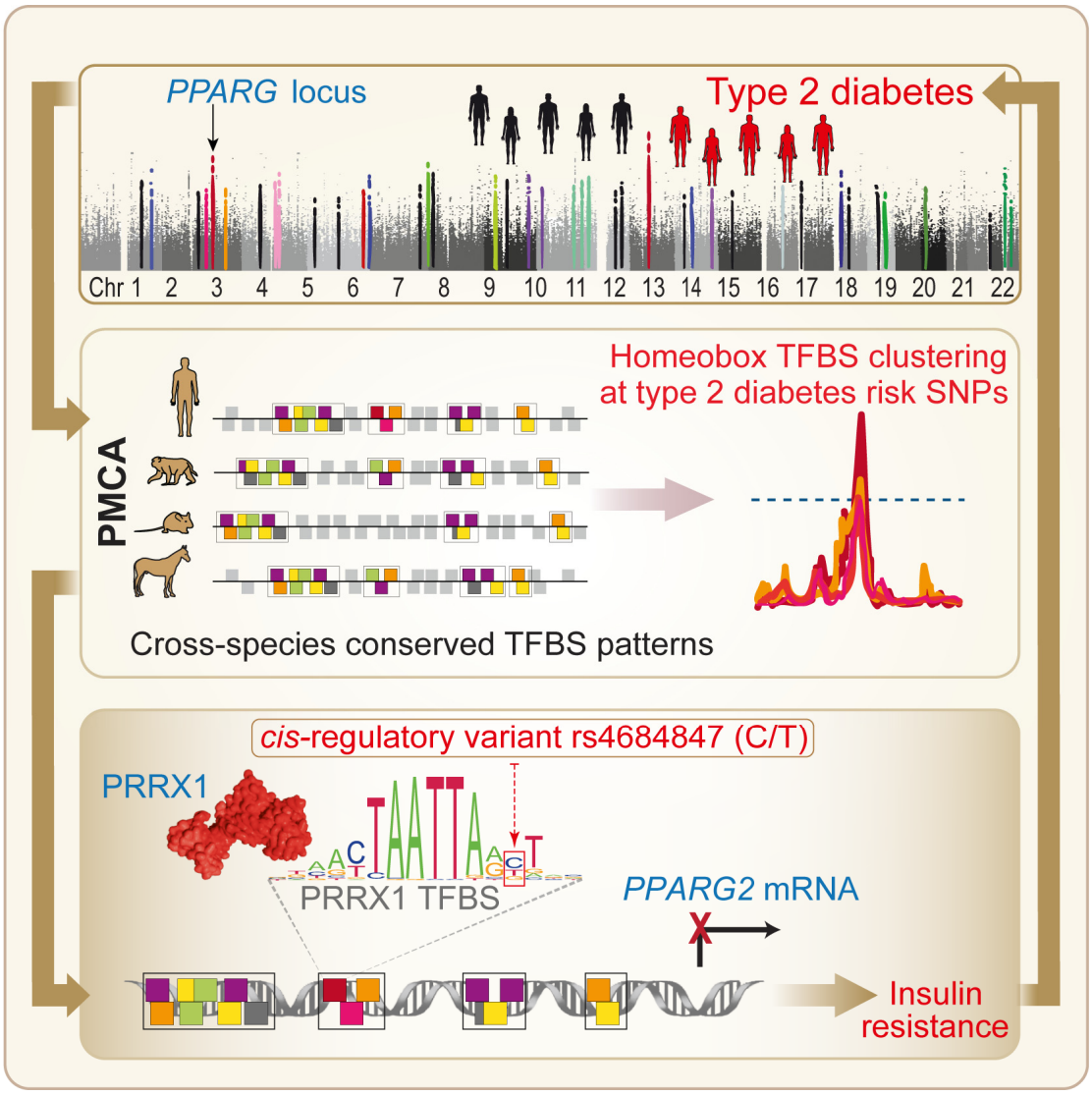
2.2.1 General Summary of the Study

The results of this study can be briefly summarized in the following Highlights and visualized in the Graphical Abstract shown below:

Highlights:

- ► Cross-species analysis of co-occurring TFBS predicts *cis*-regulatory variants.
- ► Analysis of diabetes-associated loci reveals clustering of distinct homeobox TFBS.
- ► The rs4684847 diabetes risk allele, by binding the homeobox TF PRRX1, represses *PPARG2* mRNA.
- ► PRRX1 perturbs lipid metabolism and insulin sensitivity dependent on rs4684847.

Graphical Abstract:



Graphical Abstract of the study: Genome-wide association studies (GWAS) have revealed a plethora of disease-associated risk loci in the non-coding genome, but the disease causal variants remain unknown in most cases. This study introduces a new approach for pinpointing *cis*-regulatory variants and their underlying disease mechanisms, based on phylogenetic conservation of co-occurring transcription factor binding sites (TFBS). The image illustrates patterns of TFBS conserved across humans and other vertebrate species, leading the way to *cis*-regulatory sequence variants, i.e. variants which affect gene expression and thereby disease risk. lllustration by Michael Pütz.

The majority of T2D affecting loci map to non-coding regions in the human genome (Hindorff et al., 2009), though the affected transcript isoforms responsible for mediating the effect are mostly unknown. Indeed, the specific *cis*-regulatory variants in the loci identified by GWAS approaches have rarely been pinpointed except for the *TCF7L2* and the *WFS1* loci (Gaulton et al., 2010; Stitzel et al., 2010), and the diverse sets of mechanisms by which they may increase disease risk are poorly characterized. In this proof-of-principle study, I applied PMCA on T2D genome-wide associated loci.

A feature of the ability to scan genome wide-associated non-coding genomic regions for functionality is the capability of agnostically identifying potentially informative *cis*-regulatory variants and their binding regulators in the absence of previous biological information. An important finding emerging from this study is the unexpected specific homeobox TFBS clustering at T2D regulatory risk SNPs, which was inferred from the PMCA approach and which distinguished T2D from different etiological trais: applying the computational inference procedure to all 47 GWAS-identified T2D susceptibility loci revealed a trait-specific clustering of distinct homeobox TFBS matrix families CART, PDX1, NKX6, HOMF, HBOX, BCDF. The specific clustering of homeobox TFBS matrices at T2D SNP positions in complex regions was in strong contrast to non-complex regions and distinguished T2D from other traits, i.e. asthma and Crohn`s disease. To evolve the biological impact of PMCA inferences, i.e. computational prediction of *cis*-regulatory sequence variants and distinctive homeobox TFBS enrichment in regard to T2D pathogenesis, we pursued different lines of evidence:

(1) In-depth experimental validation in primary human adipose cells at the *PPARG* locus

The missense SNP rs1801282 (Pro12Ala) at the *PPARG* locus is one of the T2D susceptibility loci reported by many candidate gene association studies as well as GWAS. The confounding

factor arising from association studies becomes clear when considering that there are 23 non-coding variants in high LD ($r^2 \geq 0.7$, ranging from position Chr3, 12393125 to position Chr3, 12396955; hg19; 1000G Project) with the Pro12Ala SNP, among them 18 in complete LD ($r^2 = 1$). Though the GWAS tagSNP rs1801282 encodes a missense mutation (Pro12Ala) implying functionality, the Pro12Ala mutation could not explain the GWAS association so far. Rather, GWAS association results and functional studies appeared contradictory: the minor 12Ala allele, associated with enhanced insulin sensitivity in humans, paradoxically blunts transcriptional activity of the insulin-sensitizing PPARγ2 factor (Deeb et al., 1998). During establishing the PMCA method, I therefore selected the *PPARG* locus for proof-of-principle analysis arguing that a novel regulatory variant located in high LD with Pro12Ala, by allele-dependent *PPARG2* up-regulation, might compensate for the decreased transcriptional activity caused by 12Ala thereby explaining the association signal. Causation due to dysregulation of gene expression has not been considered so far for the *PPARG* locus. Using the PMCA inference procedure, I could pinpoint one complex variant, rs4684847 (C/T), overlapping with the T2D-distinct clustering of the homeobox TFBS matrices. The TFBS matrix overlapping with rs4684847 belongs to the homeobox CART matrix family ($-\log10(p) = 13.00$, the highest score that we found among T2D-distinct TFBS matrix families), and is predicted to bind the homeobox transcription factor PRRX1. I have particularly made an effort to provide high-confidence proof at multiple levels that PMCA enables to inform on GWAS association signals in non-coding regions by extracting (a) the functional *cis*-regulatory variant rs4684847, which modulates *PPARG2* disease gene regulation and (b) its upstream binding transcription factor PRRX1. In a great collaborative effort with Simon Dankel at University of Bergen, Norway, we performed a variety of wet lab studies, including reporter gene assays, electrophoretic mobility shift assay, qRT-PCR, allele-specific primer extension assay (in collaboration with Bernhard Horsthemke, Essen,

Germany), siRNA-mediated knockdown studies in primary human adipose cells. In each of those assays, I specifically addressed the spatio-temporal context of the rs4684847 T2D risk allele (C allele) and could show its cell type-specific and cell stage-dependent repressive effect on endogenous *PPARG2* gene expression in primary human adipose stromal cells during early differentiation. Those studies specifically unveil the role of PRRX1, via increased binding to rs4684847, as a novel key repressor affecting *PPARG2* gene expression, a gene with a crucial impact on adipoycte differentiation and insulin senstitivity (Zhang et al., 2004; Medina-Gomez et al., 2005).

This work could further show that PMCA computational inferences at T2D risk loci lead to the discovery of a novel molecular mechanism underlying the rs4684847-phenotype association: Gene expression profiling in primary adipose stromal cells from rs4684847 homozygous CC risk allele carriers with both PRRX1 and concurrent PRRX1/PPARG knockdown (performed by Simon Dankel, Bergen, Norway), and further wet lab functional assays in rs4684847 genotyped BMI-matched patient cells, revealed that lipid handling in terms of glyceroneogenesis and free fatty acids (FFA) release is perturbed in a rs4684847 genotype-dependent manner. siRNA experiments showed that this effect was mediated by the homeobox factor PRRX1. Both an rs4684847-dependent association and a *PRRX1* mRNA dependent correlation with serum FFA levels was confirmed for the rs4684847 risk allele (C allele) in a cohort of 67 BMI- and body fat matched patients. Glyceroneogenesis is a crucial metabolic pathway in adipocytes, regulating the re-esterification of FFA to triglycerides (TG) in the fasting state (Ballard et al., 1967), thereby controlling systemic FFA homeostasis and insulin sensitivity (Millward et al., 2010). Rosiglitazone, a synthetic ligand of PPARγ2 (Lehmann et al., 1995), pharmacologically increases insulin sensitivity largely via enhancing glyceroneogenesis and subsequently decreasing FFA release (Cadoudal et al., 2007). A recent study has shown that T2D patients harboring the *PPARG* risk haplotype had a reduced

therapeutic response to treatment with rosiglitazone (Kang et al., 2005), which, based on this work, might be explained by PRRX1 binding to the rs4684847 risk allele and thereby decreasing PPARγ2 levels in homozygous CC risk allele carriers. When studying the reported insulin-sensitizing effect of rosiglitazone on glyceroneogenesis (Cadoudal et al., 2007) in patient samples, I observed a marked impaired response to rosiglitazone treatment in terms of glyceroneogenesis-mediated suppression of FFA from homozygous CC risk relative to heterozygous CT non-risk allele patient adipose samples. Importantly, PRRX1 silencing in homozygous risk patient samples was sufficient to fully abolish the reduced responsiveness to rosiglitazone, making PRRX1 an interesting target for pharmacological T2D therapy. Furthermore, we performed a homology directed repair genome editing approach (CRISPR/Cas) to genetically engineer human SGBS adipocytes which harbor the homozygous risk haplotype. By introducing solely the minor rs4684847 non-risk allele (T allele) in the risk haplotype genetic background we could show that that specific substitution of the rs4684847 non-risk allele for the risk allele in SGBS adipocytes is sufficient to induce the altered *PPARG2* expression dependent on PRRX1. For providing the final *in vivo* evidence for T2D causality, I was very lucky getting the possibility to establish valuable collaborations with different research groups, i.e. Gunnar Mellgren (Bergen, Norway), Simon Dankel (Bergen, Norway), Leif Groop (Malmö, Sweden), Matthias Blüher (Leipzig, Germany) and Peter Arner (Stockholm, Sweden). Using three different measures for insulin resistance, i.e. triglyceride/HDL ratio, HOMA-IR, and glucose infusion rate measured by euglycemic hyperinsulinemic clamp studies we could confirm rs4684847 genotype-, and *PRRX1* mRNA-dependent phenotypic changes in patient cohorts. In summary, those data link the *in silico* inferred variant rs4684847, via PRRX1, to perturbed lipid handling in primary human adipose stromal cells that provokes increased plasma free fatty acid levels and systemic insulin resistance.

(2) Enrichment analyses in GWAS data on insulin reistance and impaired insulin secretion

In a collaborative project with Yi-Hsiang Hsu (Institute for Aging Research, Harvard Medical School, Boston, US), we could compute the enrichment of predicted *cis*-regulatory T2D risk SNPs that specifically localize within binding sites belonging to the group of homeobox TFBS for the two T2D features, i.e. insulin resistance and impaired insulin secretion. We interrogated GWAS SNP imputation data for insulin resistance (HOMA-IR) and impaired insulin secretion (HOMA-B) reported from the MAGIC consortium (Dupuis et al., 2010). The empirical p-values for the enrichment of SNPs that localize in close proximity (SNP +/-20bp) to at least one of the homeobox TFBS matrix clusters were computed using 1,000 permutations on the phenotype (95% confidence level). We confirmed an enrichment of the inferred homeobox TFBS-targeting SNPs for both impaired insulin secretion (mean=1.09×10^{-6}; CI: 9.59×10^{-7}–9.51×10^{-3}, p=3.28×10^{-4}; mean: permutation background; CI: 95% confidence interval) and insulin resistance (mean=9.45×10^{-4}, CI: 5.37×10^{-4}–1.34×10^{-2},p = 1.29×10^{-7}).

(3) Co-expression analyses using RNA-seq data assayed in human islets of Langerhans from 51 healthy and 8 T2D deceased donors

In a collaborative project with Leif Groop`s group at Lund University, Malmö, Sweden we were able to perform look-ups in their unpublished RNA-seq data from pancreatic β-cells and could implicate inferred homeobox transcription factors in aberrant insulin secretion. RNAseq data revealed a significant 1.3 - 10.5-fold increase (7.28×10^{-9} < p-value < 4.02×10^{-4}) in mRNA expression of eight inferred homeobox factors *RAX*, *PRRX2*, *BARX1*, *PITX1*, *EMX2*, *NKX6-3*, *BARX2* and *MSX2* - and a 11.7-fold decrease for *PDX1* in islets from T2D subjects compared with non-T2D controls (FDR < 1%). In order to identify potential target genes in human islets for these nine T2D-specific differentially expressed TFs, a co-expression analysis was

performed in the 51 non-T2D controls against all expressed genes in the RNAseq data set (FDR $\leq$ 5%). A pathway analysis (KEGG) of the potential target genes revealed that the "metabolic pathway" was found among the top 5 significantly (hypergeometric test, FDR corrected p-value<0.05) enriched pathways for all but one differentially regulated homeobox TFs (for *MSX2*, *BARX2*, *PDX1*, *PRRX2* and *NKX6-3*), other top 5 enriched pathways includes "insulin signaling" (for *NKX6-3* and *PRRX2*), "MAPK signaling" (for *MSX2*, *PDX1* and *BARX2*), "Notch signaling" (for *PRRX2*), "Calcium signaling" (for *PITX1*) and "Pancreatic secretion" (for *MSX2*), all related to T2D pathophysiology. Seven of the analyzed differentially regulated homeobox TFs had as a potential target the insulin gene, i.e. *RAX*, *PRRX2*, *BARX1*, *PITX1*, *EMX2*, *NKX6-3* and *BARX2*, and could be regarded as novel candidates for regulation of proinsulin production. Finally, we evaluated the effects of the nine identified homeobox TFs on insulin secretion *in vitro* and found that knocking-down each of the nine TFs using siRNA in a pancreatic β-cell line (INS1 cells) significantly inhibited glucose-stimulated insulin secretion from β-cells.

2.2.2 Main Text of the Study

In the following, the main text of the manuscript is attached and the Supplemental Experimental Procedures Section is attached in Appendix.

2.2.2.1 Abstract

Genome-wide association studies have revealed numerous risk loci associated with diverse diseases. However, identification of disease-causing variants within association loci remains a major challenge. Divergence in gene expression due to *cis*-regulatory variants in non-coding

regions is central to disease susceptibility. We show that integrative computational analysis of phylogenetic conservation with a complexity assessment of co-occurring transcription factor binding sites (TFBS) can identify *cis*-regulatory variants and elucidate their mechanistic role in disease. Analysis of established type 2 diabetes risk loci revealed a striking clustering of distinct homeobox TFBS. We identified the PRRX1 homeobox factor as a repressor of *PPARG2* expression in adipose cells, and demonstrate its adverse effect on lipid metabolism and systemic insulin sensitivity, dependent on the rs4684847 risk allele which triggers PRRX1 binding. Thus, cross-species conservation analysis at the level of co-occurring TFBS provides a valuable contribution to the translation of genetic association signals to disease-related molecular mechanisms.

2.2.2.2 Introduction

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

Experimental DHS, RNA, and ChIPseq approaches have been used to prioritize candidate *cis*-regulatory variants (Maurano et al., 2012; The ENCODE Project Consortium, 2012; Ward and Kellis, 2012b). However, such functional approaches require access to appropriate human

tissues and are further hampered by the spatial, temporal, environmental and epigenetic complexity of gene regulation. These limitations emphasize the need for bioinformatics approaches that reliably assess the regulatory role of non-coding variants. So far, phylogenetic conservation has been a common denominator in the search for non-coding regulatory regions (Chinwalla et al., 2002; Pennacchio et al., 2006; The ENCODE Project Consortium, 2007; Visel et al., 2009b; Blow et al., 2010; Lindblad-Toh et al., 2011; The ENCODE Project Consortium, 2012). However, intra- and cross-species differences in gene expression are often driven by changes in transcription factor binding sites (TFBS), and their rapid evolutionary turnover results in lineage-specific regulatory regions that are functionally conserved but have low phylogenetic conservation (Ward and Kellis, 2012a), thus challenging the use of these algorithms. Importantly, gene regulatory regions in eukaryotes tend to be organized in *cis*-regulatory modules (CRMs), comprising complex patterns of co-occurring TFBS for combinatorial binding of transcription factors (TFs) (Arnone and Davidson, 1997; Pennacchio et al., 2006; Visel et al., 2013). CRMs integrate upstream signals to regulate expression of coordinated gene sets, making them a prime target to achieve phenotypic changes as a result of adaptive evolution (Junion et al., 2012). Despite the critical importance of CRMs, no algorithms have so far been developed to harness the potential power of conserved TFBS patterns within CRMs to predict regulatory variants in disease genetics.

We show that cross-species conservation at the level of the CRMs – rather than at the level of the regulatory sequence that comprises them – identifies *cis*-regulatory variants within disease-associated GWAS loci. Exploiting phylogenetic conservation of TFBS co-occurrences, we found homeobox TFBS as a *cis*-regulatory feature of T2D risk loci, for which the specific causal variants have rarely been pinpointed (Stitzel et al., 2010). Detailed analysis at the *PPARG* risk locus revealed the rs4684847 risk allele and, by changing binding of the

homeobox TF PRRX1, its genotype-dependent effect on *PPARG2* expression and insulin sensitivity.

2.2.2.3 Results

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

We developed a method, PMCA, which leverages conserved co-occurring TFBS patterns within CRMs to predict *cis*-regulatory variants, i.e. variants affecting gene expression (Figure 1A, Supplemental Experimental Procedures). To systematically identify *cis*-regulatory variants at GWAS risk loci, we extracted GWAS tagSNPs, and consequently all non-coding (nc) SNPs that are in high LD with these tagSNPs. PMCA individually tests each nc variant by analyzing the flanking region for cross-species conserved TFBS patterns, regardless of global sequence conservation. This requires first the extraction of the region surrounding a nc SNP (±60bp) from the human genome, and consequent identification of orthologous regions in 15 vertebrate species. Within each SNP-specific set of orthologous regions, phylogenetically conserved TFBS, TFBS modules (a cross-species conserved pattern of two or more TFBS occurring in the same order and in a certain distance range), and TFBS in those TFBS modules were identified and then counted. SNP-flanking regions with a significant enrichment of phylogenetically conserved TFBS modules are classified as *complex regions*, as compared to *non-complex regions* (example in Figure 1B) wherein the occurrence of TFBS modules does not exceed expectation by chance. To compute this enrichment we estimate background probabilities using randomizations of orthologous sets (details on scoring cut-offs in Supplemental Experimental Procedures).

We applied PMCA to a set of eight GWAS T2D risk loci (*MTNR1B*, *TCF7L2*, *PPARG*, *CENTD2*, *FTO, GCK, CAMK1D*, *KLF14*) (Dupuis et al., 2010; Voight et al., 2010) covering strong and weaker GWAS signals, and reflecting the different T2D features, i.e. insulin resistance and impaired insulin secretion (Doria et al., 2008). Using non-coding sequence data we defined 200 SNPs in LD with the tagSNPs ($r^2 \geq 0.7$, 1000G) (Figure S1A). PMCA predicted 64 complex and 136 non-complex regions (Figure 1C-G, Table S1). We ranked complex regions based on the count of TFBS in conserved TFBS modules (Table S2), and examined the allele-dependent *cis*-regulatory potential of the 25% highest scoring SNPs using *in vitro* EMSA and reporter assays. As predicted, SNPs in complex regions significantly differed in allele-dependent *cis*-regulatory activity compared to non-complex regions (Figure 1H-I, Table S3). Indeed, the regulatory variants revealed effects ranging from 3.1- to 101-fold change in DNA-protein binding and 1.3- to 3.5-fold change in reporter activity. Moreover, the identified variants operated in a cell type-specific manner (Figure S1B).

To examine if the identified *cis*-regulatory variants in complex regions associate with T2D *in vivo*, we performed look-ups in the MAGIC and DIAGRAM cohorts (Dupuis et al., 2010; Voight et al., 2010). The variants in complex regions revealed a similar or stronger association compared to the initial GWAS signal (Table S4), and a look-up in a recent fine-mapping study (Maller et al., 2012) confirmed that our *cis*-regulatory SNPs belong to the predicted T2D-disease SNP set. GWAS signals are enriched for regulatory variants (Nica et al., 2010). Comparing random SNPs from the 1000G Project to a limited representation of GWAS signals for 19 human diseases (Hindorff LA, accessed June 2013) (Table S5A), we found a 1.12-fold overall enrichment of SNPs in complex regions ($p=1.9x10^{-4}$, binomial distribution) (Table S5B-C), reflecting disease-conferring and low effect *cis*-regulatory variants. Finally, we applied PMCA on reported *cis*-regulatory SNPs associated with diverse disease-related traits, including cancer, myocardial infarction, thyroid hormone resistance,

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

Functional conservation beyond sequence conservation

Given that TFBS turnover is characteristic of CRM evolution (Blow et al., 2010; Ward and Kellis, 2012a), the utility of sequence conservation in deciphering *cis*-regulatory variants may be limited. To assess the power of harnessing TFBS patterns beyond sequence conservation, allowing for sequence variability, we tested complex and non-complex regions for correlations with evolutionary constrained elements detected by the SiPhy-π-method (Lindblad-Toh et al., 2011). For this analysis, we extended our initial PMCA analysis of eight T2D loci to a set of 47 T2D risk loci comprising all GWAS-reported autosomal variants (Hindorff LA, accessed June 2012) including 487 complex and 978 non-complex regions (Table S7). Non-complex regions were depleted of constrained elements in their close proximity (Figure 2A). Conversely, complex regions were enriched for nearby constrained elements, consistent with a 1.37-fold enrichment of GWAS SNPs relative to HapMap SNPs (Lindblad-Toh et al., 2011). Although complex regions overlapped 1.88-fold more with constrained elements than non-complex regions ($p=2.4x10^{-9}$, hypergeometric distribution, right sided), strikingly the majority of complex regions lacked an overlap with constrained elements (Figure 2B, Table S8). This lack of overlap was true for all variants that we

experimentally characterized as *cis*-regulatory (example in Figure 2C). In essence, considering sequence conservation helps to prioritize genomic regions that harbor potential causal variants, yet seems insufficient to pinpoint them. This underscores the importance of exploiting conservation in terms of a complexity assessment of co-occurring TFBS, in the search for *cis*-regulatory variants involved in human diseases.

To further support PMCA predictions at T2D risk loci, we merged our analysis with functional genomics data from The ENCODE Project Consortium 2011 (chromatin state and TF binding). We found complex regions highly enriched for both DHSseq peaks ($p=3.52x10^{-10}$) (Figure 2D) and ChIPseq peaks ($p=4.68x10^{-6}$) (Figures 2E, Table S9). Additionally, crossing our regulatory predictions for T2D risk SNPs with RegulomeDB, a data repository of multiple types of functional ENCODE data (Schaub et al., 2012), confirmed that complex regions are significantly enriched for functional annotations ($p=3x10^{-24}$, hypergeometric distribution, right sided) (Table S10).

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

TFBS clustering relative to transcription start sites indicates biological significance (FitzGerald et al., 2004), and TFBS combination coupled with the TFs recruited to a CRM determines CRM function (Zinzen et al., 2009). Thus, we sought evidence for a discerning T2D functional feature by exploring TFBS characteristics in evolutionary conserved complex regions at T2D risk loci. Given a SNP genomic region we used positional bias analysis, scanning 1,000bp with the SNP at midposition for the occurrence of putative TF binding sequences (883 TFBS matrices grouped in 192 TFBS matrix families) (Table S11). First, for the set of eight T2D risk loci selected for in-depth analysis above, we observed a significant

positional bias for distinct TFBS families (-log10(p)>6) exactly at SNP positions of complex contrary to non-complex regions (Figure 3A). This striking SNP-directed overrepresentation in T2D complex regions was restricted to specific TFBS in the homeobox superfamily, including the matrix families CART (-log10(p)=6.52) and PDX1 (-log10(p)=6.18) (Table S12A). To test whether these findings could be retrieved in a larger set of T2D-associated variants, we extended TFBS clustering analysis to the set of 47 GWAS T2D risk loci (Hindorff LA, accessed June 2012). Indeed, this comprehensive analysis reproduced co-localization of T2D risk SNPs exclusively with homeobox TFBS matrices in complex regions as opposed to non-complex regions (Figure 3B, Table S12B). We again found specific clustering of the CART (-log10(p)=13.00) and PDX1 families (-log10(p)=6.78) together with the homeobox matrix families NKX6 (-log10(p)=8.50), HOMF (-log10(p)=8.94), HBOX (-log10(p)=8.54) and BCDF (-log10(p)=7.24). No other TFBS matrices showed a significant peak in the bias profile at SNP positions. Importantly, when applying PMCA on risk loci of T2D non-related traits, asthma and Crohn's disease (Moffatt et al., 2010; Schaub et al., 2012) (Figure S3B-C, Table S13), we observed disease-distinctive TFBS at SNP positions (Table S12C-D). Both complex and non-complex regions lacked a clustering of homeobox TFBS at asthma risk SNPs (Figure 3C). The specific clustering of the Early Growth Response Factor (EGRF) matrix family for asthma risk SNPs in complex regions (-log10(p)=8.50, Figure 3D) was in strong contrast to T2D (-log10(p)=3.97, Figure 2E) and Crohn`s (-log10(p)=2.07, Figure S3D). Of note, the EGRF-binding factor EGR1 regulates asthma-related IL13-induced inflammation (Cho et al., 2006).

Homeobox TFs are known to be involved in tissue developmental processes including β-cell development (Jorgensen et al., 2007). However, except for the MODY gene *PDX1* (Fajans et al., 2001) and the common T2D-associated loci *HHEX1* and *ALX4* (Sladek et al., 2007), the PMCA-inferred homeobox factors have not been implicated in T2D pathogenesis.

T2D is marked by insulin resistance and impaired insulin secretion (Doria et al., 2008). To evaluate a functional role of the homeobox TFBS matrix families in T2D pathogenesis, we extracted data for insulin resistance (HOMA-IR) and impaired insulin secretion (HOMA-B) (Dupuis et al., 2010), to compute the enrichment of predicted *cis*-regulatory T2D risk SNPs that localize in close proximity to those homeobox TFBS (±20bp, permutations on the phenotypes, n=1,000, 95% confidence interval, Supplemental Experimental Procedures). We verified a significant enrichment of SNPs that localize ±20bp at inferred homeobox TFBS for both insulin resistance (mean=$1.09x10^{-6}$; CI:$9.59x10^{-7}$–$9.51x10^{-3}$, p=$3.28x10^{-4}$; mean permutation background; CI:95% confidence interval) and impaired insulin secretion (mean=$9.45x10^{-4}$; CI:$5.37x10^{-4}$–$1.34x10^{-2}$, p=$1.29x10^{-7}$). Furthermore, we elucidated a potential effect of their binding TFs on impaired insulin secretion. Assessing mRNA levels in human islets from 51 healthy and 8 T2D deceased donors by RNAseq (L. Groop, unpublished data), we found a marked expression difference for *RAX, PRRX2, BARX1, PITX1, EMX2, NKX6-3, BARX2, MSX2* and *PDX1* in islets from T2D patients compared to healthy controls ($7.28x10^{-9}$<p<$4.02x10^{-4}$, FDR<1%) (Table S14). By genome-wide co-expression analysis we found significantly co-regulated gene sets (p<$5.02x10^{-3}$; FDR<5%, n=51 healthy donors) (Table S15). Except for the gene set co-regulated with *PITX1*, we found *metabolic pathways* among the top 5 significantly enriched pathways (hypergeometric test, FDR corrected p<0.05) (Figure S3E). Other top 5 enriched pathways included *insulin signaling, MAPK signaling, notch signaling, calcium signaling* and *pancreatic secretion*. Knock-down of each candidate homeobox TFs in pancreatic INS-1 β-cells significantly perturbed glucose-stimulated insulin secretion (Figure S3F). Moreover, except for *PDX1* and *MSX2* (corrected FDR, p=0.96 and p=0.89), all PMCA-inferred homeobox TFs were significantly co-expressed with the insulin gene in islets of 26 hyperglycemic individuals (HbA1C>6) (Table S16). Although the result for *PDX1* was borderline non-significant it is a well-known regulator of insulin expression

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

The T2D identified variant rs4684847 regulates *PPARG2* gene expression

To establish the informative value of TFBS pattern analysis for pinpointing the *cis*-regulatory variant and binding TF underlying GWAS association signals, we chose the *PPARG* locus for detailed study. PPARγ is crucial in adipogenesis, lipid metabolism and systemic insulin sensitivity (Rosen et al., 1999; Medina-Gomez et al., 2005), and exists as two isoforms: PPARγ1 (*PPARG1*, *PPARG3* mRNA) and PPARγ2 (*PPARG2* mRNA) (Fajas et al., 1998), the latter mainly expressed in adipocytes (Tontonoz et al., 1994). There is a robust association of *PPARG* with T2D (Deeb et al., 1998; Heikkinen et al., 2009; Dupuis et al., 2010; Voight et al., 2010). The T2D GWAS association comes from an LD region mainly tagged by the coding missense mutation Pro12Ala (Figure 4A, upper panel). However, the minor 12Ala allele, associated with enhanced insulin sensitivity in humans, paradoxically blunts the transcriptional activity of the insulin-sensitizing PPARγ2 TF (Deeb et al., 1998). Hypothesizing that the elusive *PPARG* T2D signal instead arises from a regulatory variant that affects *PPARG2* expression, we first confirmed - before analyzing variants at the *PPARG* locus with PMCA - a risk allele-dependent 3.8-fold decrease of *PPARG2* mRNA in human adipose stromal cells (hASCs) ($p=1.0x10^{-3}$) (Figure 4B). This effect was specific for *PPARG2*, as there was no effect on *PPARG1* expression (Figure 4C).

First, to narrow-down the variants that could explain the decrease in *PPARG2* expression and thereby the underlying T2D association, we applied PMCA to each of the 23 correlated non-coding variants at the *PPARG* locus ($r^2 \geq 0.7$, 1000G Project) (Figure 4A). Seventeen variants were ruled out being located in non-complex regions (Figure S4A, Table S17).

Among the six variants in complex regions, five had either activating or repressing *cis*-regulatory activity (Figure 4D), which may reflect gene regulatory dependency on the tissue/cell-type and the spatial, temporal, environmental and epigenetic context. In fact, while the qPCR data in undifferentiated hASCs showed a suppressive effect specific for the *PPARG2* mRNA isoform, adipose tissue eQTL data showed an up-regulation of total *PPARG* mRNA in risk allele carriers (p=0.01) (Figure S4B).

Second, to pinpoint the functional variants that may explain the GWAS-reported T2D association, we scrutinized the complex regions for those TFBS showing a clustering at T2D risk SNP positions (drawn from the overall TFBS clustering analysis in complex regions, Figure 3), pursuing the variants overlapping a TFBS matrix in the disease-distinctive cluster. As shown above, our comprehensive cross-species TFBS pattern analysis of 47 T2D risk loci unveiled a clustering of specific homeobox TFBS families as a characteristic feature of T2D risk SNPs (Figure 3B). Among the six non-coding variants at the *PPARG*, only one variant, rs4684847 (C/T), overlaps with the T2D-distinct clustering of the homeobox TFBS matrix. The TFBS matrix overlapping with rs4684847 belongs to the CART matrix family (-log10(p)=13.00, the highest score among TFBS matrix families), and is predicted to bind the homeobox TF PRRX1. The other five non-coding variants showed no homeobox TFBS match (Figure 4A, lower panel).

Third – as an independent approach to confirm rs4684847 mediating the *PPARG2* suppression – we examined the cellular context of genotype-dependent *PPARG2* suppression, and epigenomic profiling data that allow for temporal chromatin state-dependent regulatory functional annotations. By allele-specific primer extension analysis in heterozygous undifferentiated hASCs genotyped for rs4684847, where each allele serves as an internal control for the other, we first confirmed a striking allelic imbalance with 5.4-fold lower *PPARG2* mRNA expression from the C risk allele ($p=6.0x10^{-4}$) (Figure 4E). Given the role of

PPARG2 in adipogenesis, we then tested whether the rs4684847 C risk allele might affect *PPARG2* mRNA expression during adipogenesis. The allele-specific primer extension analyses in hASCs from heterozygous risk allele carriers revealed that the risk allele-dependent suppression of *PPARG2* mRNA diminished with progression of adipogenesis ($p<0.001$) (Figure S4C). These data suggest a highly temporal context-specific effect of the risk allele on *PPARG2* suppression in the undifferentiated state. Given the availability of cell-stage dependent open chromatin data in hASCs reported by Mikkelsen et al, 2010, we sought supportive evidence for rs4684847 as the variant underlying the cell-stage dependent allelic *PPARG2* expression. We integrated all six variants in complex regions at the *PPARG* locus with genome-wide temporal regulatory annotations estimated by H3K27ac data. Among those six, only the flanking region rs4684847 (C/T) showed consistent cell stage-dependent H3K27ac density distributions (Figure S4D). Thus, the rs4684847-specific match with the T2D homeobox TFBS clustering, informed by conserved TFBS pattern analysis, could be confirmed by cell-stage dependent regulatory regions estimated by chromatin state data.

Finally, we performed a host of *in vitro* and *in vivo* analyses to prove that the rs4684847 risk allele (C allele) mediates the suppression of *PPARG2* mRNA expression via the transcriptional regulator PRRX1. By affinity chromatography and LC-MS/MS we could demonstrate a 2.3-fold increased binding of PRRX1 to the rs4684847 risk relative to non-risk allele (Supplemental Experimental Procedures). Moreover, by EMSA we found rs4684847 risk allele-specific DNA-protein binding (Figure 4F), and competition EMSA and supershift experiments confirmed that PRRX1 was responsible for this allele-specific DNA-protein binding (Figure 4G). Furthermore, consistent with the GWAS signal for insulin resistance rather than insulin secretion (Voight et al., 2010), in luciferase reporter assays we observed rs4684847 cell type-specific effects in 3T3-L1 adipose cells, C2C12 myocytes and Huh7 hepatocytes, whereas pancreatic INS-1 β-cells and 293T cells lacked allelic activity (Figure

S4E). Luciferase activity in 3T3-L1 preadipocytes was 5.2-fold lower for the C risk allele ($p=1.0x10^{-4}$, Figure 4H). This repressive effect was independent of 5´-vs. 3´-orientation to the reporter gene (p=0.03) and forward-reverse orientation (p=0.03) (Figure S4F), suggesting enhancer function for the non-risk allelic complex region. Importantly, perturbing the PRRX1 consensus sequence without affecting the SNP position itself fully abrogated the C risk allelic repression of reporter gene activity (Figure 4H), whereas overexpressing PRRX1 enhanced it ($p=2.0x10^{-4}$, Figure 4I).

We then sought final proof that the rs4684847 risk allele – independent of correlated sequence variants – causes the suppression of endogenous *PPARG2* expression. We used an adopted CRISPR/Cas homology-directed repair genome editing approach (Wang et al., 2013) to introduce the rs4684847 non-risk allele in human SGBS preadipocytes, replacing the endogenous risk allele. Notably, the rs4684847 non-risk allele was sufficient to increase *PPARG2* transcript levels 5.4-fold (p=0.005) (Figure 4J, left) (*PPARG1* unaffected) (Figure S4G). In parallel experiments we performed PRRX1 knockdown and confirmed that 1) risk allele-driven suppression of *PPARG2* expression was reversed by PRRX1 silencing (p=0.005) and 2) PRRX1 silencing did not affect *PPARG2* expression in non-risk allele cells (Figure 4J, right).

rs4684847 via PRRX1 binding affects FFA homeostasis and insulin sensitivity

The SNP rs1801282 (Pro12Ala) in *PPARG* associates with BMI, fasting insulin, and insulin sensitivity (Deeb et al., 1998; Voight et al., 2010). rs4684847 is located 6.5kb upstream of the *PPARG2* specific promoter and is in complete LD ($r^2=1.0$) with rs1801282. Via PMCA, we found that PRRX1 binds at the rs4684847 C risk allele and thus inhibits *PPARG2* expression. On the other hand, the T allele of rs4684847 (minor allele frequency 6.5% in Caucasians)

reduces the binding ability of PRRX1 and thus maintains a higher level of *PPARG2* expression. Further *in vivo* evidence was obtained in primary human adipose stromal cells (hASC) isolated from BMI-matched subjects, showing rs4684847-dependent *PPARG2* mRNA expression ($p=1.4x10^{-20}$, n=32). PPARγ2 is crucial for maintaining insulin sensitivity: adipose-specific *Pparg2* knockout mice develop insulin resistance independently of affecting body weight (Medina-Gomez et al., 2005), and PPARγ is target of the thiozolidinedione (TZD) class of insulin-sensitizing drugs such as Rosiglitazone (Rosi) (Lehmann et al., 1995). Indeed, we observed rs4684847-dependent association with lower T2D risk (Voight et al., 2010) (OR=0.89, 95% CI=0.86-0.92, $p=3.75x10^{-11}$, n=80,648). Further, in hASC we found rs4684847-dependent increase in adipocyte insulin sensitivity ($p=1.5x10^{-7}$, ratio insulin-stimulated/basal 2-deoxyglucose uptake, Pearson's correlation, n=32). We confirmed a significant interaction between the rs4684847 risk allele and adipose *PRRX1* mRNA levels to HOMA-IR, independent of BMI (p=0.044, n=38, interaction model, Supplemental Experimental Procedures). In addition, we observed rs4684847-dependent correlations of *PRRX1* mRNA levels with BMI, TG/HDL ratio, and BMI-adjusted HOMA-IR, and with glucose infusion rate (GIR) measured by euglycemic hyperinsulinemic clamp in a cohort of 67 BMI- and body fat-matched obese patients (Table 1, Figure S4H).

To further examine PRRX1 as mediator of the repressive rs4684847 risk allele (C allele) effect on *PPARG2* expression, we performed knockdown of PRRX1 in hASCs and found that PRRX1 silencing was sufficient to revert the risk allelic suppression ($p=3.3x10^{-15}$) (Figure 5A, Table 2). Then, to inform on the cellular processes by which PRRX1 may contribute to T2D, we studied the impact of PRRX1 on PPARγ-regulated genes in hASCs from homozygous rs4684847 CC risk allele carriers by microarray analysis (n=9). We found 2,258 transcripts regulated by PRRX1 knockdown ($q<0.2$), 336 of which were reversely regulated by concomitant *PPARG* knockdown (Figure 5B). Gene Set Enrichment Analysis (GSEA)

highlighted an enrichment of those anti-regulated genes among the most differentially expressed genes after PRRX1 knockdown (FDR=0, Figure 5C), revealing that PPARγ2 mediated the primary PRRX1 effect on global gene expression. Ingenuity Pathway Analysis (IPA) showed the strongest enrichment for lipid metabolism ($p=2.81x10^{-14}$) followed by adipose tissue function, glucose homeostasis, nutritional disease and insulin resistance (Figure 5D). Accordingly, an inverse relationship between PRRX1 and adipocyte triglyceride (TG) accumulation was observed in PRRX1-overexpressing SGBS adipocytes (Figure 5E).

By qPCR we confirmed rs4684847 allele-dependent dysregulation of genes in the identified biological pathways. Notably, the gene with the strongest risk allele-dependent decrease in mRNA levels was *PEPCKC* (Table 2). The top scoring IPA interaction network reinforced a central role for *PEPCKC* (Figure 5F). PEPCK-C is the enzyme controlling the first committed step of glyceroneogenesis, a crucial metabolic process in adipocytes regulating the re-esterification of free fatty acids (FFA) to TG (Ballard et al., 1967). Glyceroneogenesis limits FFA release from adipocytes in the fasting state thereby controlling systemic FFA homeostasis and insulin sensitivity (Millward et al., 2010). In the 67 BMI- and body fat-matched obese subjects we confirmed rs4684847 risk allele association with increased serum FFAs levels (p=0.049) and risk allele-dependent association of *PRRX1* mRNA with FFA levels (p=0.015, Table 1). To prove that rs4684847, by determining PRRX1 binding, affects glyceroneogenesis and subsequent FFA release, we monitored pyruvate incorporation in TG (Ballard et al., 1967). We confirmed a PRRX1-dependent suppression of glyceroneogenesis in CC risk allele carriers, marked by a robust correlation with *PRRX1* mRNA levels (Figure 5G) and a risk allele-dependent increase in FFA release (Figure 5H). In a parallel experiment, we also found that PRRX1 silencing was sufficient to restore cellular insulin sensitivity in risk allele carriers (Figure 5I). Importantly, the PPARγ ligand Rosi pharmacologically promotes insulin sensitivity largely via control of FFA homeostasis

through glyceroneogenesis (Cadoudal et al., 2007), and (Kang et al., 2005) reported impaired Rosi response in risk haplotype carriers. In our analysis of glyceroneogenesis in hASCs we observed an impaired response to Rosi-mediated suppression of FFA release dependent on the risk allele (Figure 5J). Strikingly, PRRX1 silencing in CC risk-allele patient samples was sufficient to abolish the reduced Rosi responsiveness, making PRRX1 a potential target for pharmacological T2D intervention.

In summary, by PMCA we demonstrate a clustering of specific homeobox TFBS at T2D risk SNPs. We specifically unveil a novel role of homeobox TF PRRX1 as a repressor of *PPARG2* via its enhanced binding at the rs4684847 C risk allele, thereby provoking dysregulation of FFA turnover and glucose homeostasis (Figure 5K).

2.2.2.4 Discussion

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

Using T2D as a showcase we demonstrate the utility of PMCA for the generic prediction of distinct homeobox TFBS at T2D risk SNPs, which is important for understanding disease regulatory circuits when we consider that interactions in a regulatory network involve numerous genes and a rather small set of TFs (Califano et al., 2012). Pursuing the results

emerging from our comprehensive T2D analysis, we show that identification of the *cis*-regulatory variant rs4684847 at the *PPARG* locus enabled linking the molecular upstream factor PRRX1 to aberrant downstream mechanisms of impaired lipid handling and insulin sensitivity, explaining the GWAS association with T2D. Notably, PRRX1 was recently implicated in adipogenesis (Du et al., 2013), yet the regulated genes remain elusive.

Here, we restricted the analysis to SNPs in LD with GWAS SNPs. However, the approach could be applied to any other kind of variability, such as somatic mutations in cancer, without loss of generality. Certain issues will require consideration, e.g. analyzing genomes of closely related species to refine scoring criteria, and extending our analysis to whole genome sequencing studies, including rare variants data, should further inform on the genetic underpinnings of phenotypic diversity in humans. Our *in silico* scoring results predict varying numbers of regulatory SNPs per LD block. Studies have now found evidence for allelic heterogeneity (Maller et al., 2012; Schaub et al., 2012), yet the number of causal variants within a disease locus is elusive. We propose an integrative framework where computational TFBS modularity analysis may be synergistically combined with functional genomics and population genetics data.

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

2.2.2.5 Experimental Procedure

The detailed experimental procedures are included in the Supplemental Experimental Procedures, which is attached as Appendix.

Definition of LD blocks

Tag SNPs were derived from reported disease risk loci identified by GWAS (references listed in Tables S1, S7 and S8). For each tag SNP, LD blocks were defined (CEU, The 1000 Genomes Project Consortium, 2010, $r^2 \geq 0.7$, NCBI GRCh37/hg19) using the SNAP viewer tool (Johnson et al., 2008), Broad Institute.

Phylogenetic Module Complexity Analysis

Our bioinformatics methodology analyzes the presence of complex patterns of evolutionarily conserved TFBSs in a *cis*-regulatory module (CRM), within genomic regions surrounding a SNP (SNP region) to predict its *cis*-regulatory functionality. For each SNP the 120 bp sequence with the SNP at central position (SNP region) was extracted from the human genome (NCBI GRCh37/hg19). Orthologous sequences were searched in 16 vertebrate species (ortholog set). The ortholog set was at first analyzed for the occurrence of transcription factor binding sites (TFBSs) that could be found in a defined input set of orthologous regions within the ortholog set on any strand (common TFBSs). We retrieved common TFBSs to identify TFBS modules that consist of at least two TFBSs co-occurring in the same orientation and distance range across a defined input set of orthologous regions. Generally, the complexity of TFBS modularity within the ortholog set is assessed based on the precise determination of the occurrence/number (#) of three measures, (1) *#common TFBSs*, (2) *#common TFBS modules* and (3) *#TFBSs in modules*. Simulation on random sets was performed to separate complex SNP regions from non-complex SNP regions by estimating the probability for random occurrence of all three measures. Our bioinformatics methodology is described in detail in Supplemental Experimental Procedures and Figure S1.

Positional Bias Analysis

120 bp genomic regions with SNP at central position were scanned by MatInspector (Genomatix) for presence of TFBS matrix family matches at SNP position, and positional bias of TFBS matrix families was calculated as described for *de novo* detection of motifs by (Hughes et al., 2000) using overlapping 50 bp sliding windows in steps of 10 bp. Positional bias was calculated as binominal P value for each matrix family and each window. For additional details see the Extended Experimental Procedures.

Correlation of SNP regions with evolutionary constraint, DNase-seq and ChIP-seq regions

Genomic regions surrounding a candidate SNP were classified as complex and non-complex and were correlated to evolutionary constrained regions (Lindblad-Toh et al., 2011) or DNase-seq and ChIP-seq peaks (The ENCODE Project Consortium, 2012). From midpoint (+/-500bp) of constrained regions as anchor, the overlapping positions (correlation) with complex and non-complex SNP regions (SNP +/- 60bp) were counted, and resulting correlations were plotted *versus* position relative to the anchor. From complex and non-complex genomic regions surrounding a SNP (+/-500bp) as anchor, the overlapping positions of DNase-seq and ChIP-seq regions (correlation) with complex and non-complex SNP regions (SNP +/- 60bp) were counted and plotted versus position relative to the anchor. For constrained, DNase-seq and ChIP-seq regions information and for additional details see the Extended Experimental Procedures.

Human adipose tissue and primary human adipose stromal cells

Human abdominal adipose tissue was obtained with informed consent from healthy male and female subjects undergoing lipoaspiration or surgical excision of subcutaneous tissue. Informed consent was obtained from all patients before the surgical procedure. Procedures for the different studies were approved by the ethical committee of the Faculty of Medicine of the

Technical University of Munich (Germany), the Regional Committee for Medical Research Ethics REK (Bergen, Norway), the Ethics committee of the University of Leipzig (Leipzig, Germany) or the ethics committees at Lund University (Sweden). Primary hASCs and mature human adipocytes were isolated as described and differentiated using a protocol modified from (Veum et al., 2011). Primary cells were genotyped using MassARRAY (Sequenom).

Expression Analysis by qRT-PCR, allele-specific primer extension and eQTL

Total cell RNA was prepared using TRIzol (Invitrogen), primary adipocytes with RNeasy Lipid Tissue Mini Kit (Qiagen). qRT-PCRs were performed using cDNA Reverse Transcription kit (Applied Biosystems) or SuperScript® VILOTM cDNA Synthesis Kit (Invitrogen), SYBR-Green or Universal ProbeLibrary (UPL) (Roche), and Mastercycler Realplex (Eppendorf) or LightCycler480 (Roche, Germany). For allele-specific primer extension SNP surrounding regions were amplified from cDNA, purified by agarose gel, primer extension performed using SNaPshot Kit (ABI Prism), genomic DNA amplified using GoTaq DNA Polymerase Kit (Promega), products analyzed by gel capillary electrophoresis on ABI 3100 DNA Analyzer and electropherograms analyzed using Gene Mapper software (ABI). For eQTL analysis total PPARγ expression levels from GeneChip® Human Gene 1.0 ST whole transcript based array (Affymetrix) and rs7638903 variant (r^2 = 1.0 to rs4684847, and to Pro12Ala) genotyped by Omni express (Illumina) were compared. For additional details including primer information see the Supplemental Experimental Procedures.

Cell Culture and Reporter Assays

Culturing of Huh7 hepatocytes, INS-1 β-cells, 293T cells and differentiation of C2C12 myocytes, 3T3-L1 adipocytes and SGBS adipocytes as described (Fischer-Posovszky et al., 2008; Laumen et al., 2009).

Genomic sequences surrounding the respective SNPs were synthesized (MWG, Invitrogen) and cloned into pGL4.22 (Promega) containing TK-promoter. Huh7, INS1, 3T3-L1 and C2C12 cells were transfected using Lipofectamine 2000 reagent (Invitrogen), renilla-luciferase pRLUbi was cotransfected for normalization (Laumen et al., 2009). Luciferase activities were measured using LuminoscanAscent (Thermo) or Sirius luminometer (Berthold).

Gene knock-down by siRNA

Primary hASCs and SGBS cells were treated with PRRX1 ON-TARGETplus human siRNA SMARTpool or non-targeting control (Dharmacon) using HiPerFect (Qiagen). For additional details see the Supplemental Experimental Procedures.

Electrophoretic mobility shift assay (EMSA)

EMSA with 42bp SNP-adjacent regions was performed with annealed Cy5-labelled oligonucleotide probes (MWG) and native protein extracts from the respective cell lines, modified protocol from (Laumen et al., 2009). For supershift experiments cell extracts from 293T cells transfected with pCMV-PRRX1-flag vector were pre-incubated with αPRRX1 (M. Kern) or control IgG (Santa Cruz Biotechnology), competition experiments with excess of unlabeled probe. Quantification of DNA-binding complexes was performed with ImageJ Software (http://rsbweb.nih.gov/ij/).

DNA-Protein affinity chromatography, LC-MS/MS

To identify rs4684847-specific binding proteins DNA-protein affinity chromatography was performed using Streptavidin magnetic beads (Invitrogen) and allele-specific biotinylated DNA-probes (MWG), followed by tryptic digest, LC-MS/MS using Ultimate3000 nano HPLC (Dionex) online coupled to a LTQ OrbitrapXL mass spectrometer (Thermo Fisher

Scientific) and data analysis based on the Ensembl mouse protein database (Version NCBI m37) using Progenesis LC-MS software v.2.5 as described (Hauck et al., 2010).

Statistical Analysis

All statistical analyses were done using the Graph Pad Prism software v5.02, Pearl or the Statistical Software R v2.14.2.

2.2.2.6 Tables and Figure Legends

Table 1

Correlation of adipose tissue *PRRX1* mRNA expression with T2D traits in rs4684847 risk allele carriers.

		***PRRX1* mRNA**		***PRRX1* mRNA**		***PRRX1* mRNA**	
rs4684847 genotypes		**All**		**CC**		**CT and TT**	
		β	**p**	**β**	**p**	**B**	**p**
a)		n=38		n=20		n=18	
log(BMI)	-	1.32	0.05	1.23	0.19	1.43	0.23
	Age	1.45	0.03	1.23	0.19	1.96	0.09
log(TG/HDL)	-	6.92	**7.54×10^{-4}**	6.40	**0.02**	6.35	0.07
	Age	6.97	**7.36×10^{-4}**	6.14	**0.02**	6.81	0.07
	age/ BMI	4.86	**8.3×10^{-3}**	5.00	**0.07**	2.64	0.33
log(HOMAIR)	-	2.77	**3.52×10^{-3}**	3.13	**8.3×10^{-3}**	1.80	0.29
	Age	2.77	**3.77×10^{-3}**	3.12	**8.6×10^{-3}**	1.70	0.34
	age/ BMI	1.41	**0.028**	2.1	**4.6×10^{-3}**	-0.55	0.63
b)		n=67		n=54		n=13	
log(GIR)	age/ BMI	-0.51	**1.83×10^{-7}**	-0.78	**3.30×10^{-8}**	-0.38	0.28
log(FFA)	age/ BMI	0.25	0.014	0.27	0.015	-0.009	0.99

Gene expression and phenotypes were measured in a) adipose tissue from a lean/obese patient cohort (mean±SD 24.2±9.1 kg/m^2), and b) adipose tissue samples from BMI-matched obese patients (mean±SD 43.2±3.1 kg/m^2) characterized by hyperinsulinemic euglycaemic clamp. rs4684847 risk-allele and non-risk allele genotypes were determined by Sequenom-assay. BMI, body mass index; HOMA-IR, homeostasis model assessment of insulin resistance; TG, triglyceride; HDL, high density lipoprotein; GIR, glucose infusion rate of hyperinsulinemic euglycemic clamp; FFA, free fatty acids. p-values and β-estimates from linear regression analysis of *PRRX1* mRNA expression levels with phenotype measures are shown.

Table 2

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

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

	±0.48	±0.30			±0.37	±0.44						
GLUT4	1.57 ±0.35	0.99 ±0.24	1.58	**6.15×10^{-5}**	1.62 ±	1.50 ±0.31	1.09	0.26	1.03	0.67	1.50	**1.08×10^{-4}**
THRSP	0.99 ±0.28	1.61 ±0.39	0.61	**8.18×10^{-5}**	1.53 ±0.33	1.60 ±0.32	0.95	0.57	1.55	**1.38×10^{-4}**	0.99	0.93

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

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

(A) Workflow of the PMCA methodology: (1) The flanking region of a non-coding SNP is extracted from the human reference genome; (2) orthologous regions are searched in the genomes of 15 vertebrate species; (3) TFBS are identified in each orthologous sequence; (4) TFBS modules are identified in the set of orthologous sequences (TFBS modules defined as all, two or more TFBS occurring in the same order and in certain distance range in all or a subset of the orthologous sequences); (5) phylogenetically conserved TFBS Ω_{TFBS}, TFBS modules $\Omega_{modules}$, and occurrences of TFBS in TFBS modules $\Omega_{TFBS\_in\_modules}$ are counted; (6) repeated counting for different numbers of input sequences weighs the degree of cross-species conservation and the number of TFBS in modules. Computation of conserved TFBS with more restricted parameters Ω_{restr_TFBS} accounts for genomic regions with low numbers of orthologs; (7) steps 3-6 are repeated using randomized input sequences (randomization of sequences is done using local shuffling in order to conserve local nucleotide frequency distributions) to estimate; (8) the probability p-est of observing a given Ω_{TFBS}, Ω_{restr_TFBS}, $\Omega_{modules}$, and $\Omega_{TFBS\_in\_modules}$ and to calculate the overall scoring criterion; (9) input sequences are classified as complex and non-complex regions; (10) complex regions harboring a trait-related TFBS at SNP position are selected for functional evaluation (trait-related TFBS are drawn from overall TFBS clustering analysis as described in text related to Figure 3). Supplemental Experimental Procedures.

(B) Representative complex region (rs4684847) and non-complex region (rs13064760). Conserved TFBS and conserved TFBS in modules occurring in more than 2 vertebrate species are shown to illustrate TFBS modularity across species.

(C-G) Classification of SNP regions for a set of eight T2D risk loci (Table S1, Figure S1).

(C-E) Box-Whisker plots (IQR 50%) show the counts of conserved TFBS Ω_{TFBS} (C), conserved TFBS modules $\Omega_{modules}$ (D) and occurrences of TFBS in TFBS modules $\Omega_{TFBS\_in\_modules}$ (E) for complex regions (red lines) and non-complex regions (black lines). Data points covered by the IQR and the whiskers values were added as rug at the sides of the plot. Note that values vary over a large range with higher median for complex regions for all criteria (at 47 T2D loci we find a median of 354.5/470.46 and 310/382.35 for $\Omega_{TFBS\_in\_modules}$ in complex/non-complex regions).

(F-G) Scoring of SNP regions is illustrated by histograms showing the probability p-est of observing Ω_{TFBS} across species (F) and showing the overall scoring criterion S_{all} (G). *Blue*

curve: empirical density function of the histogram data. *Red dashed line:* cut-off scores separating complex from non-complex regions (-log10 p-est$_{TFBS}$=1.12, S_{all}=6.5); SNP regions with a value to the left of the red line were defined as non-complex.

(H-I) *Cis*-regulatory activity of SNP regions. Non-complex regions include regions matched for TFBS density of complex regions (TFBS median=88). The allele-dependent change in DNA-binding activity from EMSAs (n=4) (H) and luciferase reporter activity (n=10) (I) is shown for each SNP. Mean±SD, p from linear mixed-effects model. See also Tables S2-3.

Figure 2. Correlations of *cis*-regulatory predictions at 47 T2D risk loci with evolutionary constrained elements and functionally annotated genomic regions (Table S7-9).

(A) Correlation of PMCA results with evolutionary constrained regions. The occurrences of 487 complex and 978 non-complex T2D-associated regions within constrained regions from SiPhy-π algorithm (Lindblad-Toh et al., 2011). Localization of SNPs relative to transcription start site in Figure S2A-B.

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

(C) Complex regions at the *PPARG* locus (Figure 4E) lack an overlap with constrained regions. Zoom-in: the rs4684847 *cis*-regulatory region does not map to a constrained region (393bp upstream of nearest constrained element). A representative TFBS module ($\Omega_{TFBS\_in\_module}$ = 3) is shown and its TFBS module conservation for a given quorum of five species is visualized by a sequence logo.

(D-E) Correlation of complex (red line) and non-complex (black line) T2D-associated SNP regions to DHSseq (D) and ChIPseq (E) peaks. From the midpoint of 487 complex and 978 non-complex regions, 1,000bp in both directions were scanned for DHSseq and ChIPseq peaks (Supplemental Experimental Procedure). For each position, the sum of occurrences was plotted. T2D complex regions were significantly enriched for overlaps with DHSseq and ChIPseq regions, displayed as a central peak (correlations with Crohn's-associated regions in Figure S2C-D).

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

Distribution of TFBS matrices relative to SNP positions (SNP±500bp) at T2D compared to asthma risk loci, calculated using positional bias analysis. 1,000bp genomic regions with SNPs at midposition were scanned for the occurrence of TFBS matches for 192 TFBS matrix families (sliding 50bp windows, p from binomial distribution model, Supplemental Experimental Procedure).

(A-B) TFBS family distribution in a set of eight and an extended set of 47 T2D risk loci. Complex regions reveal clustering of distinct homeobox TFBS matrix families at T2D risk SNP positions (±20bp, grey dashed lines). All TFBS families displayed equal distributions within T2D non-complex regions, represented (a subset of representative TFBS families is shown).

(C) TFBS family distribution in a set of eight asthma risk loci. Asthma complex and non-complex regions lack a positional bias at SNP positions for the homeobox TFBS matrix families clustering in complex regions at T2D risk SNPs (details on Crohn Figure S3).

(D-E) TFBS family distribution in asthma risk loci revealed a specific EGRF matrix family clustering in complex regions at asthma risk SNPs (D). T2D complex regions lack a clustering of EGRF matrices at SNP positions (E).

Figure 4. The non-coding SNP rs4684847 (C/T), by binding the homeobox factor PRRX1, represses *PPARG2* expression at the *PPARG* diabetes risk locus.

(A) Top panel: A LD regional plot of the *PPARG* locus. *Diamonds* tagSNP Pro12Ala and pairwise correlation of SNPs in LD (MAF≥1%) against genomic position; *Blue PPARG* gene and exons; Middle/lower panel: Classification of SNPs in complex regions (red lines) and non-complex regions (grey lines) (PMCA steps 1-9, Figure 1A); Scanning of *PPARG* complex regions for T2D-distinct homeobox TFBS matrix families (CART, HOMF, HBOX, NKX6, BCDF, PDX1; Figure 3B) pinpoints rs4684847 (C/T), based on its overlap with the CART binding matrix for PRRX1 (step 10, Figure 1A). *Zoom-in* human *PPARG* gene; *arrows* TSS of *PPARG1-3* mRNA isoforms; *boxes* coding exons (filled) and untranslated exons (open); *lines* introns; *Second zoom-in* CRM at rs4684847: the PRRX1 matrix co-occurs with

diverse TFBS matrices in consistent orientation and distance range across species, exemplarily illustrated by one conserved TFBS module ($\Omega_{TFBS\_in\_modules}$ = 3; TFBS matrices: PRRX1, TEF, LHXF).

(B-C) Genotype-dependent mRNA expression in undifferentiated hASCs genotyped for Pro12Ala and rs4684847 (r^2=1.0). qPCR of *PPARG1* and *PPARG2* mRNA isoforms (standardized to *HPRT*) homozygous CC risk (n=9) and CT non-risk allele carriers (n=5) normalized to mean for CC. Mean±SD, t-test.

(D) Validation of *cis*-regulatory predictions for complex regions at the *PPARG* locus. Quantified change in reporter activity in 3T3-L1 adipocytes is shown for each SNP, using luciferase constructs harboring the risk or non-risk alleles, representing an activating or repressing effect of the risk-allele on transcriptional activity. Mean±SD, n=3-14, paired t-test.

(E) Allele-specific primer extension analysis in hASCs of heterozygous rs4684847 carriers (n=6) normalized to mean risk allele levels (D). Mean±SD, Mann Whitney U test.

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

(H) Reporter assays with constructs harboring the rs4684847 risk and non-risk allele in 3T3-L1 preadipocytes. Truncation of the PRRX1 matrix without affecting rs4684847 reveals abrogated allelic *cis*-regulatory activity. Mean±SD, n=9, paired t-test.

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

(J) Regulation of *PPARG2* mRNA expression in SGBS adipocytes with the CC risk allele, or TT non-risk allele introduced by CRISPR/Cas9 genome editing approach. siPRRX1 and siNT transfection concurrent with induction of differentiation, *PPARG2* mRNA assessed by qRT-PCR, standardized to *HPRT*. Mean±SD, n=12, t-test. siNT, non-targeting siRNA. See also Figure S4 and Table S17.

Figure 5. Binding of PRRX1 at the rs4684847 risk allele in human adipose cells affects lipid metabolism and insulin sensitivity.

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

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

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

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

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

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

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

(J) rs4684847-dependent rosiglitazone-mediated suppression of FFA-release during glyceroneogenesis. Pearson's correlation comparing siNT vs. siPRRX1. Mean±SD, t-test. See also Figure S4G,H and Table 1-2.

(K) The rs4684847 risk allele (C allele) promotes PRRX1 binding 6.5kb upstream of the *PPARG2* specific promoter, leading to suppression of *PPARG2* mRNA expression and perturbated lipid handling in adipose cells, increased circulating FFA levels, insulin resistance, and risk of T2D.

FC, fold change; FFA, free fatty acids.

2.2.2.7 Figures

Figure 1

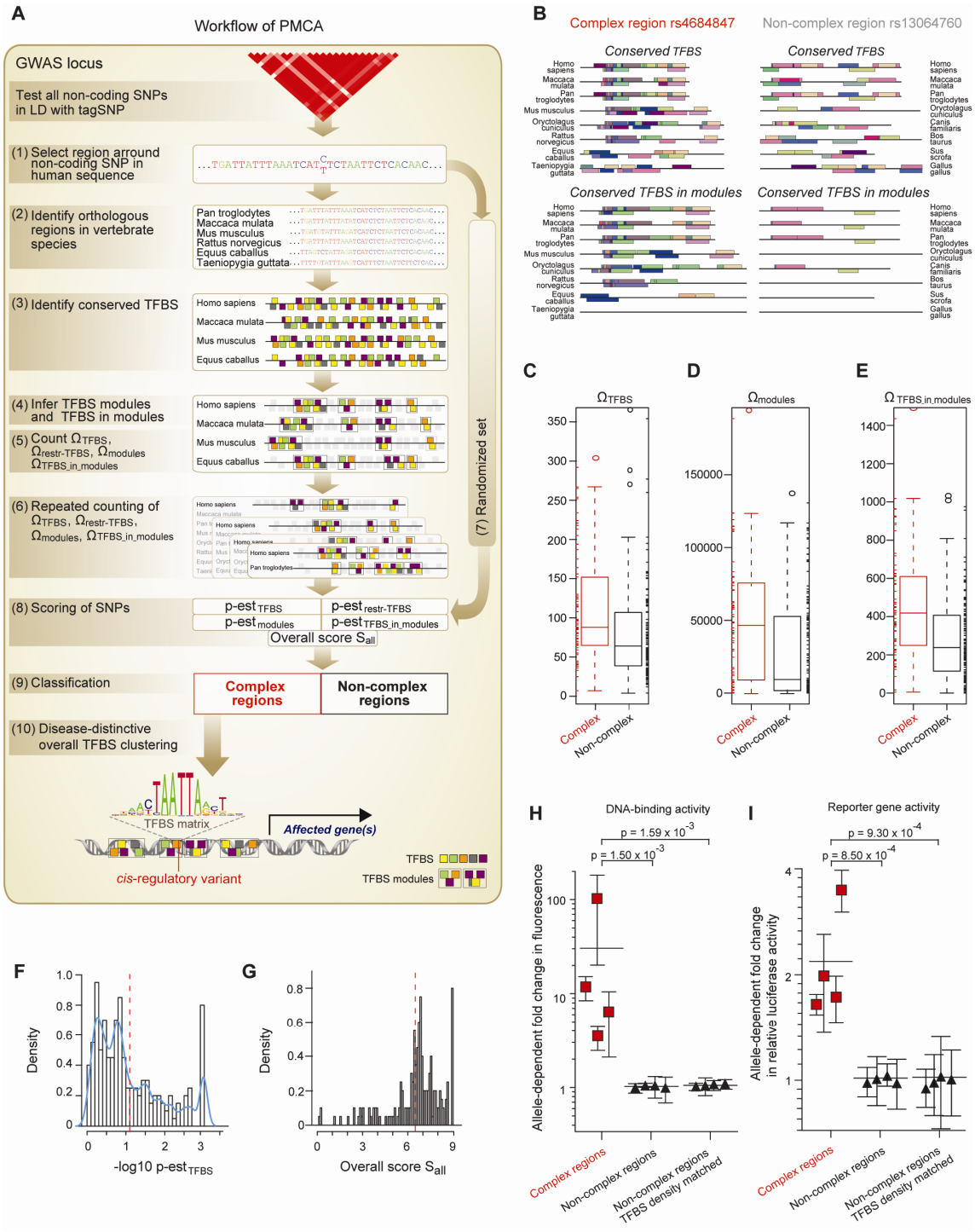


Figure 2

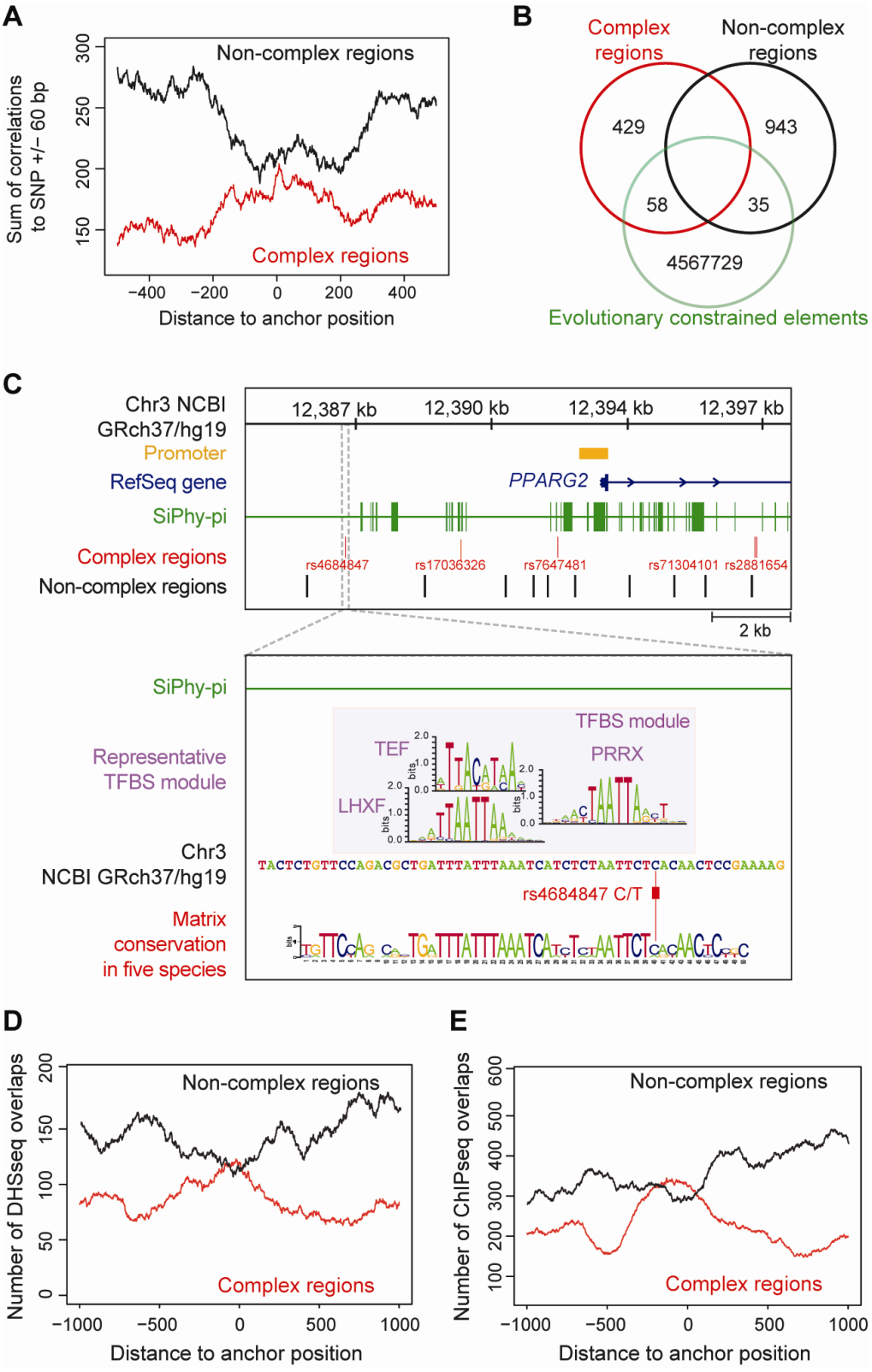

A
Non-complex regions
Complex regions
Sum of correlations to SNP +/– 60 bp
Distance to anchor position
B
Complex regions
Non-complex regions
429
943
58
35
4567729
Evolutionary constrained elements
C
Chr3 NCBI GRch37/hg19
12,387 kb
12,390 kb
12,394 kb
12,397 kb
Promoter
RefSeq gene
PPARG2
SiPhy-pi
Complex regions
Non-complex regions
rs4684847
rs17036326
rs7647481
rs71304101
rs2881654
2 kb
SiPhy-pi
Representative TFBS module
TFBS module
TEF
PRRX
LHXF
Chr3 NCBI GRch37/hg19
rs4684847 C/T
Matrix conservation in five species
D
Non-complex regions
Complex regions
Number of DHSseq overlaps
Distance to anchor position
E
Non-complex regions
Complex regions
Number of ChIPseq overlaps
Distance to anchor position

Figure 3

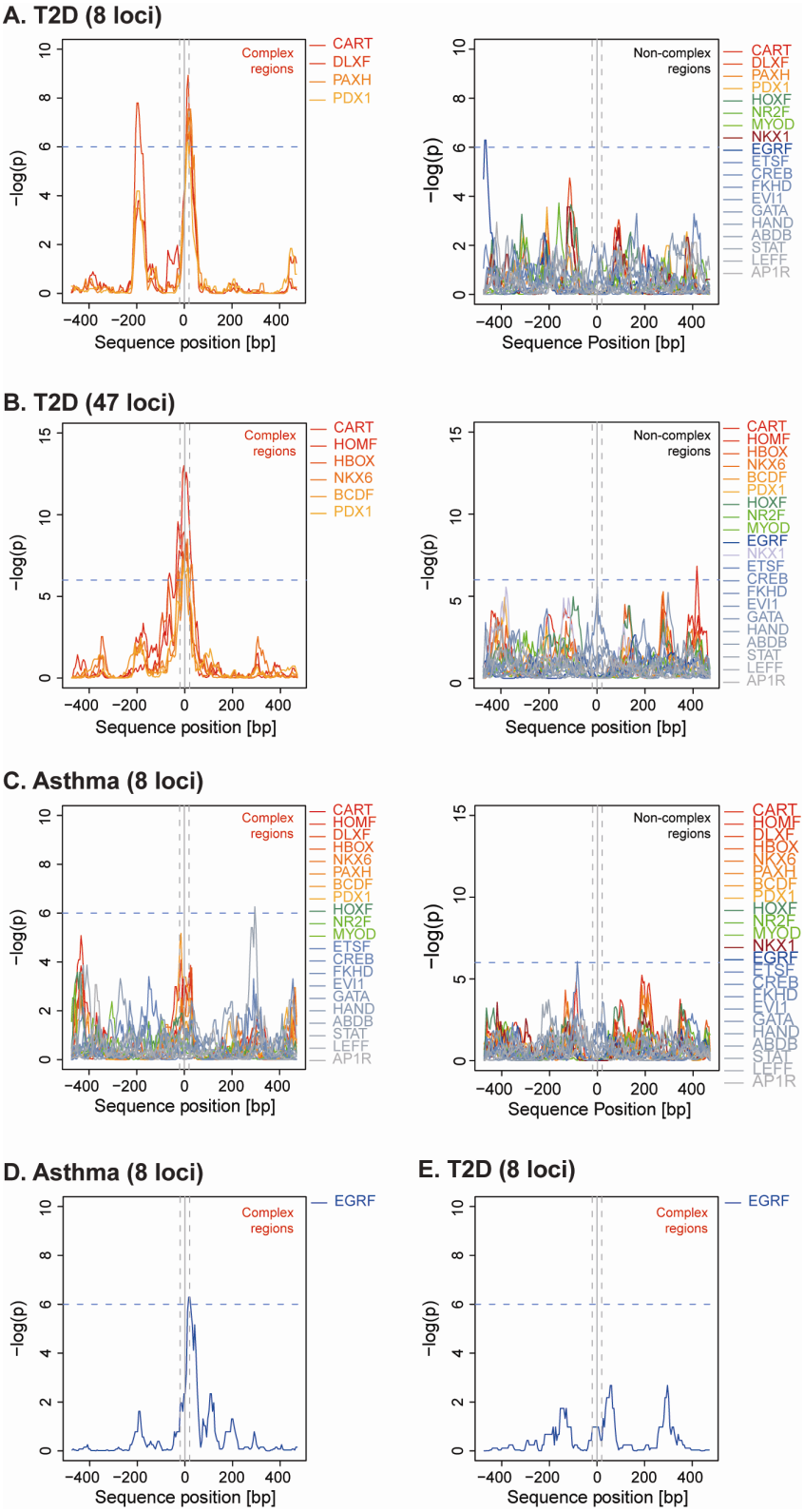

A. T2D (8 loci)
Complex regions
CART
DLXF
PAXH
PDX1
Non-complex regions
HOXF
NR2F
MYOD
NKX1
EGRF
ETSF
CREB
FKHD
EVI1
GATA
HAND
ABDB
STAT
LEFF
AP1R
−log(p)
Sequence position [bp]
Sequence Position [bp]
B. T2D (47 loci)
HOMF
HBOX
NKX6
BCDF
C. Asthma (8 loci)
D. Asthma (8 loci)
E. T2D (8 loci)

Figure 4

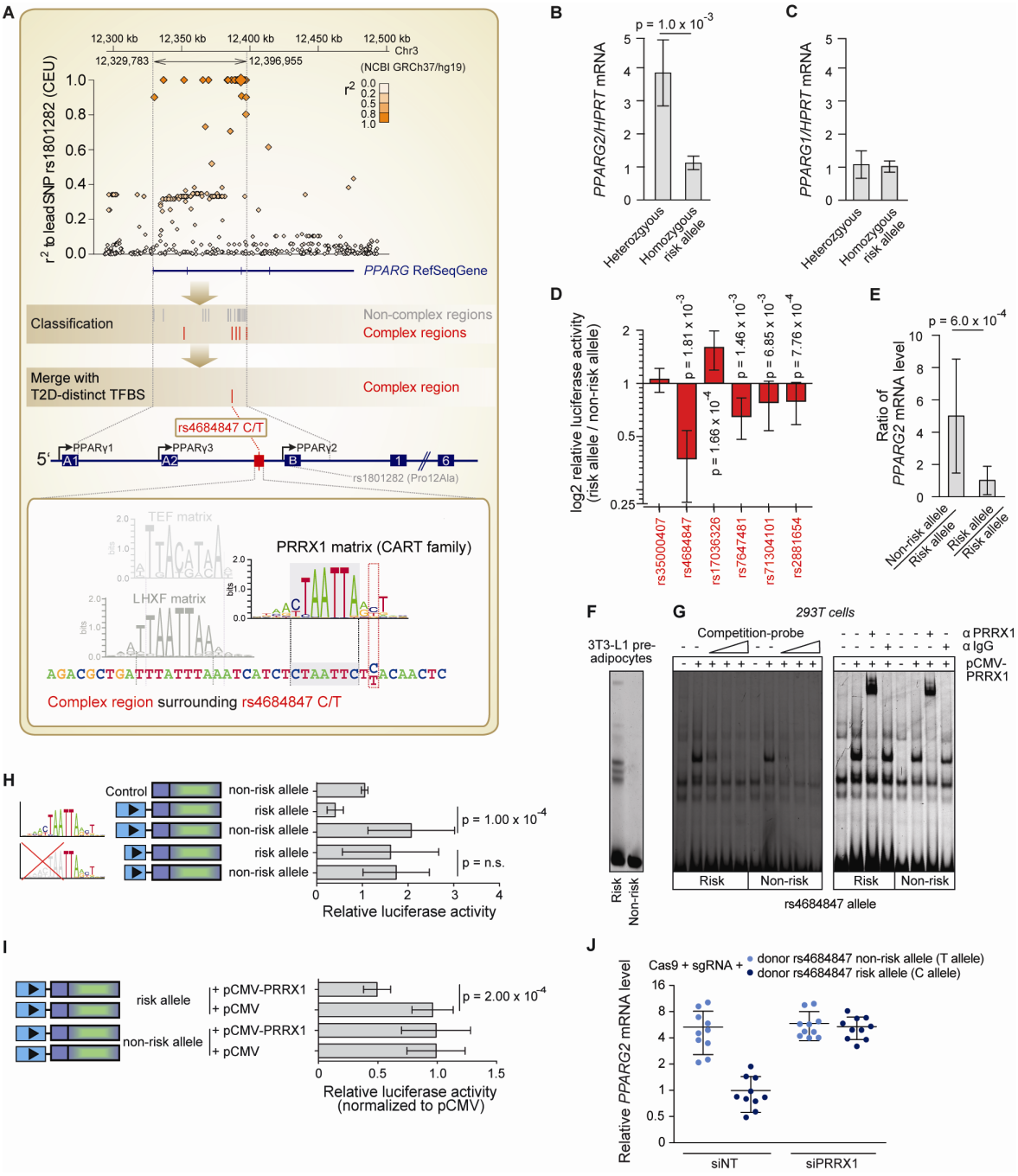
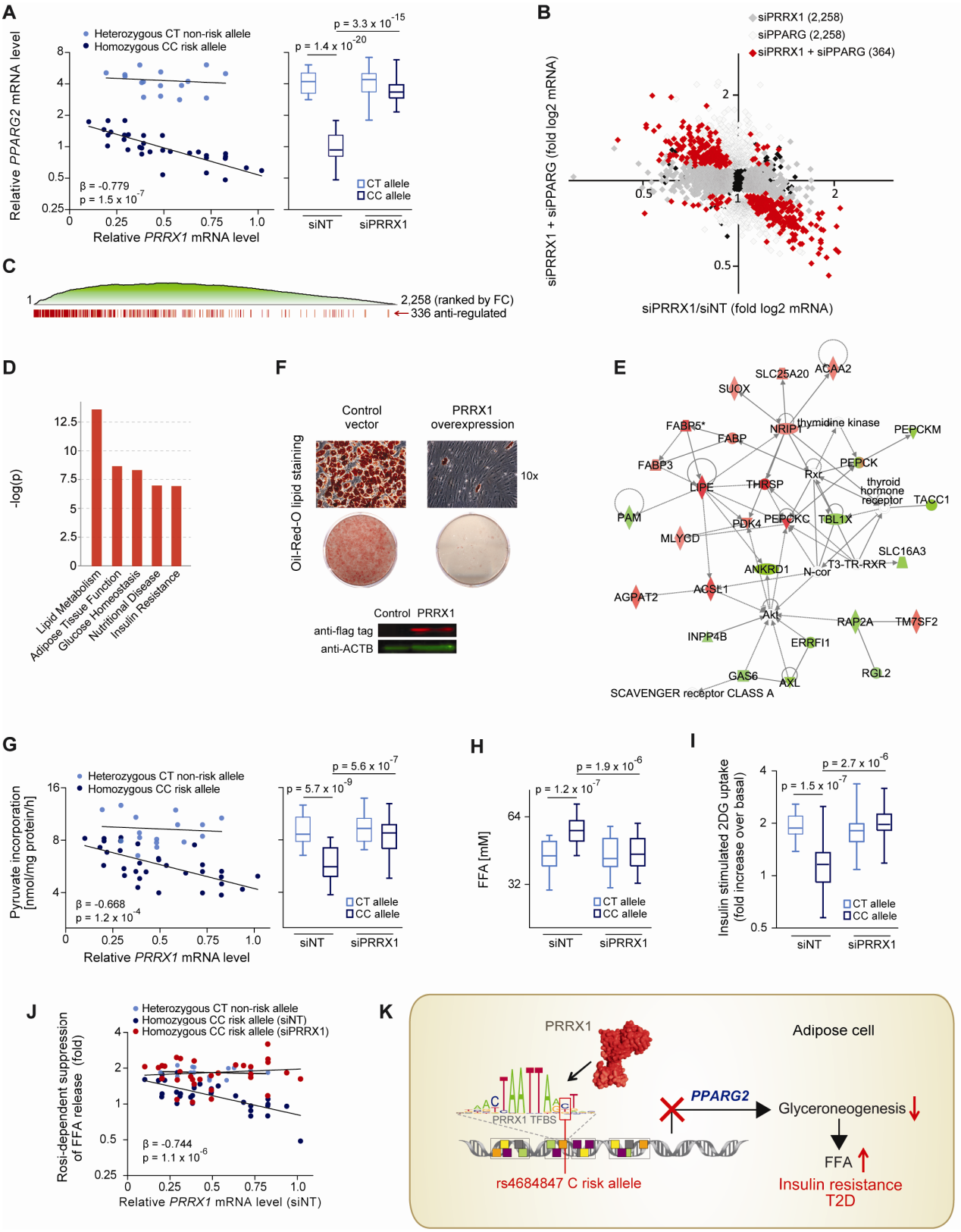

A
12,300 kb
12,350 kb
12,400 kb
12,450 kb
12,500 kb
Chr3
12,329,783
12,396,955
(NCBI GRCh37/hg19)
r2 to lead SNP rs1801282 (CEU)
PPARG RefSeqGene
Classification
Non-complex regions
Complex regions
Merge with T2D-distinct TFBS
Complex region
rs4684847 C/T
PPARγ1
PPARγ3
PPARγ2
rs1801282 (Pro12Ala)
TEF matrix
LHXF matrix
PRRX1 matrix (CART family)
AGACGCTGATTTATTTAAATCATCTCTAATTCTCACAACTC
Complex region surrounding rs4684847 C/T
B
p = 1.0 x 10-3
PPARG2/HPRT mRNA
Heterozygous
Homozygous risk allele
C
PPARG1/HPRT mRNA
D
log2 relative luciferase activity (risk allele / non-risk allele)
p = 1.81 x 10-3
p = 1.66 x 10-4
p = 1.46 x 10-3
p = 6.85 x 10-3
p = 7.76 x 10-4
rs35000407
rs4684847
rs17036326
rs7647481
rs71304101
rs2881654
E
p = 6.0 x 10-4
Ratio of PPARG2 mRNA level
Non-risk allele
Risk allele
F
3T3-L1 pre-adipocytes
Risk
Non-risk
G
293T cells
Competition-probe
α PRRX1
α IgG
pCMV-PRRX1
Risk
Non-risk
rs4684847 allele
H
Control
non-risk allele
risk allele
p = 1.00 x 10-4
p = n.s.
Relative luciferase activity
I
risk allele
non-risk allele
+ pCMV-PRRX1
+ pCMV
p = 2.00 x 10-4
Relative luciferase activity (normalized to pCMV)
J
Cas9 + sgRNA +
donor rs4684847 non-risk allele (T allele)
donor rs4684847 risk allele (C allele)
Relative PPARG2 mRNA level
siNT
siPRRX1

Figure 5


A
Heterozygous CT non-risk allele
Homozygous CC risk allele
Relative PPARG2 mRNA level
β = -0.779
p = 1.5 x 10-7
Relative PRRX1 mRNA level
p = 3.3 x 10-15
p = 1.4 x 10-20
CT allele
CC allele
siNT
siPRRX1
B
siPRRX1 (2,258)
siPPARG (2,258)
siPRRX1 + siPPARG (364)
siPRRX1 + siPPARG (fold log2 mRNA)
siPRRX1/siNT (fold log2 mRNA)
C
2,258 (ranked by FC)
336 anti-regulated
D
-log(p)
Lipid Metabolism
Adipose Tissue Function
Glucose Homeostasis
Nutritional Disease
Insulin Resistance
F
Control vector
PRRX1 overexpression
Oil-Red-O lipid staining
10x
Control PRRX1
anti-flag tag
anti-ACTB
E
SLC25A20
ACAA2
SUOX
FABP5*
FABP
NRIP1
thymidine kinase
PEPCKM
FABP3
PEPCK
LIPE
THRSP
Rxr
thyroid hormone receptor
TACC1
PAM
PDK4
PEPCKC
TBL1X
MLYCD
SLC16A3
ANKRD1
N-cor
T3-TR-RXR
AGPAT2
ACSL1
Akt
RAP2A
TM7SF2
INPP4B
ERRFI1
GAS6
AXL
RGL2
SCAVENGER receptor CLASS A
G
Pyruvate incorporation [nmol/mg protein/h]
Heterozygous CT non-risk allele
Homozygous CC risk allele
β = -0.668
p = 1.2 x 10-4
Relative PRRX1 mRNA level
p = 5.6 x 10-7
p = 5.7 x 10-9
CT allele
CC allele
siNT
siPRRX1
H
FFA [mM]
p = 1.9 x 10-6
p = 1.2 x 10-7
CT allele
CC allele
siNT
siPRRX1
I
Insulin stimulated 2DG uptake (fold increase over basal)
p = 2.7 x 10-6
p = 1.5 x 10-7
CT allele
CC allele
siNT
siPRRX1
J
Rosi-dependent suppression of FFA release (fold)
Heterozygous CT non-risk allele
Homozygous CC risk allele (siNT)
Homozygous CC risk allele (siPRRX1)
β = -0.744
p = 1.1 x 10-6
Relative PRRX1 mRNA level (siNT)
K
PRRX1
Adipose cell
PRRX1 TFBS
PPARG2
Glyceroneogenesis
FFA
Insulin resistance
T2D
rs4684847 C risk allele

2.3 Computer implemented method for identifying regulatory regions of regulatory variations and diagnostic means and methods for type 2 diabetes

Large parts of the manuscript described in and Chapter 2.1 and Chapter 2.2 are the basis for two patents that have been filed and published by the Technical University of München. Those patents are based on my novel computational concept of exploiting cross-species conservation in terms of a complexity assessment of co-occurring TFBS within CRM, regardless of the Phylogenetic conservation at nucleotide level. . I was largely involved in writing the patents.

The patent "COMPUTER IMPLEMENTED METHOD FOR IDENTIFYING REGULATORY REGIONS OR REGULATORY VARIATIONS" includes the bioinformatics PMCA methodology as described in Chapter 2.1 and can be viewed online (http://patentscope.wipo.int/search/en/detail.jsf?docId=WO2013024173&recNum=87&docAn=EP2012066144&queryString=obesity&maxRec=32487 ; Hyperlink: Claussnitzer_2013-1). The invention relates to the analysis of patterns of co-occurring TFBS across species for the computational identification of regulatory elements and regulatory sequence variants in the human genome.

The patent "DIAGNOSTIC MEANS AND METHODS FOR TYPE 2 DIABETES" relates to the application of the PMCA methodology on genetic susceptibility loci that have been associated with T2D. The patent can be viewed online (http://patentscope.wipo.int/search/en/detail.jsf?docId=WO2013024175&recNum=217&docAn=EP2012066150&queryString=%28%20& ; hyperlink: Claussnitzer_2013-2). The patent was filed for a total of 41 *cis*-regulatory variants that were, by using the PMCA framework, computationally pinpointed. I validated the *cis*-regulatory functionality of those variants with a series of wet lab experimental approaches as described in Chapter 2.2

2.4 Functional characterization of promoter variants of the adiponectin gene complemented by epidemiological data.

The content of this chapter has been published in Laumen et al, 2009, and is described below. For this manuscript, I could apply parts of the computational cross-species TFBS analysis on sequence variants that have been statistically associated with circulating adiponectin levels. The computational analysis served to select potential *cis*-regulatory SNP variants within the adiponectin locus, which might explain the statistical genotype association with circulating adiponectin levels. The *in silico* analysis pinpointed SNP variants, i.e. rs16861194, rs17300539, rs266729, that localize in a potential CRM within in the adiponectin promoter region.

2.4.1 Introduction

Adipose tissue produces and releases a variety of factors, which may be directly involved in the pathophysiology of obesity-associated insulin resistance (Hauner, 2005). One of the most interesting candidates with respect to the development of metabolic syndrome and type 2 diabetes is the *APM1* gene that encodes the abundantly expressed protein adiponectin. Circulating adiponectin concentrations are negatively associated with insulin resistance and atherosclerosis and are decreased in humans with type 2 diabetes, coronary artery disease, or obesity (Matsuzawa, 2006). Animal experiments showed that administration of adiponectin reduces blood glucose levels, improves insulin resistance, and directly ameliorates endothelial dysfunction (Yamauchi et al., 2001; Berg et al., 2001; Ouedraogo et al., 2007). Furthermore,

low adiponectin levels are associated with other components of the metabolic syndrome, such as hypertension and dyslipidemia (Kadowaki et al., 2006).

APM1 maps to chromosome 3q27, a region known to be linked to type 2 diabetes and the metabolic syndrome (Vionnet et al., 2000). In view of the important role of circulating adiponectin in the pathogenesis of major metabolic disorders, several studies have addressed the correlation of *APM1* SNPs with adiponectin levels. They revealed a significant correlation between two SNPs, rs266729 and rs17300539, and adiponectin levels (Vasseur et al., 2000; Vasseur et al., 2002; Heid et al., 2006). In one of these studies, the functional activity of both SNPs for transcriptional regulation as promoter elements was analyzed by luciferase assay (Bouatia-Naji et al., 2006). However, these experiments were performed in COS7 cells that do not express adiponectin and hence do not represent an ideal cell system for this type of analysis. We performed transfection experiments using mutated promoter constructs in 3T3-L1 adipocytes expressing endogenous adiponectin and analyzed DNA binding activity of different haplotype combinations of three promoter SNPs. Two of the selected SNPs are known to be associated with adiponectin levels and the third one lies in close proximity. This prompted us to assume that all three may be located in a transcriptionally functional element that may be altered by one or all SNPs. The relevance of these SNP haplotypes for human adiponectin levels was investigated in 1,692 participants of the MONICA/KORA (Cooperative Health Research in the Region of Augsburg) S123 cohort as well as in 696 participants of the KORA S4 cohort.

2.4.2 Research Design and Methods

SNP selection.

We searched for SNPs in the promoter region of the *APM1* gene that *1*) co-localize with putative transcription factor binding sites, *2*) have been reported to be associated with

adiponectin level or other adiponectin-related traits, and *3*) lie in close genomic proximity. The first criterion is based on the assumption that SNPs may interfere with the functionality of a binding site, and the second should ensure previous epidemiological SNP association with adiponectin or related parameters. The rationale for the third criterion is the fact that transcription factor binding sites are often found in close proximity and build a functional module; the combination of different transcription factor binding sites is usually essential for regulation of transcription. If potential regulatory SNPs are found in such a potential module, it may enhance the probability that they are indeed functional SNPs. We hypothesized that SNPs combining both properties were most likely to alter a functional module in humans. The SNP2 (rs17300539, G>A) showed the strongest association with adiponectin levels in several studies. This SNP has been chosen together with two additional SNPs that both have shown association with type 2 diabetes (SNP1 = rs16861194 A>G; SNP3 = rs266729 C>G; in MONICA/KORA S123, rs1648707 A>C has served as proxy for SNP3 with linkage disequilibrium values of $r = 0.84$ and $D' = 1$). All three together are located in a small 80-bp part of the adiponectin promoter/enhancer region. Furthermore, all three SNPs lie within putative transcription factor binding sites that are shown in Figure 1. These putative binding sites have been predicted using the Genomatix software (Genomatix, Munich, Germany).

Functional studies

Cell culture.

The mouse preadipocyte cell line 3T3-L1 was cultured as described (Laumen et al., 2008). To promote adipose differentiation, Dulbecco's modified Eagle's medium containing 10% fetal calf serum was supplemented with 250 nmol/l dexamethasone and 0.5 mmol/l isobutyl-methylxanthine for the first 3 days and 66 nmol/l insulin throughout the whole differentiation period.

Transfection of cells.

3T3-L1 cells were transfected on day 0, 6, and 8 of differentiation, respectively, using the Lipofectamine 2000 transfection reagent (Invitrogen, Karlsruhe, Germany). A total of 2 µg DNA and 2 µl transfection reagent were mixed according to the manufacturer's instructions and added to the cells for 4 h. Then 24–48 h after transfection, luciferase activity was measured using the dual-luciferase reporter assay (Promega, Mannheim, Germany). In all transfections, 0.2 µg ubiquitin-promoter renilla luciferase vector was cotransfected to normalize for transfection efficiency.

Cloning and mutagenesis of adiponectin promoter luciferase vectors.

A 2,100-bp adiponectin promoter (bases −2,125 to +41) was PCR amplified from genomic DNA using an iProof High-Fidelity PCR Kit (Bio-Rad, Germany) and cloned into the pGL3basic luciferase vector (Promega, Germany) as described recently (Kita et al., 2005), using the primers depicted in Table S1 in the supplemental material (found in an online-only appendix at http://dx.doi.org/10.2337/db07-1646). The haplotype configuration of the cloned promoter was determined by sequencing and was shown to carry the major allele (M) of the three SNPs described above (MMM-luc). Using the QuickChange Multi Site-Directed Mutagenesis Kit (Stratagene, Germany) and the primers listed in Table S1 (supplemental material), the variants of the three SNPs were introduced in all remaining seven possible combinations (for primers, see Table S2). All vectors were sequenced to confirm the correct SNP variation combination.

Electrophoretic mobility shift assay.

Probes for electrophoretic mobility shift assay (EMSA) were amplified by PCR from the above-described luciferase vectors carrying the eight different SNP combinations using the primers EMSA 5′ and 3′ (Table S1). The resulting 80-bp probe spans the APM1 gene corresponding to chromosome 3 position 188042104–188042183 (for sequence, see Figure S1

in the supplemental material). Primers contained a synthetic *Hind*III site, and resulting probes were cut to enable radioactive Klenow fill-in. EMSA was performed with 2–4 µg nuclear protein extract and with 30,000–50,000 cpm of a ^{32}P-labeled probe as described previously (Schorpp et al., 1995).

Statistical analysis of transfection studies.

Overall, statically comparisons were performed using the Kruskal-Wallis test followed by pair-wise testing using the Dunn's multiple comparison test.

Epidemiological investigation

The KORA S4 and the MONICA/KORA S123 sample.

The KORA Survey S4 (formerly known as S2000) is a population-based study of adults recruited from 1999 to 2001 conducted under the same conditions as the previous three surveys (S1, S2, S3) with patients recruited during the years 1984–1995 in the World Health Organization MONICA project. Details of the surveys are reported elsewhere (Loewel et al., 2005). Study participants from all four surveys were from the study region of Augsburg (German nationality). Measures of weight and height were available to compute the BMI. All participants gave their written informed consent.

A subsample of the KORA S4 survey including 696 subjects aged 55–74 years with ~50% men was designed to address objectives regarding pre-diabetic stages. Adiponectin was measured in these subjects using the human adiponectin radioimmunoassay from Linco Research (St. Charles, MO) as described previously (Rathmann et al., 2006; Herder et al., 2006).

From the above-stated MONICA/KORA surveys S1, S2, and S3, a number of the 1,692 subjects aged 35–74 years with equal gender distribution were selected randomly from each survey as a subcohort sample (MONICA/KORA S123 sample) (Thorand et al., 2005).

Adiponectin levels were measured using the human adiponectin enzyme-linked immunosorbent assay from Mercodia (Uppsala, Sweden). The intra- and interassay coefficients of variation of control sera were 3.2% and 5.8%, respectively.

Genotyping.

PCR primers were designed by Sequenom's MassARRAY Assay Design program. Genotyping analyses were carried out by means of matrix-assisted laser desorption ionization–time of flight analysis of allele-dependent primer extension products as described elsewhere (Weidinger et al., 2004). Genotyping calls were made in real time with MassARRAY RT software (Sequenom, San Diego, CA). Negative controls were included in all assays. In the 12.5% of randomly selected samples genotyped in duplicate, the discordance rate was 0.3%.

Statistical analysis of epidemiological data.

Statistical SNP and haplotype association analysis was performed using the SAS procedure SURVEYREG to account for the sampling scheme in the MONICA/KORA S123 sample in the estimation of the standard errors of association estimates; in the KORA S4 sample, linear regression was applied using the GLM procedure. The logarithm of adiponectin was used as the outcome variable to yield a normal distribution. All analyses were adjusted for age, sex, and BMI. An additive as well as a dominant genetic model was applied. The minor allele frequencies of SNPs were computed, and linkage disequilibrium was assessed. SNPs were tested for Hardy-Weinberg equilibrium.

Haplotypes were estimated from genotypes via the expectation-maximization algorithm using the *R* statistics package (haplo.em) or SAS version 9.1 (Schaid et al., 2002). Haplotypes were used in the regression models including all haplotypes except the most common haplotype to compute the association with adiponectin per copy of a haplotype adjusted for the other

haplotypes compared with the group of subjects with two copies of the most common haplotype.

2.4.3 Results

SNP variants and haplotypes in the two epidemiological KORA samples.

In the two epidemiological cohorts KORA S4 and MONICA/KORA S123, we analyzed the three SNPs that we had selected for our investigation as candidates for an adiponectin-regulating role because of their nearby location in the APM1 promoter, because of their possible interference to transcription factor binding sites, and because of previous reports about association with adiponectin or related phenotypes. The SNPs analyzed were rs16861194 (SNP1), rs17300539 (SNP2), and rs266729 (SNP3) in KORA S4 and rs1648707 in MONICA/KORA S123 as proxy for SNP3 (as described in RESEARCH DESIGN AND METHODS). The r^2 values as a measure of linkage disequilibrium were 0.006 (0.009) and 0.024 (0.179) for SNP1 compared with SNP2 and SNP3, respectively, and 0.038 (0.049) for SNP2 compared with SNP3 in the KORA S4 sample (and the MONICA/KORA S123 sample); hence, they are not in linkage disequilibrium. We statistically reconstructed the haplotypes; of the theoretically possible eight haplotypes across the three SNPs, we observed five (MMM, MmM, MMm, mMM, and mMm, with M and m indicating the major or minor allele, respectively). SNP and haplotype characteristics are given in Table 1 and Table 2, respectively, indicating a large consistency of allele frequencies in the two cohorts.

APM1 promoter activity during adipocyte differentiation.

The MMM adiponectin promoter construct containing the major allele M in all three SNPs was transfected into 3T3-L1 preadipocyte cells (d0) and 3T3-L1 adipocytes 6 and 8 days after induction of differentiation (d6 and d8). We measured a significant induction of luciferase

activity during adipocyte differentiation (threefold on day 6 and fivefold on day 8, respectively) demonstrating the functionality of the promoter (Figure 2).

Next, we focused on the five haplotypes observed in the KORA samples (MMM, MmM, MMm, mMM, and mMm). A significant overall difference between promoter activities was observed (Figure 2A), and each promoter construct revealed significant induction of luciferase activity upon differentiation compared with respective transfections in undifferentiated cells in pair-wise comparisons. On day 6 of differentiation, we found a tendency of 50% higher promoter activity of the MMM promoter compared with the mMM, MmM, and MMm promoters. On day 8 of differentiation, the most striking difference was observed between the MMM and the mMM promoter, with MMM showing a threefold higher activity than mMM ($P < 0.05$). Notably, both promoters with the minor allele at the SNP1 position (mMM and mMm) revealed impaired basal promoter activity compared with the MMM promoter already in preadipocytes.

Next, we transfected cells with promoters regulated by the theoretically possible, using nonexisting haplotypes in the epidemiological samples (Mmm, mmM, and mmm), and observed the strongest impact by the threefold major to minor allele alteration (mmm promoter) with a complete loss of basal promoter activity in preadipocytes. Additionally, the mmm promoter was almost resistant to transcriptional activation during differentiation supporting the importance of these sites for transcription of the *APM1* gene. Interestingly, all promoters with the minor allele at the SNP1 position showed the strongest reduction of basal promoter activity or altered kinetic of activity during differentiation compared with the MMM promoter, suggesting a crucial role of SNP1 for promoter activation.

Impact of rosiglitazone on APM1 activation depending on SNP variant combinations.

To investigate whether the different haplotype constructs had an impact on the inducibility of the *APM1* gene promoter, we transfected 3T3-L1 adipocytes on day 6 after induction of differentiation and determined luciferase activity in the presence or absence of rosiglitazone. Promoters with MMM, MmM, MMm, or Mmm haplotypes revealed a two- to fivefold induction of luciferase activity after treatment with rosiglitazone compared with control treated cells (Figure 3). In contrast, all other haplotypes with the minor variant at the SNP1 position (mMM, mMm, and mmm) showed no response upon rosiglitazone treatment.

Influence of APM1 SNP variations on DNA binding activity.

To analyze whether these haplotypes have an impact on DNA binding activity of nuclear proteins, we performed EMSAs using nuclear extracts from undifferentiated 3T3-L1 (preadipocytes) and in vitro differentiated 3T3-L1 adipocytes (day 6 after induction of differentiation) and DNA probes with all eight possible haplotypes. We found one major complex and some minor slower migrating complexes using the DNA probe with the major SNP variants (MMM) and nuclear extracts from preadipocytes. Nuclear extracts from differentiated adipocytes revealed a slight decrease of the major complex and increased binding of a slower migrating complex. Most haplotypes showed similar patterns of DNA binding compared with MMM. In contrast, the mMM probe revealed strongly reduced DNA binding activity with nuclear extracts from preadipocytes, but comparable binding of major and minor complexes in differentiated adipocytes. Finally, we performed EMSAs with nuclear extracts of preadipocytes and adipocyte cultures that were induced with rosiglitazone to investigate whether stimulation affects DNA binding activity. Rosiglitazone treatment of preadipocytes abolished protein binding to DNA probes with MMM and the most other haplotypes (MmM, MMm, mmM, Mmm, and mmm), whereas no inhibition of DNA binding was found for the mMm variant. Moreover, we detected restored DNA binding activity for the

mMM variant. Surprisingly, rosiglitazone treatment of differentiated adipocytes had no major impact on DNA binding activity (Figure 4).

Association of SNPs and haplotypes with circulating adiponectin in the epidemiological samples.

Table 3 and Table 4 summarize the results of SNP and haplotype association analyses in the MONICA/KORA S123 (1,692 participants) and the KORA S4 (696 participants) samples. Subjects carrying the minor allele of SNP1 showed consistently lower circulating adiponectin levels in both cohorts, which was statistical significant in the larger MONICA/KORA S123 sample ($P = 0.001$), but not in the smaller KORA S4 sample. Consistent to the SNP1 finding, all haplotypes found in the studies containing the minor allele for SNP1 showed reduced adiponectin level, which was statistically significant for the more frequent mMm ($P = 0.009$). This observation of lower adiponectin level for the SNP1 minor allele and the respective haplotypes is in line with the promoter assay finding of a reduced activity for these haplotypes. Subjects carrying the minor allele in SNP2 showed a significant increase in adiponectin levels in both KORA samples ($P = 0.00005$ and $P < 10^{-9}$), which was in line with haplotype analysis for haplotype MmM ($P < 0.0001$ and $P = 0.0002$), but did not fit with the promoter activity assays. SNP3 showed decreased adiponectin levels in all studies, which was statistical significant in the larger MONICA/KORA S123 sample ($P = 0.00001$). Three haplotypes (mmM, Mmm, and mmm) were neither present in any subject of the KORA S4 nor in the MONICA/KORA S123 sample. Given the sample size of 1,676 (696) in the MONICA/KORA S123 (KORA S4) sample and the haplotype frequency of 0.0025 (0.0015) as expected from the minor allele frequencies of the three SNPs, the finding of zero subjects with the mmm haplotype was statistically significantly different from what would have been expected by chance ($P= 0.0167$ in MONICA/KORA S123, $P = 0.3534$ in KORA S4, $P =$

0.0059 for both samples combined). This observation is in line with the observation of a complete loss of promoter activity.

2.4.4 Discussion

Recent epidemiological studies support the concept that SNPs in the *APM1* gene are associated with type 2 diabetes and other metabolic disorders in several populations (Ouedraogo et al., 2007). In the current study, we investigated three different SNPs (SNP1 = rs16861194, SNP2 = rs17300539, and SNP3 = rs266729 or rs1648707) in the *APM1* gene promoter region located within an 80-bp region of the promoter that are known for their association with circulating adiponectin levels or related phenotypes (Vasseur et al., 2002; Heid et al., 2006; Bouatia-Naji et al., 2006). We applied an approach of combining functional experiments with epidemiological data and showed that these SNPs influence basal and inducible *APM1* promoter activity in 3T3-L1 adipocytes accompanied by alterations in DNA binding activity. In human epidemiological studies, we presented SNP and haplotype association analyses of two population-based samples of the MONICA/KORA studies, which was consistent with most of our functional findings.

Intriguingly, the constructed promoter with the minor allele (mmm) in all SNPs was almost completely inactive with regard to basal activity and differentiation- or rosiglitazone-induced activity. Our results clearly demonstrate the functional relevance of these SNPs for activation of the *APM1* promoter by interfering with transcription factor binding sites. Indeed, we found specific binding of nuclear proteins to a DNA probe containing all three minor alleles as well as changes in the pattern of DNA-protein complexes upon adipocyte differentiation and partly also upon rosiglitazone stimulation. Given the highly reduced promoter activity and low circulating adiponectin levels being associated with increased risk of severe diseases such as type 2 diabetes and coronary heart disease (Yamauchi et al., 2001; Berg et al., 2001;

Tschritter et al., 2003; Stefan et al., 2003; Hotta et al., 2000; Kumada et al., 2003; Fruebis et al., 2001; Yamamoto et al., 2002; Iwashima et al., 2004), one may speculate that the mmm haplotype affects adiponectin expression in vivo to an extent that might be disadvantageous. This hypothesis is supported by our epidemiological finding that none of the 2,340 subjects in our analysis carried this haplotype, which was highly significantly different from what would have been expected by chance ($P = 0.0059$). Yet, further studies in humans are necessary to support this hypothesis. The importance of haplotype combination has also been shown for SNP2 and SNP3, which in combination, increases the risk of diabetes (Schwarz et al., 2006).

The transcription factors involved in the regulation of the *APM1* promoter were analyzed by several groups (Qiao et al., 2005; Kim et al., 2006; Iwaki et al., 2003); however, most studied promoter regions do not contain the SNPs analyzed here. One study demonstrated slightly higher promoter activity upon deletion of the promoter region containing these SNPs (Kita et al.; 2005). However, such deletion of promoter regions removes all regulatory sites and hence does not allow a SNP-specific analysis concerning the influence on binding characteristics of transcriptional activators or repressors. Indeed, our EMSA experiments revealed the existence of specific DNA binding complexes that are affected by adipocyte differentiation and rosiglitazone stimulation. At least two minor alleles in the haplotype (mMM or mMm) exhibited obvious alterations in DNA binding complexes. Surprisingly, the presence of relevant DNA binding factors was already found in preadipocytes. It has to be considered that epigenetic mechanisms may also be involved in the regulation of the analyzed promoter. It is known that transcription of the*APM1* gene is regulated by histone acetylation (Musri et al., 2006). Interestingly, DNA binding activity upon rosiglitazone stimulation critically depends on the combination of several SNP variants. This supports the view that this promoter region represents a functional module with binding of various proteins that interact and build a more complex structure. As an example, we found alterations in DNA binding activity using the

DNA probe with the minor SNP1 allele and nuclear extracts from preadipocytes. Further introduction of a minor allele in SNP2 restored normal DNA binding, whereas introduction of a minor allele in SNP3 resulted in a different pattern of DNA binding activity.

A potential limitation of our study is that mouse 3T3L1 cells may differ from human adipocytes regarding the presence of transcription factors. However, several transcription factors such as C/EBPs, SREBP, and PPARs were previously shown to regulate both the human and mouse adiponectin promoter, and the binding sites were well characterized (Qiao et al., 2005; Kim et al., 2006; Iwaki et al., 2003) (Figure 1). Bioinformatics binding site prediction revealed putative binding sites for the large family of homeodomain proteins and zinc finger proteins, yet clearly no binding sites for the so far known regulators of adiponectin. Although we could clearly show that the SNPs modulate DNA binding activity, the exact binding factors remain to be identified. A recent publication also suggested SNP3 by bioinformatics prediction to modify a zinc finger protein site (Zhang et al., 2008), but in this work, no attempt was made to analyze the influence of the SNP3 on DNA binding activity.

The correlation between elevated circulating adiponectin levels and the presence of the minor allele in SNP2 is in line with published data and with a recent report of increased promoter activity in COS-7 cells (Bouatia-Naji et al., 2006). However, COS-7 is not an adipocyte cell line and may not express an appropriate set of transcription factors expected in adipocytes. Direct adiponectin measurement of endogenous adiponectin is not possible in the available cell models, since the cell line does not contain the different genomic haplotypes, but our transfection studies represent a good model to address this aspect.

The minor allele of SNP2 resulted in a higher inducibility by rosiglitazone only in the combination with the major allele in SNP1. Even more for some minor allele constructs, this inducibility was increased from a lower basal level, whereas the haplotype with three minor alleles was not inducible at all. This serves as an additional hint for a functional interaction

between these two SNPs and furthermore that a functional analysis of SNPs should also take into account the activation state of cells. Peroxisome proliferator–activated receptor (PPAR)-γ agonists such as rosiglitazone are known to induce adiponectin expression in adipocytes (Gustafson et al., 2003; Yang et al., 2002). Furthermore, treatment of type 2 diabetic patients with rosiglitazone improves insulin sensitivity but stimulates fat accumulation (Hauner, 2002). The response to glitazones in humans could possibly differ depending on the promoter haplotype, which pinpoint a potential relevance of the APM1 promoter SNPs for improved individualized treatment.

We found obvious changes in the promoter and DNA binding activity when the minor allele in SNP1 was present. These findings are in line with the known association of SNP1 with hypoadiponectinemia (Vasseur et al., 2002; Heid et al., 2006; Bouatia-Naji et al., 2006) and the significantly lower adiponectin levels in our MONICA/KORA S123 sample. Moreover, the epidemiological haplotype data extend these findings, with adiponectin being downregulated by the minor allele of SNP1 and upregulated by the minor allele of SNP2. An important challenge for the future characterization of this functional module is the identification of the nuclear factors whose binding is affected by SNP1 and SNP2.

We used a combined functional and epidemiological approach and thus were able to overcome the drawback of each approach separately: On the one side, in epidemiological studies, it is not clear whether a significant SNP association is derived from the analyzed SNP directly or from a latent SNP in linkage disequilibrium. Significant haplotype associations can pinpoint a certain haplotype of interest, but it would remain to be shown which specific allele combination—and possibly including latent alleles between genotyped loci—would be of functional relevance. This is overcome by our functional haplotype promoter studies where all effects are clearly attributed to the distinct alterations analyzed, since polymorphisms in linkage disequilibrium have not been mutated; hence a major functional impact is exerted by

the combination of the here analyzed SNPs. On the other hand, functional studies alone do not allow the drawback to effects in humans. This limitation is overcome by adding epidemiological data, which support the functional findings regarding the regulation by SNP1 and SNP2, the effects of haplotype combination, and a potential negative selection of the haplotype with minor alleles in all three SNPs due to suppressed adiponectin promoter activity.

In conclusion, the present study on the *APM1* gene is the first one analyzing the functional activity of *APM1* regulatory SNPs in a cell model expressing endogenous adiponectin and shows the importance to consider SNP haplotypes. The epidemiological data support the functional findings and thereby underscore the relevance in humans. Our results demonstrate that promoter variants in the *APM1* gene are relevant for the regulation of adiponectin transcription. Furthermore, our study represents a suitable approach by combining functional and epidemiological methods to characterize the role of gene variants.

2.4.5 Tables and Figure Legends

Table 1.

Characteristics of SNPs in the sample of the KORA S4 study (*n* = 696) and the MONICA/KORA S123 study (*n* = 1,692)

	rs number	Position*	Call rate† (S4/S123)	Hardy-Weinberg equilibrium *P* value‡ (S4/S123)	Minor allele frequency (S4/S123)
SNP1	rs16861194	−11426	0.966/0.988	0.494/0.865	0.059/0.083
SNP2	rs17300539	−11391	0.951/0.990	0.347/0.897	0.091/0.090
SNP3§	Rs266729	−11377	0.967/0.989	0.773/0.990	0.278/0.333

Table 2.

Characteristics of haplotypes in the sample of the KORA S4 study (*n* = 696) and the MONICA/KORA S123 study (*n* = 1,653*)

	SNP1	SNP2	SNP3†	Frequency (S4/S123)
MMM	A	G	C	0.571/0.577
mMM	G	G	C	0.059/0.0003
MmM	A	A	C	0.091/0.090
MMm	A	G	G	0.278/0.249
Mmm	G	A	G	0/0
Mmm	A	A	G	0/0
mMm	G	G	G	0/0.083
mmM	G	A	C	0/0

Haplotypes are given by stating m or M for each of the three SNPs in a row indicating whether the haplotype exhibits the minor (m) or the major (M) allele at the SNP location. *For complete data for all three SNPs.

†Depicted is the genotype C>G of rs266729 measured in KORA S4; the proxy rs1648707 with genotype A>C (not depicted) was measured in MONICA/KORA S123.

Table 3.

SNP association analysis in the KORA S4 sample (n = 696) and in the MONICA/KORA S123 sample (n = 1,692)

					Coefficient (P)	
	SNP	Genotype	n	Mean[*] (µg/ml)	Additive[†]	Dominant[‡]
S123 (n = 1,692)		AA	1407	11.2062	Reference	Reference
	SNP1	AG	253	10.5341	−0.0602 (P = 0.0014)	−0.0634 (P = 0.002)
		GG	12	10.1812		
		GG	1388	10.7692	Reference	Reference
	SNP2	AG	274	12.7478	0.1665 ($P < 10^{-9}$)	0.1748 ($P < 10^{-9}$)
		AA	13	14.6926		
		AA	745	11.4570	Reference	Reference
	SNP3[§]	CA	743	10.8979	−0.0502 (P = 0.00001)	−0.0598 (P = 0.0001)
		CC	185	10.3596		
S4 (n = 696)		AA	596	8.929	Reference	Reference
	SNP1	AG	73	8.536	−0.0362 (P = 0.507)	−0.0248 (P = 0.6677)
		GG	3	4.967		

	SNP	Genotype	*n*	Mean* (µg/ml)	Coefficient (*P*)	
					Additive†	Dominant‡
		GG	544	8.536	Reference	Reference
	SNP2	AG	115	10.164	0.1897 (*P* = 0.00005)	0.2042 (*P* = 0.00003)
		AA	3	9.7		
		CC	349	9	Reference	Reference
	SNP3§	GC	274	8.876	−0.0287 (*P* = 0.3273)	−0.0185 (*P* = 0.6167)
		GG	50	8.422		

Data are from linear regression on log(adiponectin), adjusted for age, sex, and BMI and survey (for the S123 sample) using an additive or a dominant genetic model. *Geometric mean of adiponectin concentrations in micrograms adiponectin per milliliter serum.

†Mean change in log(adiponectin) per copy of the minor allele.

‡Mean change in log(adiponectin) for subjects to the indicated reference (e.g., SNP1 with the AG or GG compared with the AA).

§For SNP3 in the case of KORA S4, the genotype C>G of rs266729 is depicted; in the case of MONICA/KORA S123, the genotype A>C of the proxy rs1648707 is depicted.

Table 4

Haplotype association analysis in the KORA S4 sample (*n* = 696) and the MONICA/KORA S123 sample (*n* = 1,653*)

	Haplotype		*n*	Geometric mean†	Coefficient
S123 (*n* = 1,676)	MMM	0/1/2	300/799/554	11.343/11.035/11.029	Reference
	MmM	0/1/2	1,368/272/13	10.761/12.738/14.683	0.15641 (*P* < 0.0001)

	Haplotype		*n*	Geometric mean†	Coefficient
	MMm	0/1/2	927/628/98	11.298/10.832/10.792	−0.022 (*P* = 0.1009)
	mMM	0/1/2	NA	NA	NA
	mMm	0/1/2	1389/252/12	11.203/10.520/10.174	−0.0489 (*P* = 0.0091)
S4 (*n* = 696)	MMM	0/1/2	119/356/221	9.003/8.878/8.758	Reference
	MmM	0/1/2	578/115/3	8.597/10.164/9.7	0.1775 (*P* = 0.0002)
	MMm	0/1/2	352/294/50	8.993/8.777/8.422	−0.0155 (*P* = 0.603879)
	mMM	0/1/2	620/73/3	8.918/8.536/4.967	−0.019 (*P* = 0.733164)
	mMm	0/1/2	NA	NA	NA

Results from linear regression models on log(adiponectin), adjusted for age, sex, and BMI, survey (for the S123 sample), and the other haplotypes, with MMM being the reference using an additive genetic model. Haplotypes are depicted by m and M for the minor or major allele, respectively, in SNP1, SNP2, and SNP3. 0/1/2 = number of reconstructed haplotype copies.

*For complete data for all three SNPs.

†Geometric mean of adiponectin concentrations (μg/ml) per copy of the reconstructed haplotypes.

Figure 1. Schematic overview of the used promoter constructs.

A: Schematic overview of the luciferase reporter vectors used in this study for transfections. Genomic location of the here analyzed SNP1 (rs16861194), SNP2 (rs17300539), and SNP3 (rs266729) are marked (*B*). All experimentally verified transcription factor binding sites are shown for the human (*B*) and mouse locus (*C*), and the here analyzed SNPs are all located upstream of these sites. The genomatix-predicted putative binding sites are depicted. SNP1 interferes with a putative CART binding site and SNP2 with a putative NKXH binding site (both sites for different families of homeobox proteins), and SNP3 interferes with a zinc-finger binding site.

Figure 2. *APM1* promoter activity during differentiation.

Transient transfection of 3T3-L1 cells at different stages of adipogenic differentiation with the indicated *APM1* promoter constructs is shown. A total of 1 µg of the indicated promoter construct (MMM = *APM1* promoter with the three described SNPs in the major configuration, m = minor variant, M = major variant) was transfected into 3T3-L1 cells at the indicated day of differentiation (day 0 = preadipocytes, day 6 and 8 = 6 or 8 days after induction of differentiation). A total of 0.1 µg ubiquitin-renilla vector was cotransfected for normalization of the transfection. The haplotypes observed in KORA samples are depicted separately from the theoretically existing but not observed haplotypes. Cells were harvested 24 h after transfection. Results are shown as the ratio of firefly-/renilla-luciferase activity and the mean of minimal five independent experiments ± SD. The Kruskal-Wallis overall comparison of all constructs and observed/theoretical haplotypes is indicated with*P* values; comparison of the day 0, 6, and 8 values for each construct were $P < 0.001$ for MMM, $P < 0.001$ for MmM, $P < 0.05$ for MMm, $P < 0.001$ for mMM, $P < 0.001$ for mmm, $P < 0.005$ for mmM, $P > 0.05$ for Mmm, and $P < 0.01$ for mMm. The significance of Dunn's multiple comparison test

comparing the day 0 value for each construct with its values at day 6 and day 8, respectively, is indicated with asterisks: *$P < 0.05$ and **$P < 001$.

Figure 3. Inducibility of different haplotypes by rosiglitazone.

Transient transfection of 3T3-L1 cells with the indicated adiponectin promoter constructs at day 6 after induction of differentiation is shown. A total of 1 μg of the indicated promoter construct was transfected into 3T3-L1 cells. A total of 0.1 μg ubiquitin-renilla vector was cotransfected for normalization of the transfection. The haplotypes observed in MONICA/KORA S123 or S4 survey are separately depicted from the theoretical existing, but was not observed in patients. At 24 h after transfection, cells were induced with 1 μmol/l rosiglitazone for 24 h as indicated. Cells were harvested 24 h after transfection. Results are shown as the ratio of firefly-/renilla-luciferase activity and the mean of minimal three independent experiments ± SD. Kruskal-Wallis overall comparison of all constructs and of frequent/theoretical haplotypes is indicated with a *P* value, followed by the Dunn's multiple comparison test comparing the uninduced (−) with the respective rosiglitazone-induced (+) cells for each construct, as indicated with **$P < 0.001$.

Figure 4. DNA binding activity of different SNP variant combinations. An 80-bp fragment, described in RESEARCH DESIGN AND METHODS, was radioactively labeled, incubated with 2 μg of the indicated protein extracts, and separated on a gel as described. AC, 3T3-L1 adipocyte; PAC, 3T3-L1 preadipocyte; cells were induced with 1 μmol/l rosiglitazone (+) or with DMSO control (−).

2.4.6 Figures

Figure 1

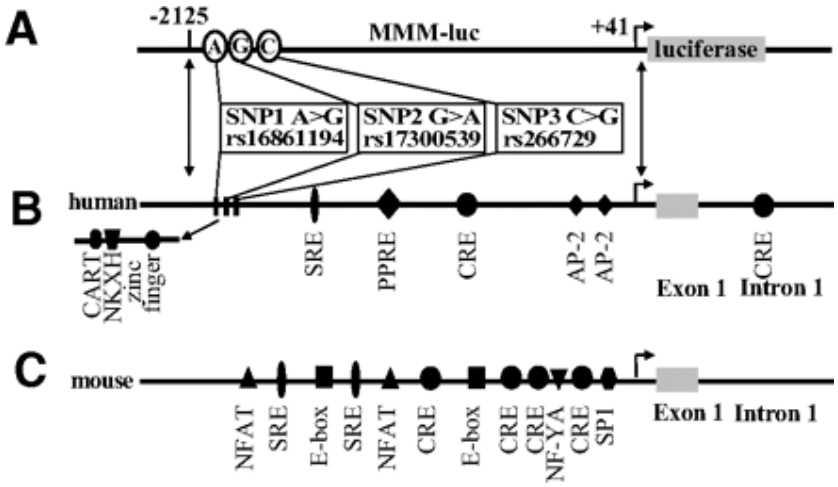


Figure 2

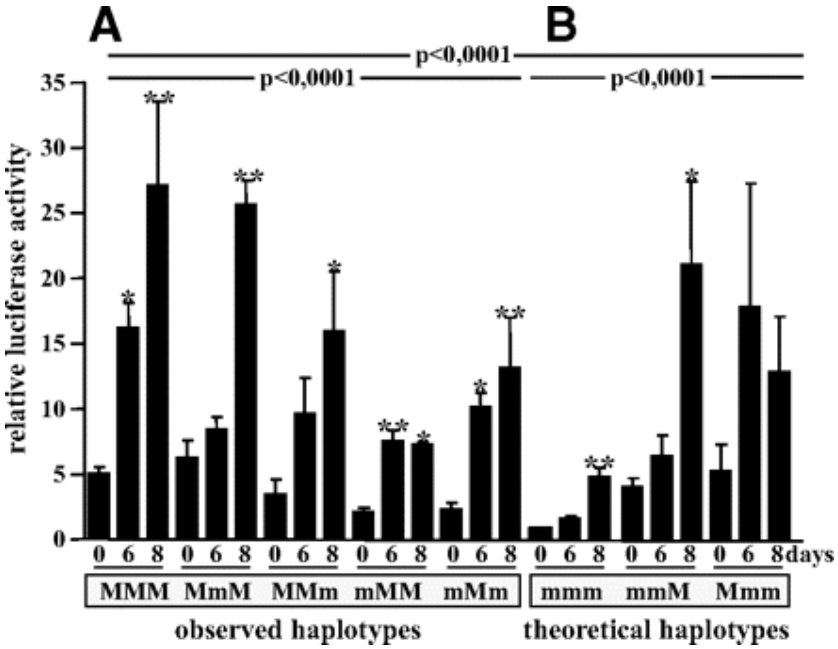


Figure 3

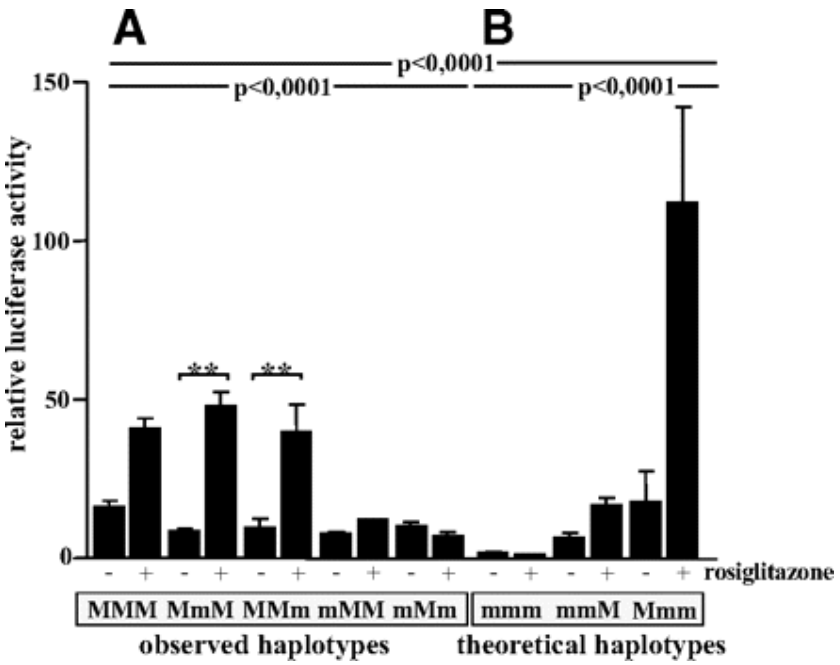
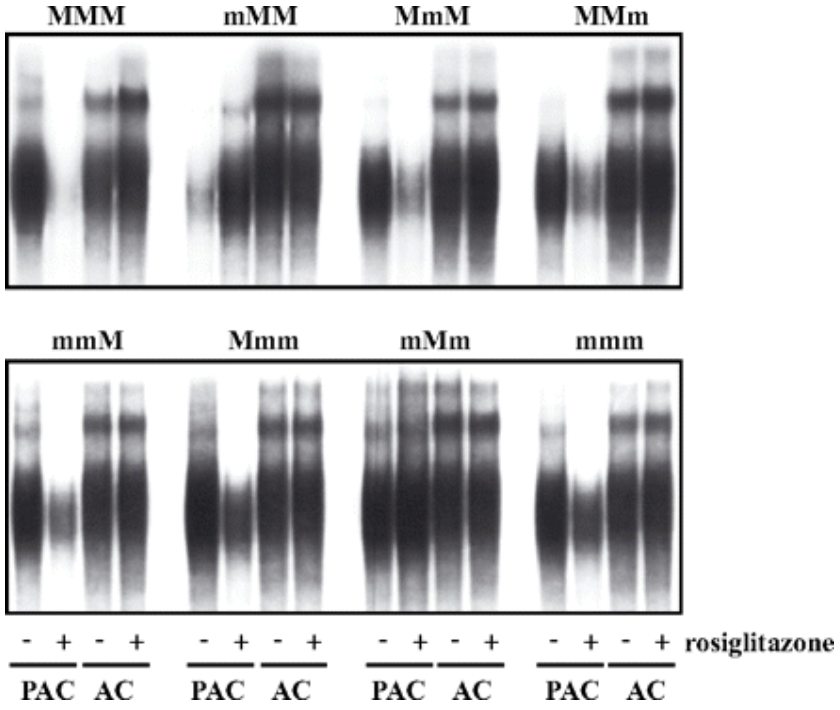

A
B
p<0,0001
p<0,0001
p<0,0001
150
100
50
0
relative luciferase activity
**
**
- + - + - + - + - + - + - + - + - + - +
rosiglitazone
MMM MmM MMm mMM mMm
mmm mmM Mmm
observed haplotypes
theoretical haplotypes

Figure 4


MMM
mMM
MmM
MMm
mmM
Mmm
mMm
mmm
- + - + - + - + - + - + - + - +
rosiglitazone
PAC AC PAC AC PAC AC PAC AC

2.5 Octamer-dependent transcription in T cells is mediated by NFAT and NF-κB.

The content of this chapter has been published in Mueller et al., 2013, and is described below. For this manuscript, I performed a computational analysis of the *BOB.1/OBF.1* promoter across a set of vertebrate species with a specific focus on putative NFAT and NF-κB transcription factor binding sites. The analysis includes parts of the PMCA computational framework. The hypothesis was that highly conserved NFAT and NF-κB within a promoter CRM might indicate a functionally relevant regulation of the BOB.1 gene via binding of NF-κB and NFAT transcription factors. NFAT transcription factors are known to harbor an imperfect Rel homology domain that is only capable of weak DNA binding in the monomeric and dimeric form. These factors therefore tend to interact with other transcription factors such as AP-1 (c-Jun/c-Fos), GATA-4, MEF-2, and NF-κB (Liu et al., 2012; Hogan et al., 2003) to strengthen the DNA interactions. Indeed, the computational analysis identified unknown cluster of combinatorial NFAT/ NF-κB binding sites within the *BOB.1/OBF.1* promoter.

2.5.1 Introduction

Regulated gene expression is a complex process since different signals need to be integrated in a cell-type-specific manner in accordance with the particular developmental stage as well as activation state. This complexity is achieved by the architecture of a given promoter and/or enhancer and therefore by the integrated action of different transcription factors in conjunction with recruited co-activators or –repressors. These proteins act together on

promoter DNA finally leading to the formation of specific transcriptional complexes based on the DNA sequence they bind as well on the activity of each component itself.

The octamer element ATGCAAAT is one of such DNA sequences and plays an important role in mediating promoter activity of a large array of ubiquitous as well as lymphocyte-specific genes. Octamer-dependent transcription is achieved in first line by transcription factors that belong to the Oct family. The selectivity of Oct factors to octamer sequences as well as their transcriptional activity can be enhanced by the recruitment of either ubiquitously expressed or cell type specific co-activators. For instance, the histone H2B promoter activity depends on Oct1 (Pou2f1) and its interaction with the transcriptional co-activator OCA-S, a protein complex containing GAPDH as a key component, whose expression is highly increased during the S phase of the cell cycle (Zheng et al., 2003). In lymphocytes, the transcriptional co-activator BOB.1/OBF.1 (Pou2af1) is responsible for the cell type specific octamer-dependent transcription. BOB.1/OBF.1 is recruited to DNA by the interaction with POU domains of the ubiquitously expressed Oct1 or the lymphocyte specific factor Oct2 (Pou2f2) (Gstaiger et al., 1996; Gstaiger et al., 1995; Pierani et al., 1990; Luo et al., 1992; Luo et al., 1995; Pfisterer et al., 1995; Strubin et al., 1995), the two Oct family members expressed in lymphocytes (Staudt et al., 1986).

However, not all octamer-regulated promoters depend on the presence of BOB.1/OBF.1 (König et al., 1995; Shore et al., 2002). The ability of Oct1 or Oct2 to recruit BOB.1/OBF.1 to the DNA might be conferred by different octamer sequences that favor or disfavor the ternary complex formation of these proteins at the octamer motif (Tomilin et al., 2000). In addition, we and others demonstrated that the presence of BOB.1/OBF.1 enables Oct factors to bind to

unfavorable non-consensus octamer motifs (Lins et al., 2003; Brunner et al., 2006). Together, the lymphocyte specific regulation of octamer-dependent transcription depends on an appropriate DNA sequence, on the activity of Oct1 and Oct2 transcription factors, as well as on the presence of the transcriptional co-activator BOB.1/OBF.1. Furthermore, the latter is posttranslationally modified by phosphorylation at Ser184 which is required for its constitutively or inducible transcriptional activity in B or T cells, respectively (Zwilling et al., 1997).

The importance of octamer-dependent transcription is underlined by the phenotypes of Oct1-, Oct2- and BOB.1/OBF.1-deficient mice. The deletion of the ubiquitously expressed Oct1 protein leads to embryonic lethality (Wang et al., 2004), and deletion of the lymphocyte specific Oct2 protein causes death of newborn mice shortly after birth (Corcoran et al., 1993). Fetal liver cell transfer into immuno-compromised mice revealed that Oct1 is dispensable for B cell development and function (Wang et al., 2004). In contrast, Oct2-deficient B cells are unable to differentiate into immunoglobulin secreting cells (Corcoran et al., 1993). This phenotype is similar to that observed for BOB.1/OBF.1-deficient mice. Although viable, these mice are unable to form germinal centers upon administration of T cell dependent antigens. Hence, the production of secondary immunoglobulins is severely compromised (Schubart et al., 1996; Nielsen et al., 1996; Kim et al., 1996). Beside missing germinal centers, BOB.1/OBF.1$^{-/-}$ mice show multiple defects at several stages of B cell development (Hess et al., 2001; Brunner et al., 2003; Samardzic et al., 2002). Although the relevance of Oct proteins and BOB.1/OBF.1 for B cell development and function cannot be dismissed, these proteins are also important for T cells. Functional octamer motifs could be detected within the promoter regions of the chemokine receptor *CCR5* (Moriuchi et al., 2001) as well as *IL2* (Shibuya et al., 1989; Brunvand et al., 1988; Pfeuffer et al., 1994) and *IL4* (Pfeuffer et al.,

1994; Chuvpilo et al., 1993; Li-Weber et al., 1998) genes. Also, the *IFNγ* promoter contains an octamer motif that is bound by Oct proteins together with BOB.1/OBF.1. As a consequence, the secretion of IFNγ by BOB.1/OBF.1-deficient TH1 cells is reduced to a level that disabled these mice to efficiently combat a *Leishmania major* infection (Brunner et al., 2007). Given the importance of the octamer-dependent transcription for B and T cell development and function it is, on the one hand, important to search for octamer-dependent target genes and, on the other, to understand the regulatory mechanisms underlying the octamer-dependent transcription itself.

Regulation of transcription is one major mechanism to determine the capacity of a given protein. The promoters of ubiquitously expressed *Oct1* gene or the lymphocyte-specific *Oct2* gene have not been described until today. In contrast, the *BOB.1/OBF.1* promoter was extensively studied in order to investigate its *cis*-acting elements controlling its activity in B cells (Stevens et al., 2000; Shen et al., 2007; Massa et al., 2003) where BOB.1/OBF.1 is constitutively expressed at all stages of B cell development (Schubart et al., 1996), albeit at different levels. The highest expression of BOB.1/OBF.1 was found in germinal center B cells. Accordingly, signals important for germinal center formation, like the stimulation with anti-CD40 antibodies plus IL4, are able to increase the expression of BOB.1/OBF.1 *in vitro* (Qin et al., 1998; Greiner et al., 2000). In contrast, in T cells BOB.1OBF.1 expression is inducible by treatment of T cells with PMA/Ionomycin (P/I) or by antigen receptor engagement (Zwilling et al., 1997; Sauter et al., 1997) suggesting that different signals and possibly also transcription factors might be responsible for the expression of BOB.1/OBF.1 in B versus T cells. Interestingly, also the expression of the lymphocyte specific Oct2 protein is upregulated in activated T cells with a kinetic similar to that obtained for BOB.1/OBF.1 (Kang et al., 1992; Bhargava et al., 1993). Since the expression and function of Oct proteins

together with BOB.1/OBF.1 was also found to be important for the control of TH1 and TH2 immune responses (Brunner et al., 2007) the present study focuses on the regulation of octamer-dependent transcription in T cells. Our analyses revealed the involvement of the Ca^{2+}/calmodulin-dependent phosphatase calcineurin (CN) since Oct2 and BOB.1/OBF.1 expression are efficiently suppressed in the presence of the immunosuppressant cyclosporin A (CsA) or by the siRNA-mediated suppression of CN-A. Upon T cell receptor (TCR) stimulation CN turned out to be a key signaling molecule essential for the induction of both NFAT and NF-κB transcription factors (Clipstone et al., 1992; Palkowitsch et al., 2011; Frischbutter et al., 2011). CN dephosphorylates NFAT transcription factors enabling them to translocate into the nucleus where they control the expression of numerous genes essential for T cell activation and differentiation (Serfling et al., 2004). In addition, CN is involved in the regulation of the TCR-induced NF-κB signaling pathway by regulating the complex assembly of Carma1, Bcl10, and Malt1 (CBM) by dephosphorylating Bcl10. CBM formation is an essential prerequisite for the activation of the IκB-kinase complex that in turn phosphorylates the inhibitor of NF-κB (IκB) leading to its proteasomal degradation. Consequently, cytosolic NF-κB becomes released and translocates into the nucleus where it regulates, similar to NFAT, the promoter activity of numerous genes crucial for T cell development, differentiation and function. Therefore, both signaling pathways were analyzed for their capacity to mediate Oct and/or BOB.1/OBF.1 gene expression and were identified as important regulators of octamer-dependent transcription in T cells.

2.5.2 Materials and Methods

Cell lines and culture

The Jurkat-4 x Octamer-Luc cell line was generated by electroporation of Jurkat cells with a luciferase reporter construct bearing 4 copies of the octamer motif cloned earlier (Annweiler et al., 1993) together with a plasmid expressing the Puromycin resistant gene (pSV-Puro) and selected by Puromycin. Jurkat-NEMO$^{-/-}$ was described (Harhaj et al., 2000). Jurkat cells and derivates (Jurkat-4 x Octamer-Luc, Jurkat-NEMO$^{-/-}$) and Φ-NX amphotropic retrovirus producer cells were cultured in DMEM (Gibco, Invitrogen), containing 10% FCS (Biochrome), 2 mM L-glutamine, 100 U/ml penicillin and 100 µg/ml streptomycin (Gibco, Invitrogen), and 50 µM α-mercaptoethanol at 37°C and 5% CO_2. A3.01 T cells and Namalwa B cells as well as primary CD4$^+$ cells were cultured in RPMI containing 10% FCS, 2 mM L-glutamine, penicillin/streptomycin and 50 µM α-mercaptoethanol at 37°C and 5% CO_2.

CD4$^+$ T cell isolation

Primary CD4$^+$ T cells were isolated from lymph nodes of mice either by positive or negative selection using magnetic microbead technology (Miltenyi).

Mice

C57BL/6 wild type and TNFRI/p65 double-deficient mice (and all their wildtype/heterozygous combinations) were generated and obtained from our breeding facility. The *NFATc1/c2* double knockout mice were generated by crossing the *Nfatc2*$^{-/-}$ mouse (Schuh et al., 1998) with the *Nfatc1*$^{flx/flx}$ (described in (Bhattacharyya et al., 2011) x *Cd4-Cre* mouse. Mice were analyzed 3-4 or 10-12 weeks after birth, respectively.

Cell stimulation

Cells were stimulated in the presence of the following agents: 100 ng/ml Cyclosporin A (CsA) and 200 ng/ml FK506 (calcineurin inhibitors), 1, 2 or 4 μM Bay-117082 (a NF-κB inhibitor) as indicated, 25 ng/ml PMA, 500 ng/ml Ionomycin, 200 nM OH-tamoxifen (OHT) (all obtained from Sigma-Aldrich), 15 or 30 μM SB203580 (Upstate) as indicated, 4 μg/ml αCD3 and 0.5 μg/ml αCD28 (BD). The NEMO binding peptide and the NFAT inhibitory peptide 11R-VIVIT were obtained from Calbiochem.

Promoter sequence analyses

In silico search for binding sites in BOB.1/OBF.1 promoters of different species was performed using the MatInspector tool (Matrix Family Library Version 8.3) of the Genomatrix software tool (Genomatrix Company).

Electrophoretic mobility shift assays (EMSA)

Preparation of whole cell extracts for EMSA and the protocol of the EMSA procedure have been described earlier (Brunner et al., 2007). The used oligonucleotides bearing the appropriate transcription factor binding site were annealed and subsequently labeled using ^{32}P-αdCTP in a fill in reaction. Sequences are presented in the supplementary table 1.

Promoter cloning

The 1500 and 500 bp BOB.1/OBF.1 promotor constructs were cloned into the pTKL/2 vector containing the HSV-thymidine kinase promoter (-105 to +52) from the pBLCAT2 in front of

the firefly luciferase coding region. The HSV-thymidine kinase promoter was excised by a restriction endonuclease digest using HindIII and BglII and replaced by the 1500 bp, or 500 bp BOB.1/OBF.1 promoter constructs cloned via genomic PCR using the following primers: mBOB.1/OBF.1prom 1500 bp 5' (HindIII): GCC AGG AAG CTT AGG GGT TGA G; mBOB.1/OBF.1prom 1500 bp 3' (BglII): GCC TTT TCT CTT TGA AGC AGA GAT CTT GGC TTC TTT ACT. For amplification of the 500 bp promoter fragment the same 3' primer as for the 1500 bp fragment was used in combination with the following primer: mBOB.1/OBF.1prom 500 bp 5' (HindIII): GAC CAA TGG TAA GCT TAG TCC TGC. Cloning of the BOB.1/OBF.1 promoter mBOB.1/OBF.1prom Δ500 bp construct was achieved using the following combination of primers for genomic PCR: mBOB.1/OBF.1prom 1500 bp 5' (HindIII) as indicated above together with mBOB.1/OBF.1prom Δ500 bp 3' (BglII): CCA TTT ACA GGA CAG ATC TTA CCA TTG GTC.

The BOB.1/OBF.1 promoter mBOB.1/OBF.1prom Δ500 bp + TATA construct was generated by cloning of an insert that was amplified by genomic PCR using the following primers: mBOB.1/OBF.1prom 1500 bp 5' (HindIII): GCC AGG AAG CTT AGG GGT TGA G and mBOB.1/OBF.1prom Δ500 bp 3' (HindIII): GCA GGA CTA AGC TTA CCA TTG GTC, into the HindIII site of the pTATA vector harboring the TATA box of the HSV-thymidine kinase promoter from pBLCAT2 (-38 to +52) in front of the firefly luciferase coding region.

In vitro Mutagenesis

Mutations of the predicted NFAT/NF-κB sites within the BOB.1/OBF.1 promoter were generated using the QuickChange™ Mutagenesis kit (Promega). Primers were ordered from Biomers and sequences are provided in supplementary table 4.

Transfection of cells

Transfections of Jurkat T and Namalwa B cells were performed by electroporation (Bio-Rad) with 450 V and 250 µF in PBS. The expression vector for the constitutive active version of CREB (c2CREB) is a kind gift of G. Tiel (Al Sarraj et al., 2005). Expression vectors for NFATc1 and RelA/p65 have been described (Kempe et al., 2005; Chuvpilo et al., 2002). The pRL-CMV plasmid (Renilla Luciferase control reporter vector; Promega) was cotransfected in all experiments and used for normalization of different transfection efficiencies in the individual experiments. For suppression of calcineurin A isoforms by siRNA Jurkat T cells were transfected with 75 nM of each SMARTpool siRNA (CnBa: PPP3CA # L008300-0005; CnBb: PPP3CB # L009704-0005) or with the appropriate concentration of control siRNA (OnTARGETplus # D001810-01-05) using the Nucleofection Kit V (Amaxa/Lonza). The cells were subsequently incubated for 72 h prior to analysis. Used SMARTpool siRNA were obtained from Dharmacon.

Retroviral infection of cells

For virus production the retroviral vectors expressing NF-ATc1/□A-ER, NF-ATc1/□C-ER (Nayak et al., 2009), IKK2-EE, IKK2-KD (Denk et al., 2001) or the respective empty vectors were transfected into the amphotropic ϕNX retrovirus producer cells by use of the calcium phosphate method. Supernatant containing the retrovirus was collected 24 h and 48 h after transfection and used to infect A3.01 T cells in the presence of 8 µg/ml polybrene. At two consecutive days spin infection was performed at 2700 rpm (1300 x g) and 37°C for 120 min. Positive cells were selected with Zeocin until all cells were nearly 100% positive for green fluorescent protein.

Western blots

For Western blot analysis, 5 to 10 µg of total protein extracts were separated on 12.5% polyacrylamid gels and transferred onto nitrocellulose membranes (Schleicher&Schuell). Membranes were blocked (TBS, 0.1% Tween, 5% milk), stained with anti-BOB.1/OBF.1 (Sigma), Oct2, ERK2, IKK2 (Santa Cruz Biotechnology) and NFATc1 (Alexis) antibodies followed by incubation with HRP-coupled secondary antibodies (Pierce), and visualized by enhanced chemiluminescence (Pierce).

RT-PCR

Total RNA was isolated using Trizol reagent (Qiagen) and reverse transcribed using MMLV reverse transcriptase (Roche). Quantitative PCR were performed using qPCR-SYBR-Green (Roche). Primers ordered from Biomers were designed using "Universal Probe Library Assay Design Center" Software (Roche). Primer sequences are given in supplementary table 2. Quantification of gene regulation was performed by the ΔΔCp method using *RPL13* as house-keeping gene.

Chromatin Immunoprecipitation (ChIP)

ChIP experiments were performed using the ChIP-IT® Express Chromatin Immunoprecipitation Kit as well as the Re-ChIP-IT® magnetic chromatin re-immunoprecipitation kit from Active Motif according to the manufacturer's protocol with slight modifications: after enzymatic digestion for 10 min, the chromatin was sheared by 2 cycles of sonication (10 pulses each cycle) in the same buffer. The chromatin was pre-cleared

for 2 h with protein G microbeads (Invitrogen) and then incubated with rabbit polyclonal anti-p65 antibody sc-372 (2 µg/ml; Santa Cruz Biotechnology), mouse monoclonal anti-NFATc1 antibody 7A6 (4 µg/ml; Alexis), mouse IgG (Dianova) or normal rabbit serum (Pierce) at the appropriate concentrations for 4 h followed by incubation with protein G microbeads for 1 h. The amount of precipitated DNA was evaluated by quantitative PCR using the Roche Light Cycler LC480. The used primers and primer positions are depicted in Figure 6 and presented in supplementary table 3. The relative amount of precipitated DNA was calculated using the following formula: $E^{(\text{crossing point 1/10 total input} - \text{crossing point sample})}$ and is depicted as amount of precipitated genomic DNA relative to that precipitated by control antibodies (mouse IgG or normal rabbit serum). E = efficiency of the PCR determined by serial dilutions of total input.

2.5.3 Results

NFAT and NF-κB inhibitors restrain the inducible expression of BOB.1/OBF.1 and Oct2 in T cells

To elicit the signaling pathways controlling the octamer-dependent transcription in T cells, human Jurkat T cells were stably transfected with a luciferase reporter construct harboring four consecutive sequences bearing the octamer binding motif ATGCAAAT in front of a luciferase gene. As expected, the octamer activity was strongly upregulated upon treatment with P/I but remained unaffected by stimulation with either P or I alone suggesting the requirement of a combined calcium influx and PKC activation for octamer-dependent transcription in T cells (Figure 1A). Several transcription factors are known to depend on such a combinatorial calcium and PKC signaling, including NF-κB and NFAT. To further specify the signaling pathways controlling octamer dependent transcription, we used a panel of specific pharmacological inhibitors. Indeed, complete suppression of the inducible octamer activity was seen when the cells were pretreated with the CN inhibitor CsA. Also the NF-κB inhibitor Bay11-7082 as well as the p38 inhibitor SB203580 interfered with the inducible octamer activity, although less efficiently. These data suggest that Ca^{2+}/CN, NF-κB and MAP kinase dependent signaling pathways are involved in the regulation of octamer-dependent transcriptional activity (Figure 1A).

We next wondered whether the same pharmacological inhibitors could influence the activity of the known *BOB.1/OBF.1* promoter. Therefore, a reporter construct harboring the previously described 1500 bp *BOB.1/OBF.1* promoter (Stevens et al., 2000) was transiently transfected into Jurkat T cells which were either left untreated or pretreated with inhibitors for CN, NF-κB or p38 for 30 min prior to the stimulation with P/I. These experiments revealed

that the *BOB.1/OBF.1* promoter is sensitive to all of the analyzed inducers and inhibitors, very similar to the octamer-dependent reporter (Figure 1B). To explore whether protein expression of BOB.1/OBF.1 mirrors its promoter activity, Jurkat T cells as well as primary $CD4^+$ cells isolated from lymph nodes of C57BL/6 wildtype mice were either left untreated or pretreated for 30 min with the inhibitors used in the previous experiments and subsequently stimulated with P/I. Interestingly, we found a striking co-regulation of Oct2 and BOB.1/OBF.1 in Jurkat (Figure 1 C and D) and primary $CD4^+$ T cells (Figure 1 E and F). Thus, the same inducers as well as inhibitors control the expression of Oct2 and BOB.1/OBF.1 at protein and mRNA levels. However, while the BOB.1/OBF.1 expression was completely abolished by treatment of Jurkat or $CD4^+$ T cells with CsA even at low concentrations (50 to 100 ng/ml), low levels of Oct2 expression were still detectable (Figure 1C and E). In contrast, the NF-κB inhibitor Bay11-7082 had only a moderate influence on BOB.1/OBF.1 as well as Oct2 expression in Jurkat T cells even at high concentrations (2 to 4 μM), whereas in primary $CD4^+$ T cells the inhibition of the NF-κB pathway led to an almost complete abrogation of Oct2 expression and to a marked reduction in BOB.1/OBF.1 expression at very low concentrations (1 μM) of the inhibitor. The insufficient block of BOB.1/OBF.1 and Oct2 expression by Bay11-7082 might be due to a residual NF-κB activity, as seen in EMSA experiments (Supplementary Figure 1). However, higher concentrations of the inhibitor Bay11-7082 were toxic for primary T cells. Also, pretreatment of primary $CD4^+$ T cells with CN inhibitors CsA or FK506 or by the NF-κB inhibitor Bay11-7082 completely inhibited the strong BOB.1/OBF.1 as well as Oct2 induction seen after the more physiological stimulation with anti-CD3 and anti-CD28 antibodies (Figure 1 G and H). The p38 inhibitor SB-203580 had a moderate effect on BOB.1/OBF.1 and Oct2 expression only at higher concentrations (Figure 1C) suggesting a rather minor and/or indirect effect of p38 on the regulation of octamer-dependent transcription in T cells. Interestingly, EMSA studies (Supplementary Figure 1) revealed that the NF-κB

inhibitor Bay11-7082 inhibits additionally NFAT binding to DNA and *vice versa* the CN inhibitors CsA and FK506 interfere with NFAT binding but also with the NF-κB signaling pathway, as it was described recently (Palkowitsch et al., 2011; Frischbutter et al., 2011). Therefore, the observed effect of CsA and FK506 on BOB.1/OBF.1 expression is not only mediated by inhibiting NFAT activity, since these compounds also affect the activity of NF-κB, but rather by the combined activity of NFAT and NF-κB. However, treatment of murine primary $CD4^+$ T cells or transfected Jurkat cells with specific NFAT and NF-κB signaling pathway inhibitors, like the NFAT inhibitory peptide 11R-VIVIT and the NEMO binding peptide, respectively, clearly demonstrates that both pathways are involved in the regulation of BOB.1/OBF.1 expression via the regulation of BOB.1/OBF.1 promoter activity (Figure 1I and J).

The *BOB.1/OBF.1* promoter contains several binding sites for NFAT and NF-κB

Previously, the *BOB.1/OBF.1* promoter was analyzed in order to identify regulatory elements responsible for its activity in B cells (Stevens et al., 2000; Shen et al., 2007; Massa et al., 2003). In those studies a functional CREB/ATF site could be identified which is crucial for the B lymphocyte specific activity of the promoter. In addition, a sequence related to a NFAT binding site was described, however without any functional relevance for the B cell-specific *BOB.1/OBF.1* promoter activity. In search for possible additional transcription factor binding sites using the Genomatrix Software we identified two combined NFAT/NF-κB binding sites within the *BOB.1/OBF.1* promoter. Both newly identified sites are conserved between rat, mouse and man suggesting a possible functional relevance of these regulatory elements. In contrast to B cells, where a function of the predicted NFAT binding site has not been revealed (Stevens et al., 2000), an inducible complex formation could be detected after P/I treatment of primary $CD4^+$ T cells. This complex resembles that formed at the NFAT site of the *IL2*

promoter used in control experiments (Figure 2B, lane 2 and C, lane 2). In addition, also an inducible complex formation could be observed using the newly identified combined composite and consecutive NFAT/NF-κB sites of the *BOB.1/OBF.1* promoter as probes (Figure 2D and E, lane 2). In supershift experiments antibodies against NF-κB p50 were able to compete against the observed inducible complexes formed at the predicted NFAT/NF-κB binding sites within the *BOB.1/OBF.1* promoter (Figure D, E lane 3) in a similar way like anti-p50 antibodies prevent p50 binding to a consensus NF-κB site (Figure 2F, lane 3). Interestingly, anti-p50 antibodies are also able to interfere with the inducible complex formed at the predicted NFAT site of the *BOB.1/OBF.1* promoter (Figure 2B, lane 3) as well as with the NFAT-binding observed at the *IL2* promoter site (Figure 2C, lane 3). Similarly, anti-p65 antibodies interfere with binding to the consecutive and composite NFAT/NF-κB sites, which is comparable with the competition seen at the consensus NF-κB site (Figure 2D, E and F, lane 4). Additionally, anti-p65 antibodies also interfere with the complex formation on the NFAT site of the *BOB.1/OBF.1* and *IL2* promoters (Figure 2B and C, lane 4). The use of NFATc1 antibodies reduces the binding to the NFAT and NFAT/NF-κB sites of the *BOB.1/OBF.1* promoter and generates supershifted complexes (marked by clamp in Figure 2B, D and E, lane 5). Also, NFATc2 antibodies prevent complex formation to these specific sites (Figure 2B, D and E, lane 6). Control experiments using IgG antibodies or normal rabbit serum revealed the specificity of antibody binding (Supplementary Figure 2). Moreover, unlabeled ologonucleotides bearing the consensus NFAT or NF-κB binding site could efficiently compete against labeled NFAT or NFAT/NF-κB sites identified in the *BOB.1/OBF.1* promoter (Supplementary Figure 2). Together, our data suggest that NFAT as well as NF-κB family members are able to bind to all of these newly identified potential *cis*-acting elements of the BOB.1/OBF.1 promoter *in vitro*.

Newly identified NFAT/NF-κB binding sites contribute to BOB.1/OBF.1 promoter activity

To demonstrate the importance of the potential NFAT or NFAT/NF-κB binding site for the BOB.1/OBF.1 promoter activity mutation analyses were performed. In these experiments specific point mutation were introduced to prevent either NFAT or NF-κB binding to these sites (Figure 3). Point mutation (G→T; Figure 3A) within the previously identified NFAT motif within the *BOB.1/OBF.1* promoter did not lead to alterations in complex binding to this site. However, introduction of a point mutation (G→T) within the second core motif for NFAT and NF-κB clearly prevents inducible by P/I complex formation indicating the importance of this residue within this potential transcription factor binding site (Figure 3B). Within the consecutive NFAT/NF-κB binding site of the *BOB.1/OBF.1* promoter both mutations, preventing NF-κB (GG→TT) as well as NFAT binding (CC→AA), abolish inducible complex formation upon treatment of T cells with P/I (Figure 3C). Therefore, both site possibly contribute to the inducible *BOB.1/OBF.1* promoter activity in T cells. In contrast, whereas mutations of critical residues within the NF-κB motif (GG→TT) of the composite NFAT/NF-κB site within the *BOB.1/OBF.1* promoter lead to a clear reduction of complex binding upon P/I treatment of $CD4^+$ T cells, mutations within the potential NFAT binding site (CC→TT) slightly enhance the ability of complex binding to this site (Figure 3D). For comparison, the mutated consensus NFAT and NF-κB sites were also analyzed (Figure 3E).

Additionally, the introduction of mutations into the 1500 bp *BOB.1/OBF.1* promoter construct that was cloned in front of a luciferase gene and transfected into Jurkat T cells (Figure 3F) revealed that all three newly identified NFAT and NFAT/NF-κB site are important for full *BOB.1/OBF.1* promoter activity as mutation of each reduces *BOB.1/OBF.1* promoter activity by about 15 to 20 %. Mutation of all three NFAT/NF-κB sites led to a reduction of

BOB.1/OBF.1 promoter activity by 40 %. Interestingly, additional mutation of the previously identified CREB/ATF site, important for BOB.1/OBF.1 promoter activity in B cells, almost completely abrogates BOB.1/OBF.1 promoter activity in T cells.

A 500 bp sequence is necessary and sufficient for the highest inducible *BOB.1/OBF.1* promoter activity in Jurkat T cells

For the full activity of the *BOB.1/OBF.1* promoter in B cells a sequence spanning 1500 bp was identified (Stevens et al., 2000). To define regions important for the inducible activity in T cells, additionally a shorter promoter construct was generated encompassing approximately 500 bp upstream of the start site of transcription of the *BOB.1/OBF.1* gene (Figure 4A). To compare the behavior of these constructs with respect to the basal as well as inducible BOB.1/OBF.1 promoter activity in B and T cells, the constructs were transfected into Namalwa B or Jurkat T cells. In accordance with previous findings (Stevens et al., 2000) the 1500 bp promoter showed the highest basal activity in Namalwa B cells that could be further increased (3-fold) by treatment of cells with P/I (Figure 4B and C). In B cells, the activity of 500 bp construct was approximately 3-times lower when compared to the 1500 bp fragment with respect to basal and inducible promoter activity. As expected, in T cells no basal *BOB.1/OBF.1* promoter activity could be observed. However, it was strongly induced when cells were treated with P/I (Figure 4C). Notably, in T cells the 500 bp promoter construct showed the highest inducibility after P/I treatment (Figure 4B and C). The importance of the 500 bp sequence of the *BOB.1/OBF.1* promoter was further underscored by deleting this promoter element which led to the complete loss of basal and inducible *BOB.1/OBF.1* promoter activity in B and T cells (Figure 4B and C). This loss of activity was not caused by the deletion of regulatory elements necessary for the recruitment of the basal transcriptional complex, since the fusion of the same promoter construct to a TATA-box element of the

thymidine kinase promoter did also not lead to enhanced BOB.1/OBF.1 promoter activity in T cells (Supplementary Figure 3).

In order to investigate the contribution of NFAT and NF-κB to the *BOB.1/OBF.1* promoter activity in T cells, Jurkat cells were transfected with the 500 bp or 1500 bp *BOB.1/OBF.1* promoter constructs, together with expression vectors for NFATc1 or RelA/p65. Whereas overexpression of NFATc1 had only a moderate effect, overexpression of RelA/p65 already significantly enhanced the inducible activities of the 1500 and 500 bp promoters. Co-expression of both factors together leads to a further increase of BOB.1/OBF.1 promoter activity, although this effect was modest (Figure 4D). Again, the 500 bp promoter construct was sufficient to achieve full *BOB.1/OBF.1* promoter activity induced by NFAT/NF-κB overexpression.

In B cells, the CREB/ATF site was found to be important for BOB.1/OBF.1 promoter activity. In order to investigate whether this site is also important for the inducible expression of BOB.1/OBF.1 in T cells, the 1500 bp promoter construct was transfected into Jurkat T cells together with expression vectors coding for NFATc1, RelA/p65 or a constitutive active version of CREB either alone or in different combinations of these vectors. Also in T cells, the CREB/ATF site seems to be of relevance since overexpression of a constitutive active CREB protein leads to a significant increase of *BOB.1/OBF.1* promoter activity that was further enhanced by overexpression of NFAT and NF-κB (Supplementary Figure 4). These data together with our data obtained from EMSA and transfection experiments in that the CREB site was inactivated indicate that this CREB site is also of considerable importance for *BOB.1/OBF.1* promoter activity in T cells.

Modulation of NFAT activity interferes with the expression levels of BOB.1/OBF.1 and Oct2

One major regulator of NFAT activity is the Ca^{2+}/calmodulin-dependent protein phosphatase CN that dephosphorylates cytosolic NFAT proteins. NFAT, in turn, translocates into the nucleus where it binds to DNA and regulates gene expression, a process that can be efficiently inhibited by CsA. Down modulation of CN by Amaxa transfection of siRNA directed against the isoforms α and β of the catalytic subunit CN A revealed a clear dependence of BOB.1/OBF.1 and Oct2 expression on CN activity (Figure 5A). Obviously, Amaxa transfection of the indicated siRNA itself pre-activated Jurkat T cells, since BOB.1/OBF.1 as well Oct2 expression could be detected even in the unstimulated state. However, the expression of both proteins could be further enhanced by P/I treatment. Although the down regulation of CN expression was incomplete, a clear inhibition of inducible BOB.1/OBF.1 and Oct2 expression could be observed indicating the importance of CN for BOB.1/OBF.1 and Oct2 expression in T cells.

In peripheral T cells, mainly two members of the NFAT family are expressed, NFATc1 and NFATc2. Upon T cell activation, the NFATc1 isoform NFATc1/ αA is predominantly expressed (Serfling et al., 2006). To study the influence of NFAT on BOB.1/OBF.1 or Oct2 expression, human A3.01 T cells were retrovirally infected using vectors expressing either the NF1ATc1/□A or NFATc1/□C isoform that were fused to a part of the modified hormone binding domain of the estrogen receptor □ (ER□) (54) which is controlled by OH-Tamoxifen (OHT). Whereas the OHT treatment of cells infected with the empty vector left BOB.1/OBF.1 and Oct2 expression unaffected, a significant increase in their protein expression level could be observed when NFATc1/ αA or NFATc1/ αC became functional active in the presence of OHT (Figure 5B).

To test whether the absence of endogenous NFAT protein expression would influence

BOB.1/OBF.1 as well as Oct2 expression, $CD4^+$ T cells were isolated from *Nfatc2*$^{-/-}$ x*Nfatc1*$^{flx/flx}$*xCD4-Cre* mice in which all T cells are devoid of the expression of NFATc1 as well as NFATc2. After stimulation of purified $CD4^+$ T cells with either P/I or αCD3+ αCD28 for 18 h the mRNA expression of *BOB.1/OBF.1* and *Oct2* was analyzed by quantitative PCR. The mRNA levels of *BOB.1/OBF.1* as well as *Oct2* were sensitive to the combined ablation of NFATc1 and NFATc2 protein expression (Figure 5C and D), indicating the importance of these factors for octamer-dependent transcription in T cells.

Modulation of NF-κB activity affects the expression levels of BOB.1/OBF.1 and Oct2

Next, we investigated the influence of the NF-κB activity on BOB.1/OBF.1 or Oct2 expression. First, NEMO-deficient or wildtype Jurkat cells were stimulated with P/I. Western blot analysis revealed a clear dependence of BOB.1/OBF.1 expression on the presence of NEMO in T cells (Figure 6A). NEMO is essential for the canonical NF-κB signaling since its deficiency leads to a specific block of IKK activity (Harhaj et al., 2000). Therefore, these findings clearly indicate a contribution of NF-κB to the regulation of *BOB.1/OBF.1* gene transcription.

In another approach, human A3.01 T cells were retrovirally transduced using vectors expressing either a constitutive active version (EE) or a kinase dead mutant (KD) of IKK2 (Denk et al., 2001) (Figure 6B). In EMSA, the modulation of NF-κB activity was monitored. Overexpression of a kinase dead version of IKK2 led to a clear reduction in the inducible NF-κB binding activity, whereas overexpressing a constitutive active IKK2 led to an enhanced basal as well as inducible NF-κB activity. In EMSA, a clear correlation between NF-κB and Oct2 binding activity to the appropriate consensus DNA binding sites could be observed. Again, western blot analyses revealed a clear dependence of the expression level of BOB.1/OBF.1 and Oct2 on the level of NF-κB activity.

To analyze the effect of defective NF-κB activity on BOB.1/OBF.1 and Oct2 expression *in vivo*, we have used RelA/p65-deficient mice that were bred into the TNFR type 1 (TNFRI)-deficient background to overcome the TNF-α -induced liver toxicity and embryonic lethality (Alcamo et al., 2001) observed in RelA/p65$^{-/-}$ mice (58). CD4^{+} T cells were isolated from lymph nodes of mice deficient for TNFRI and RelA/p65 (double knock out = DKO) or from mice that were wildtype or heterozygous for one or the other or both genes and stimulated with P/I for 18 h. When the expression of RelA/p65 was abolished, the expression of BOB.1/OBF.1 was almost completely restrained, whereas Oct2 protein is still detectable although the level of expression was severely reduced (Figure 6C). Together, these data indicate an essential role of NF-κB signaling for the induction of Oct2 and BOB.1/OBF.1 expression in T cells.

To address the question if the loss or gain of NF-κB or NFAT transcription factors leads to changes in expression and/or binding activity of the other factor we have used the advantage of CD4^{+} T cells obtained from mice bearing genetic mutations either for NFATc1/NFATc2 or for TNFRI/p65. Additionally, we overexpressed NFATc1A and RelA/p65 in Jurkat T cells (Supplementary Figure 5). Interestingly, these experiments revealed that the deletion or overexpression of NFAT does not significantly lead to changes neither in NF-κB p65 expression as seen in western blot analyses nor in NF-κB binding activity analysed in EMSA experiments using a consensus NF-κB binding site (Supplementary Figure 5A and C). However, deletion of RelA/p65 in primary CD4^{+} T cells leads to a clear reduction of NFATc1 expression and consequently to a considerable reduction in NFAT binding activity to the DNA motif of the *IL2* promoter. Vice versa, overexpression of RelA/p65 in Jurkat cells clearly enhances NFAT binding activity (Supplementary Figure 5B and C). These data indicate that the observed effect on BOB.1/OBF.1 and Oct2 expression in TNFRI/p65 double

deficient murine $CD4^+$ T cells and in human Jurkat or A3.01 T cells expressing various mutants interfering with NF-κB signalling (Figure 6) could be mediated, at least in part, also by changes in NFAT activity.

Analyses of contribution of NFAT and NF-κB to the *BOB.1/OBF.1* promoter activity *in vivo*

In order to clarify whether the observed NFAT and NF-κB effects on BOB.1/OBF.1 and Oct2 expression are direct or rather indirect, primary $CD4^+$ T cells were used in chromatin immunoprecipitation (ChIP) experiments in which a fragment encompassing 189 bp from the first 500 bp of the *BOB.1/OBF.1* promoter containing the already described NFAT binding site was analyzed. Additionally, a second fragment of 161 bp was analyzed containing the combined NFAT/NF-κB sites described here. The locations of analyzed binding sites as well as of primers used for amplification of genomic DNA are depicted in Figure 6A. After induction with P/I, both fragments of the *BOB.1/OBF.1* promoter could be efficiently enriched by NFATc1 and RelA/p65 antibodies (Figure 6B and C), indicating that NFAT and NF-κB transcription factors exert a direct effect on the *BOB.1/OBF.1* promoter.

2.5.4 Discussion

Stimulation of the TCR triggers the activation of different signaling cascades leading finally to the activation of NFAT, NF-κB and AP-1 which act together with several other transcription factors to orchestrate gene expression important for T cell function, survival, proliferation and differentiation. *Oct2* and *BOB.1/OBF.1* are among these induced genes in activated T helper cells. Due to their ability to act in concert at the octamer motif identified within the promoter region of the *IFNγ* gene (Brunner et al., 2007) BOB.1/OBF.1 and Oct2 are important for the balanced TH1 and TH2 cytokine secretion by T helper cells. While investigating signaling pathways, which regulate octamer-dependent transcription we found that the expression of BOB.1/OBF.1 as well as Oct2 critically depends on the expression and function of CN. Down-modulation or inhibition of the Ca^{2+}/calmodulin dependent serine/threonine-specific protein phosphatase using either siRNA or CsA attenuated the expression of both proteins. The fact that CsA affects NFAT as well as NF-κB activity (Clipstone et al., 1992; Palkowitsch et al., 2011; Frischbutter et al., 2011) prompted us to investigate the involvement of both signaling cascades in the regulation of BOB.1/OBF.1 and Oct2 expression in T cells. Using NEMO as well as NFAT inhibitory peptides, blocking the NF-κB and the NFAT pathway in a specific manner, revealed that both transcriptional activities are involved that process.

Indeed, we found several transcription factor binding sites within the *BOB.1/OBF.1* promoter region conserved between mouse and man that can be bound upon stimulation by both NFAT and NF-κB family members *in vitro* as well as *in vivo*. The most distally located "composite" NFAT/NF-κB site (see Figure 2A) contains the classical NFAT core motif TTTCC that is also found in the distal promoter of the *IL2* gene as well as an adjacent overlapping NF-κB binding motif AGGTGT that was recently described as an active NF-κB site within the human *prostatic acid phosphatase* gene promoter (Zelivianski et al., 2004). Composite NFAT/NF-κB

binding sites are described for numerous of promoters. For those identified within the *HIV LTR* and the human *IL8* promoter binding of NFATc2 as a homodimer has been detected (Giffin et al., 2003; Jin et al., 2003). Others have shown that NFATc2 competes for binding with NF-κB to the composite NFAT/NF-κB site within the *HIV LTR* (Macian et al., 1999) indicating that either NFAT or NF-κB factors bind to this site. Indeed, there is no experimental evidence for a heterodimer formation between NFAT and NF-κB family members. Although in supershift experiments antibodies against NF-κB and NFAT are able to prevent inducible complex formation on this composite NFAT/NF-κB site, introduction of specific mutations revealed that rather NF-κB than NFAT transcription factors are bound to this site.

The second combined, "consecutive" NFAT/NF-κB site also contains the classical NFAT core motif TTTCC that is followed at the 5' site by the sequence 5'-GGGAATCGCA-3'. The latter represents the canonical NF-κB decamer from position 1 to 6 and bears also a cytidine at position 9. At such consecutive sites, NFAT and NF-κB complexes can be formed simultaneously, not competing for binding among each other. However, in Re-ChIP experiments we were not able to show the mutual binding of NFATc1 together with RelA/p65 (data not shown) to this site. Yet, both the NF-κB and the NFAT motif are important for transcription factor binding as individual mutation of both motifs prevents inducible by P/I complex formation.

The third important site for the regulation of *BOB.1/OBF.1* promoter activity is the already described non-consensus NFAT binding site GAAGAAA that has no functional relevance for the *BOB.1/OBF.1* promoter activity in B cells (Stevens et al., 2000). In contrast, in T cells an inducible complex formation could be detected to this site *in vitro*. Moreover, a 500 bp *BOB.1/OBF.1* promoter construct bearing this site was sufficient for full inducible promoter activity in T cells that could be further enhanced by cotransfection of expression vectors for

NFATc1 and RelA/p65. Cotransfection of both expression vectors together could further increase the promoter activity in comparison to that achieved by an individual overexpression of each factor indicating that both cooperate in the regulation of the 500 bp *BOB.1/OBF.1* promoter activity. Additionally, supershift experiments and chromatin immunoprecipitations revealed a binding of both, NFAT and NF-κB, to this part of the promoter. Hence, a not yet identified NF-κB binding site should be present in close proximity to the mentioned non-canonical NFAT site. Because of a high homology between the DNA binding domains of NF-κB and NFAT family members, they generate a similar conformation and therefore contact identical nucleotides at the core motif GGAAA (Wolfe et al., 1997). The non-canonical NFAT site GAAGAAA within the BOB.1/OBF.1 promoter is followed by the sequence AAAAAAG (Figure 2A). Both sites share high similarity with the core sequence GGAAA to which both factors, NF-κB and NFAT, are able to bind. The second motif AAAG seems to be of biologically relevance since the point mutation G→T prevents inducible complex binding in T cells. Possibly, this part of the *BOB.1/OBF.1* promoter is essential for the cooperative action of NFAT and NF-κB necessary for the high inducible activity of the 500 bp promoter.

Additionally, the CREB/ATF binding site, important for the promoter activity in B cells (Stevens et al., 2000) seems also to be relevant in T cells, since overexpression of an active version of CREB further increased the NFAT/NF-κB induced *BOB.1/OBF.1* promoter activity in Jurkat T cells whereas mutation of this site abrogates significantly BOB.1/OBF.1 promoter activity. Therefore, the observed decreasing effect of the p38 inhibitor SB203580 on the BOB.1/OBF.1 promoter as well as on the activity of octamer-dependent transcription in general could be mediated by the inhibition of CREB and/or ATF-1 that can also bind to the same site as both factors need p38 kinase function for full transcriptional activity (Tan et al., 1996). Otherwise, an inhibition of p38 could affect NFAT proteins since p38-mediated

signaling activates NFATc1 (NFAT2) promoter activity and also increases its mRNA stability (Wu et al., 2003).

Together, here we identified three different functional NF-κB binding sites close to NFAT binding sites important for *BOB.1/OBF.1* promoter activity. One of these NF-κB sites (identified in the composite NFAT/NF-κB site) was recently identified as an alternative NF-κB binding site different from the canonical, the second (identified in the consecutive NFAT/NF-κB site) shares high homology to the canonical site, however there are variations within the second κB half-site, and the third (around the previously identified NFAT site) displays similarities to just the core motif to which NFAT and NF-κB proteins can bind. Indeed, deviations from the consensus NF-κB site were described for several promoters leading finally to the modulation of the affinity of different factors of the family to the appropriate site (Zabel et al., 1991) and also influences the composition of the interacting heterodimers bound (reviewed in (Natoli et al., 2005)). This leads to the assumption that the expression of BOB.1/OBF.1 is differentially regulated under certain conditions, like primary or secondary T cell activation or in different T helper subtypes, by different homo- or heteromeric complexes that are bound to one or the other NF-κB site within the promoter.

The fact that the here identified NF-κB sites are in close proximity to NFAT sites and thereby build either composite or consecutive combined NFAT/NF-κB binding motifs suggests that either NFAT or NF-κB complexes or both together are able to transactivate the *BOB.1/OBF.1* promoter. This conclusion is in line with our overexpression data, where the modulation of NFAT or NF-κB activity alone has already significant influence on BOB.1/OBF.1 as well as Oct2 expression. Additionally, the use of specific inhibitory peptides revealed that both the NFAT and the NF-κB signaling pathway are important for *BOB.1/OBF.1* promoter activity

and BOB.1/OBF.1 expression in T cells. Otherwise, we were not able to show simultaneous binding of both NFAT and NF-κB factors in Re-ChIP experiments. The fact that the deletion of RelA/p65 leads to a complete loss of BOB.1/OBF.1 expression in primary $CD4^+$ T cells, whereas the deletion of NFATc1/c2 decreases BOB.1/OBF.1 expression by about 50% indicates that NF-κB activation is possibly of higher importance for sufficient BOB.1/OBF.1 expression in T cells than NFAT factors. But the activity of NFAT and NF-κB depends on Ca^{2+} oscillation (Dolmetsch et al., 1998), which differs in different cell types, including TH1, TH2 and TH17 T helper subpopulations (Fanger et al., 2000; Weber et al., 2008). This leads to distinct levels of NFAT and NF-κB nuclear localization and finally to differential cytokine expression by different T helper subtypes. Therefore, it seems likely that also the availability and activity of NFAT or NF-κB family members in different T helper subpopulations or under certain differentiation conditions may determine whether the one or the other or both factors together contribute to the *BOB.1/OBF.1* promoter activity *in vivo*.

In fact, the regulation of BOB.1/OBF.1 expression in T cells seems to be complex, since in natural regulatory T cells (nTreg) its expression is down modulated (Marson et al., 2007), whereas in T cells which are anergized or suppressed by nTreg its expression was found more than 10 times higher than in non-suppressed $CD4^+$ T cells (Sukiennicki et al., 2006). Although we have already described the importance of BOB.1/OBF.1 for balanced TH1/TH2 cytokine secretion (Brunner et al., 2007), the precise regulation of BOB.1/OBF.1 and Oct expression in different T helper subpopulations that possibly affects the induction of a different set of genes, remains to be elucidated.

In summary, by an array of *in vitro* and *in vivo* approaches we provide evidences for the importance of NFAT together with NF-κB transcription factors for the regulation of octamer-

dependent transcription in T cells. Further work is required to identify the promoter region as well as regulatory *cis*-acting elements responsible for the *Oct2* gene regulation. Additionally, it will be interesting to address the question whether the same transcription factors / transcription factor binding sites, identified as important elements for the inducible BOB.1/OBF.1 expression in T cells, have also relevance for its basal as well as inducible expression in B cells.

2.5.5 Figure Legends

Figure 1. Inhibition of octamer function by CsA, p38 and NF-κB inhibitors in T cells.

A) A luciferase reporter construct bearing four copies of the consensus octamer element was stabily transfected into Jurkat T cells. Cells were either left untreated or stimulated with P, I alone or together (P/I), or were pretreated with CsA (100 ng/ml), SB203580 (20 μM) or Bay11-7082 (2 μM) prior to the stimulation with P/I. Mean values ± s.d. of 5 experiments are shown. B and J) The 1500 bp BOB.1/OBF.1 promoter construct cloned in front of a luciferase gene was transfected together with the appropriate empty vector (ev) into Jurkat T cells that were subsequently treated as described in (A) or as indicated. The determined relative luciferase activity of the empty vector without stimulation was set one and the fold increase of the promoter activity was calculated. Mean values ± s.d. of 3 (in B) and 2 (in J) independent experiments are shown J). The analyses of the pTATA promoter construct served as internal control. C to I) Jurkat or isolated primary $CD4^+$ T cells were treated as indicated. After 18 h of stimulation cells were harvested. One part was used for protein extraction and immunoblotting (C, E, G, I), and the other half was used for RNA isolation and quantitative RT-PCR (D, F, H) for the detection of *BOB.1/OBF.1* and *Oct2* mRNA expression levels that were determined relative to the expression of the housekeeping gene *RPL13*. Analyses of

HPRT mRNA expression also normalized to *RPL13* expression served as internal control (data not shown). The analyses of ERK2 expression (C, E, G, I) served as loading control.

Figure 2. Inducible complex formation on potential NFAT and NF-κB sites within the *BOB.1/OBF.1* promoter.

A) Schematic representation of the murine *BOB.1/OBF.1* promoter. Indicated are the position of the TATA-box, of the CREB/ATF-binding site, as well as the positions / sequences of the potential NFAT and NF-κB sites. B to F) Primary murine $CD4^+$ T cells were either left untreated or stimulated for 18 h with P/I. Whole protein extracts were prepared and analyzed in EMSA. In supershift experiments, the indicated antibodies were used.

Figure 3. Mutation analyses of potential NFAT and NF-κB sites within the *BOB.1/OBF.1* promoter.

A) Potential NFAT and NF-κB sites within the *BOB.1/OBF.1* promoter as well as consensus NFAT and NF-kB site were mutated as indicated and used (B to E) in EMSA experiments together with whole cell extracts of isolated murine $CD4^+$ T cells that were either left untreated or were stimulated for 18 h with P/I as indicated. F) Potential NFAT and NF-κB sites within the *BOB.1/OBF.1* promoter that was cloned in front of a luciferase gene were mutated as indicated. Mutated constructs were transfected into Jurkat T cell that were subsequently stimulated with P/I over night. The % inhibition of promoter activity was calculated relative to the activity of the wildtype 1500 bp BOB.1/OBF.1 promoter was set as 100 %.

Figure 4. The *BOB.1/OBF.1* promoter spanning 500 bp is necessary and sufficient for full inducible activity in T cells.

A) Schematic representation of the BOB.1/OBF.1 promoter. The positions of the analyzed *cis*-elements as well as of the promoter fragments used in reporter assays are indicated. B and C) Namalwa B and Jurkat T cells were transiently transfected with the empty vector (ev) or with reporter constructs bearing either the 1500 bp or the 500 bp *BOB.1/OBF.1* promoter construct or a deletion mutant of the longer version were the first 500 bp are missing (1500Δ500 bp). The relative luciferase activity (B) or the fold induction (C) of different

promoter fragments was determined without or after stimulation of cells with P/I, were in the second case the fold induction was determined relative to the luciferase activity of the empty vector without stimulation that was set to one. (D) Jurkat T cells were transiently transfected with either the empty vector or the 1500 or 500 bp *BOB.1/OBF.1* promoter constructs as indicated, either alone or together with expression vectors for NFATc1, RelA or combination of both. Transfected cells were either left untreated or stimulated subsequently with P/I. Next day the cells were harvested to determine the relative luciferase activity. The relative luciferase activity resulting from the transfection of the empty vector without induction was set to one. The fold induction relative to the empty vector was calculated and is depicted. B to D) Shown are the mean values ± s.d. of three independently performed experiments.

Figure 5. Calcineurin and NFAT factors control the inducible BOB.1/OBF.1 and Oct2 expression in T cells.

A) Jurkat T cells were nucleoporated either with scrambled siRNA or siRNA pools directed against the α and β isoforms of CN A subunit (CN A α + β). Subsequently, cells were stimulated for 8 h with P/I. Whole cell extracts were prepared and used for protein expression analyses of CN A α + β, BOB.1/OBF.1 or Oct2 by immunoblotting. The detection of ERK2 expression served as loading control. B) Human A3.01 T cells were infected retrovirally with vectors expressing either the NF1ATc1/ α A-ER or NFATc1/ α C-ER or just the empty control vector (HA-ER). Afterwards cells were either left untreated or treated with OHT, I, P/I or with combinations of these inducers as indicated. After 18 h of stimulation cells were harvested and subjected for immunoblotting using primary antibodies against NFATc1, BOB.1/OBF.1 and Oct2. The detection of ERK2 expression levels served as loading control. C and D) Primary $CD4^+$ T cells were isolated from wildtype (wt) or mutant mice in which the expression of NFATc1 and NFATc2 was simultaneously deleted in T cells specifically (NFATc1c2-DKO). After stimulation of cells with P/I or with αCD3+αCD28 antibodies for 18 h cells were harvested to analyze *BOB.1/OBF.1* as well as *Oct2* expression at mRNA levels by quantitative RT-PCR. The relative mRNA expression levels are expressed relative to that determined in unstimulated wildtype cells that were set as one. The experiments were performed twice in triplicates and using different dilutions of the cDNA reveling the same result. One representative experiment is depicted as mean values ± s.d.

Figure 6. NF-κB activity influences BOB.1/OBF.1 and Oct2 expression levels in T cells.

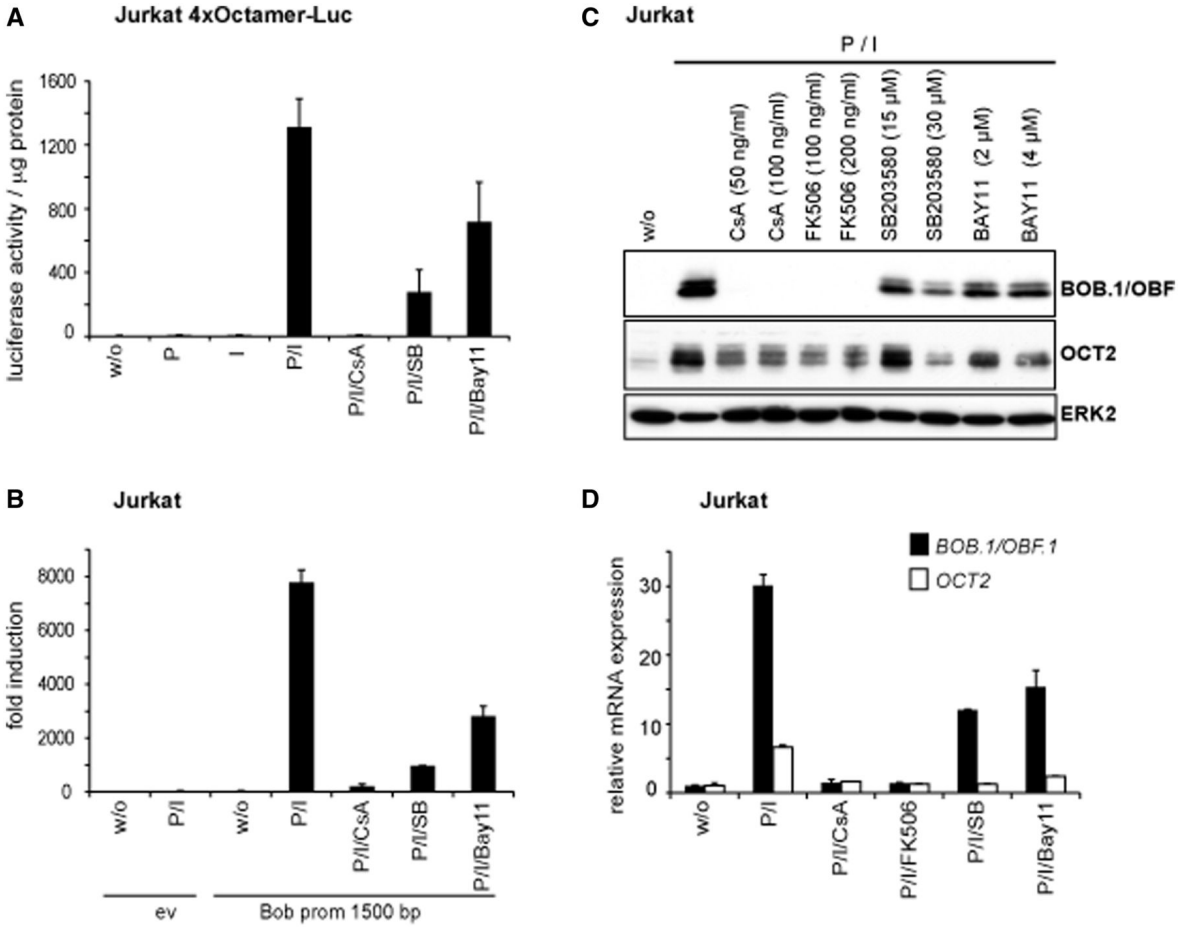
Wildtype (wt) or NEMO-deficient ($NEMO^{-/-}$) Jurkat T cells were stimulated for the time indicated with P/I or left untreated. The expression of BOB.1/OBF.1 was analyzed by Western Blotting. B) Human A3.01 T cells were retrovirally infected using vectors expressing either a constitutive active version of IKK2 (EE), a kinase dead version (KD) or the empty vector (ev). After selection, cells were stimulated with P/I for 18 h and subsequently analyzed for the protein expression levels of BOB.1/OBF.1 and Oct2 by Western Blotting (WB). The detection of IKK or ERK2 expression levels served as internal control or loading control experiments, respectively. Additionally, the same extracts were used in EMSA to monitor the NF-κB as well as Oct2 binding activity to DNA using labeled consensus sites. C) Primary $CD4^+$ T cells of mice of the indicated genotype were either left untreated or stimulated with P/I for 18 h. Afterwards, the protein expression levels of BOB.1/OBF.1 and Oct2 were analyzed in immunoblots.

Figure 7. NFAT and NF-κB transcription factors bind to different BOB.1/OBF.1 promoter regions *in vivo*.

A) The nucleotide sequence of the analyzed mouse BOB.1/OBF.1 promoter region is depicted. The NFAT as well as the predicted combined NFAT/NF-κB site are shown in red. The position of used primers for amplification of DNA fragments after chromatin immunoprecipitation are shown with bold, italic letters. Underlined letters indicate the position of the TATA box. The start site of transcription is marked as +1. B and C) Murine primary $CD4^+$ T cells were treated for 18 h with P/I. The chromatin was cross-linked, sheared and immunoprecipitated using the indicated antibodies. Immunoprecipitations using mouse IgG or normal rabbit serum served as negative controls. Immunoprecipitated DNA was purified and used as template in quantitative PCR reactions using primers as indicated in (A) for the detection of fragments bearing either the NFAT site (B) or both combined NFAT/NF-κB sites (C). (D) As an internal control, precipitated DNA was amplified using primers (Supplementary Table 3) located upstream of analyzed potential NFAT and NF-κB sites of the *BOB.1/OBF.1* promoter.

2.5.6 Figures

Figure 1.



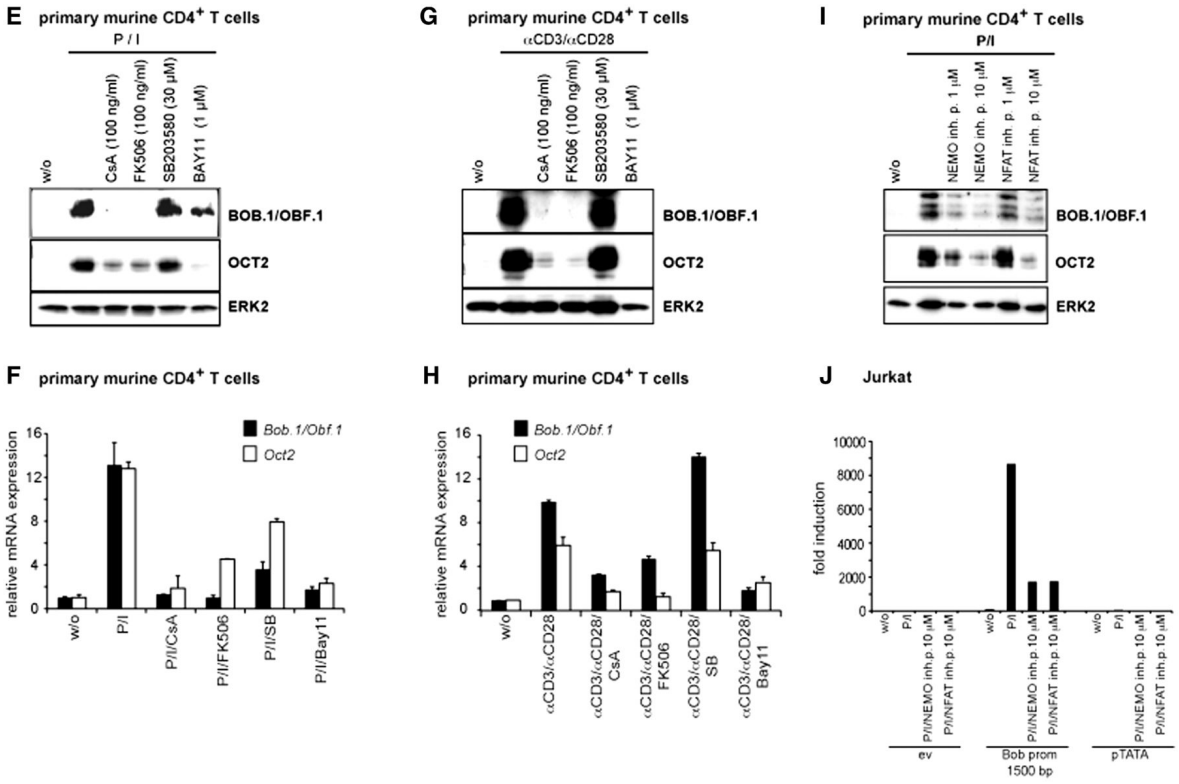

E
primary murine CD4+ T cells
P / I
w/o
CsA (100 ng/ml)
FK506 (100 ng/ml)
SB203580 (30 µM)
BAY11 (1 µM)
BOB.1/OBF.1
OCT2
ERK2
G
primary murine CD4+ T cells
αCD3/αCD28
w/o
CsA (100 ng/ml)
FK506 (100 ng/ml)
SB203580 (30 µM)
BAY11 (1 µM)
BOB.1/OBF.1
OCT2
ERK2
I
primary murine CD4+ T cells
P/I
w/o
NEMO inh. p. 1 µM
NEMO inh. p. 10 µM
NFAT inh. p. 1 µM
NFAT inh. p. 10 µM
BOB.1/OBF.1
OCT2
ERK2
F
primary murine CD4+ T cells
Bob.1/Obf.1
Oct2
relative mRNA expression
16
12
8
4
0
w/o
P/I
P/I/CsA
P/I/FK506
P/I/SB
P/I/Bay11
H
primary murine CD4+ T cells
Bob.1/Obf.1
Oct2
relative mRNA expression
16
12
8
4
0
w/o
αCD3/αCD28
αCD3/αCD28/ CsA
αCD3/αCD28/ FK506
αCD3/αCD28/ SB
αCD3/αCD28/ Bay11
J
Jurkat
fold induction
10000
8000
6000
4000
2000
0
w/o
P/I
P/I/NEMO inh.p 10 µM
P/I/NFAT inh.p 10 µM
ev
Bob prom 1500 bp
pTATA

Figure 2.

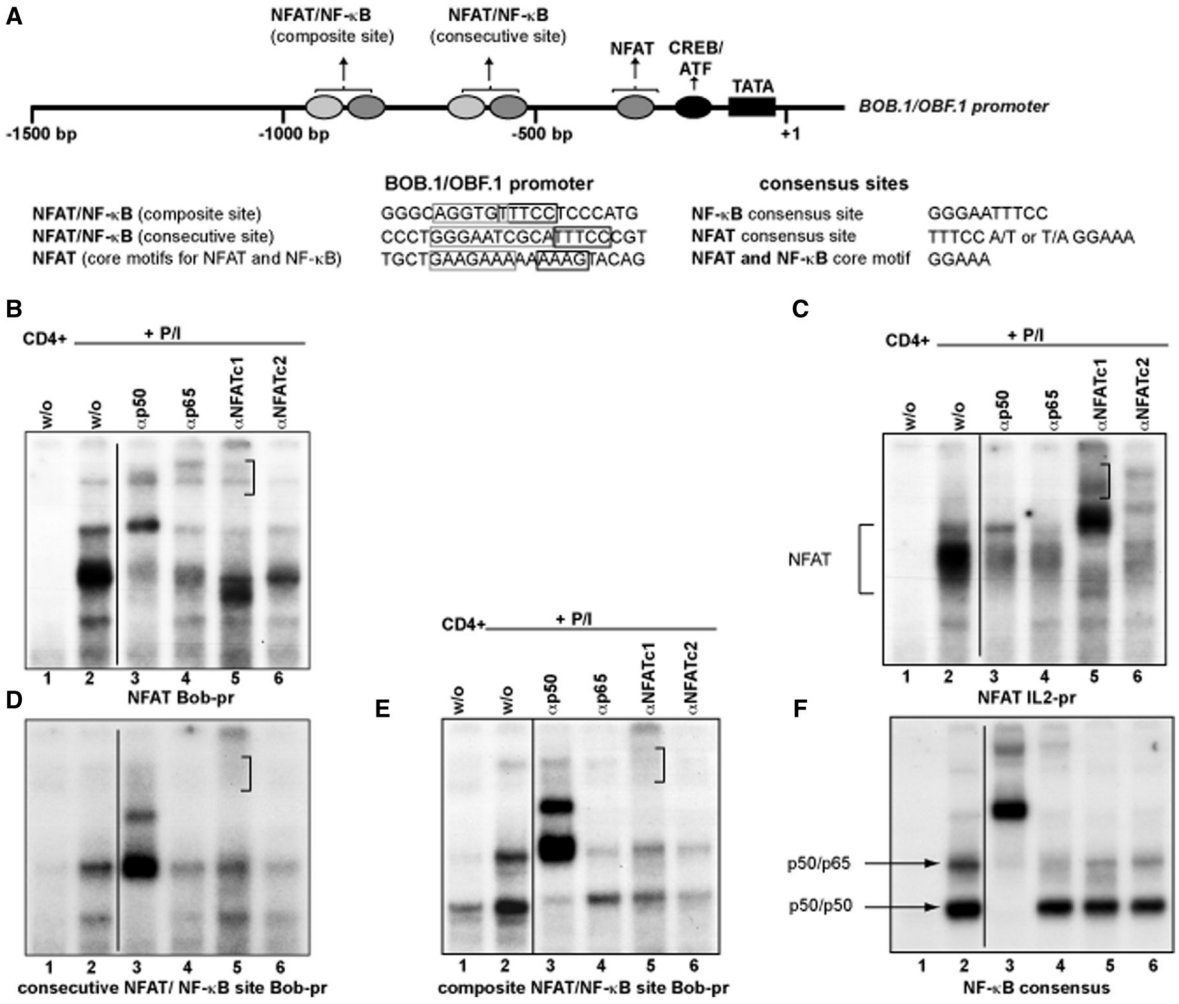

A
NFAT/NF-κB (composite site)
NFAT/NF-κB (consecutive site)
NFAT
CREB/ ATF
TATA
BOB.1/OBF.1 promoter
-1500 bp
-1000 bp
-500 bp
+1
BOB.1/OBF.1 promoter
consensus sites
NFAT/NF-κB (composite site) GGGCAGGTGTTTCCTCCCATG
NFAT/NF-κB (consecutive site) CCCTGGGAATCGCATTTCCCGT
NFAT (core motifs for NFAT and NF-κB) TGCTGAAGAAAAAAAAAGTACAG
NF-κB consensus site GGGAATTTCC
NFAT consensus site TTTCC A/T or T/A GGAAA
NFAT and NF-κB core motif GGAAA
B
CD4+
+ P/I
w/o
w/o
αp50
αp65
αNFATc1
αNFATc2
1 2 3 4 5 6
NFAT Bob-pr
C
NFAT
NFAT IL2-pr
D
consecutive NFAT/ NF-κB site Bob-pr
E
composite NFAT/NF-κB site Bob-pr
F
p50/p65
p50/p50
NF-κB consensus

Figure 3.

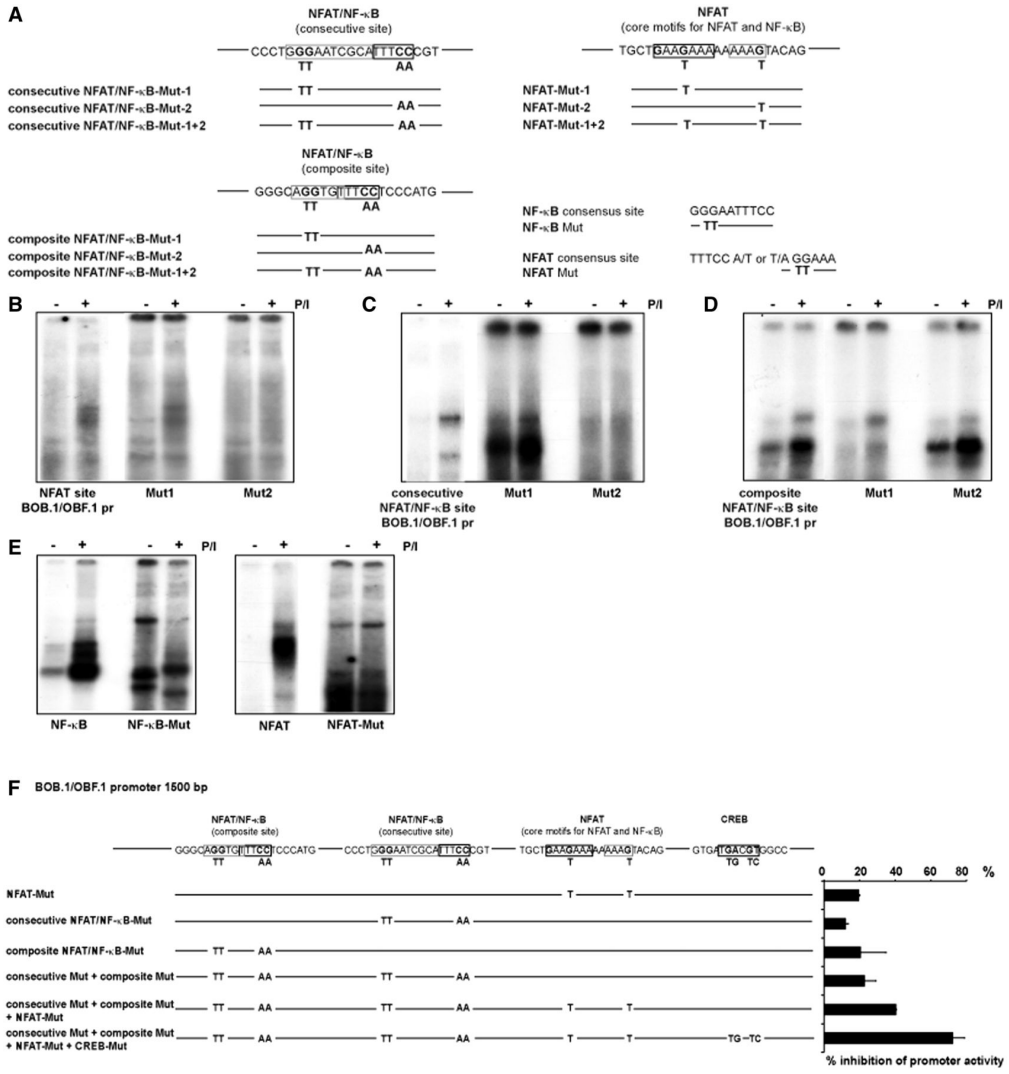

A
NFAT/NF-κB
(consecutive site)
CCCTGGGAATCGCATTTCCCGT
TT
AA
consecutive NFAT/NF-κB-Mut-1
consecutive NFAT/NF-κB-Mut-2
consecutive NFAT/NF-κB-Mut-1+2
NFAT/NF-κB
(composite site)
GGGCAGGTGTTTCCTCCCATG
composite NFAT/NF-κB-Mut-1
composite NFAT/NF-κB-Mut-2
composite NFAT/NF-κB-Mut-1+2
NFAT
(core motifs for NFAT and NF-κB)
TGCTGAAGAAAAAAAAGTACAG
T
NFAT-Mut-1
NFAT-Mut-2
NFAT-Mut-1+2
NF-κB consensus site
NF-κB Mut
GGGAATTTCC
NFAT consensus site
NFAT Mut
TTTCC A/T or T/A GGAAA
B
C
D
P/I
NFAT site
BOB.1/OBF.1 pr
Mut1
Mut2
consecutive
NFAT/NF-κB site
BOB.1/OBF.1 pr
composite
NFAT/NF-κB site
BOB.1/OBF.1 pr
E
NF-κB
NF-κB-Mut
NFAT
NFAT-Mut
F
BOB.1/OBF.1 promoter 1500 bp
CREB
TGACGT
TG TC
NFAT-Mut
consecutive NFAT/NF-κB-Mut
composite NFAT/NF-κB-Mut
consecutive Mut + composite Mut
consecutive Mut + composite Mut
+ NFAT-Mut
consecutive Mut + composite Mut
+ NFAT-Mut + CREB-Mut
0
20
40
60
80
%
% inhibition of promoter activity

Figure 4.

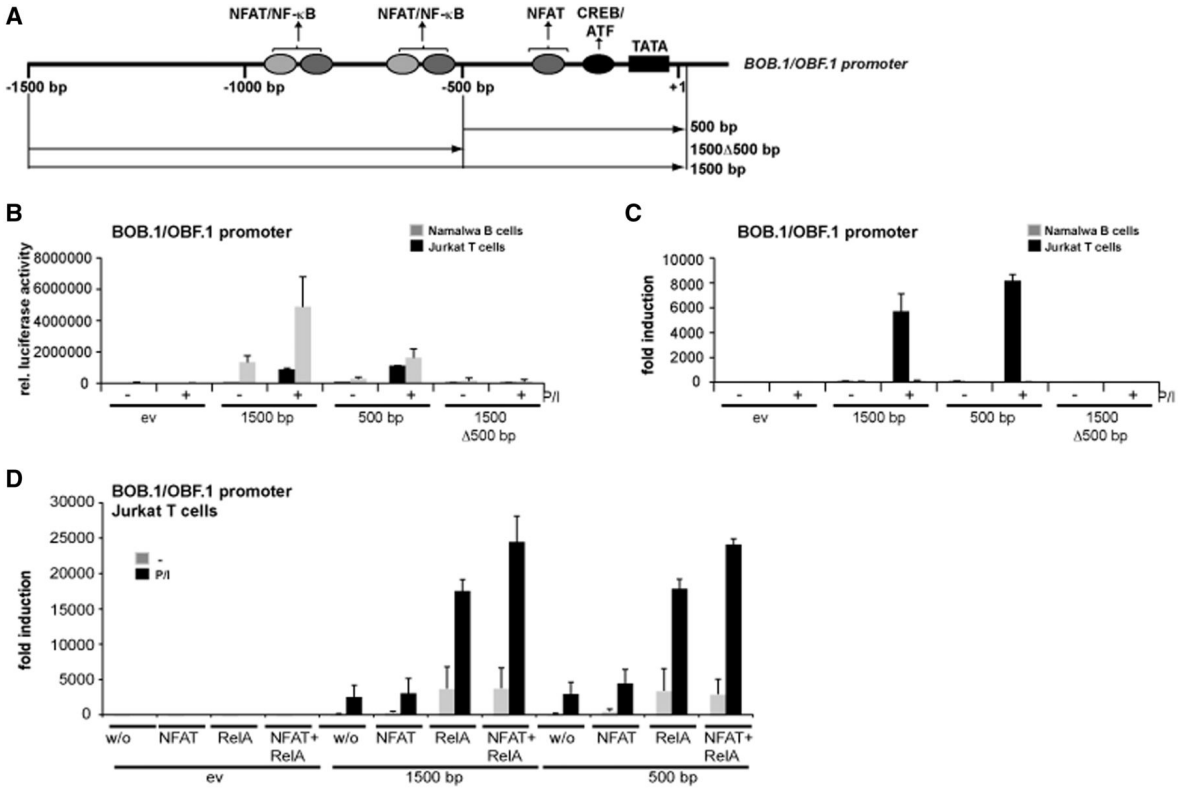

A
NFAT/NF-κB
NFAT/NF-κB
NFAT
CREB/ ATF
TATA
BOB.1/OBF.1 promoter
-1500 bp
-1000 bp
-500 bp
+1
500 bp
1500Δ500 bp
1500 bp
B
BOB.1/OBF.1 promoter
Namalwa B cells
Jurkat T cells
rel. luciferase activity
8000000
6000000
4000000
2000000
0
ev
1500 bp
500 bp
1500 Δ500 bp
P/I
C
BOB.1/OBF.1 promoter
Namalwa B cells
Jurkat T cells
fold induction
10000
8000
6000
4000
2000
0
ev
1500 bp
500 bp
1500 Δ500 bp
P/I
D
BOB.1/OBF.1 promoter
Jurkat T cells
fold induction
30000
25000
20000
15000
10000
5000
0
P/I
w/o
NFAT
RelA
NFAT+ RelA
ev
1500 bp
500 bp

Figure 5.

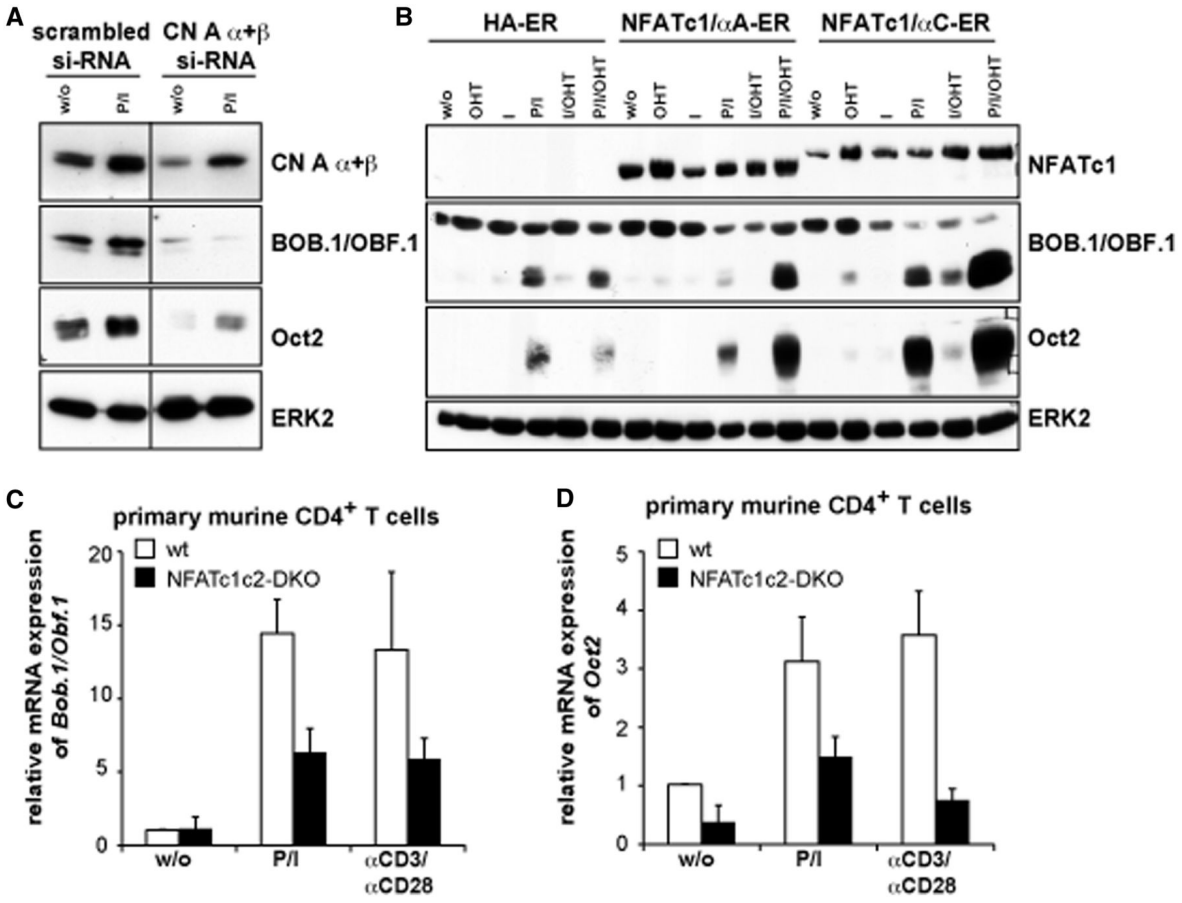

A
scrambled si-RNA
CN A α+β si-RNA
w/o
P/I
CN A α+β
BOB.1/OBF.1
Oct2
ERK2
B
HA-ER
NFATc1/αA-ER
NFATc1/αC-ER
w/o
OHT
I
P/I
I/OHT
P/I/OHT
NFATc1
BOB.1/OBF.1
Oct2
ERK2
C
primary murine CD4+ T cells
wt
NFATc1c2-DKO
relative mRNA expression of Bob.1/Obf.1
0
5
10
15
20
w/o
P/I
αCD3/ αCD28
D
primary murine CD4+ T cells
wt
NFATc1c2-DKO
relative mRNA expression of Oct2
0
1
2
3
4
5
w/o
P/I
αCD3/ αCD28

Figure 6.

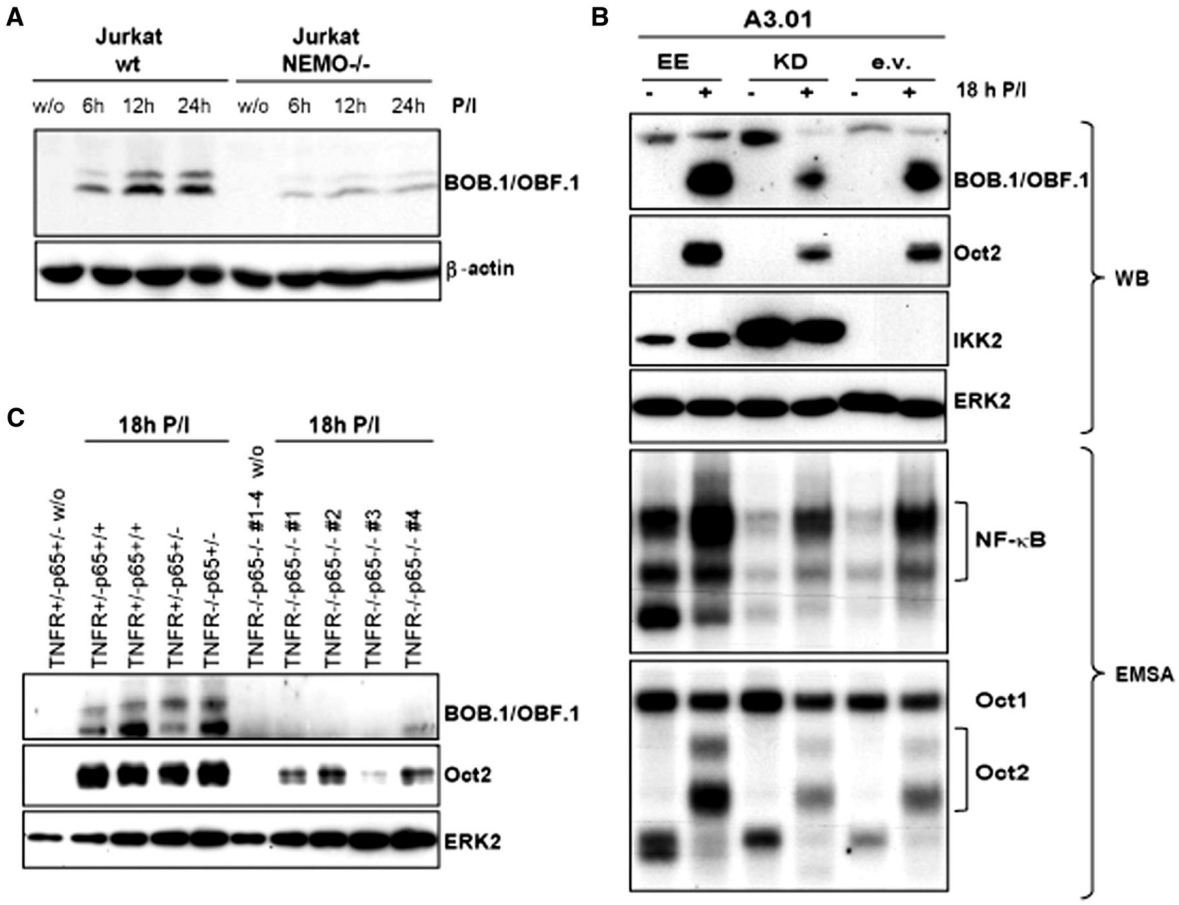
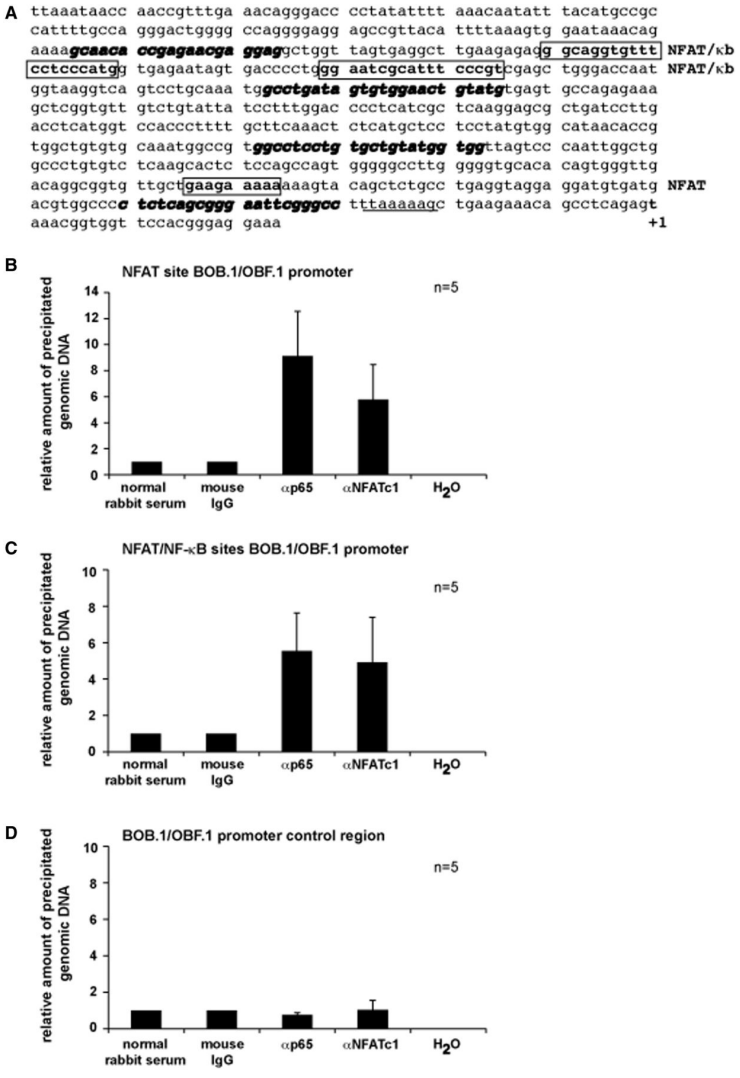

A
Jurkat wt
Jurkat NEMO-/-
w/o 6h 12h 24h w/o 6h 12h 24h P/I
BOB.1/OBF.1
β-actin
B
A3.01
EE
KD
e.v.
- + - + - + 18 h P/I
BOB.1/OBF.1
Oct2
IKK2
ERK2
WB
NF-κB
Oct1
Oct2
EMSA
C
18h P/I
18h P/I
TNFR+/-p65+/- w/o
TNFR+/-p65+/+
TNFR+/-p65+/+
TNFR+/-p65+/-
TNFR-/-p65+/-
TNFR-/-p65-/- #1-4 w/o
TNFR-/-p65-/- #1
TNFR-/-p65-/- #2
TNFR-/-p65-/- #3
TNFR-/-p65-/- #4
BOB.1/OBF.1
Oct2
ERK2

Figure 7.


A
ttaaataacc aaccgtttga aacagggacc cctatatttt aaacaatatt tacatgccgc
cattttgcca gggactgggg ccaggggagg agccgttaca ttttaaagtg gaataaacag
aaaagcaaca ccgagaacga ggaggctggt tagtgaggct tgaagagagg gcaggtgttt NFAT/κb
cctcccatgg tgagaatagt gacccctggg aatcgcattt cccgtcgagc tgggaccaat NFAT/κb
ggtaaggtca gtcctgcaaa tggcctgata gtgtggaact gtatgtgagt gccagagaaa
gctcggtgtt gtctgtatta tcctttggac ccctcatcgc tcaaggagcg ctgatccttg
acctcatggt ccacccttttt gcttcaaact ctcatgctcc tcctatgtgg cataacaccg
tggctgtgtg caaatggccg tggcctcctg tgctgtatgg tggttagtcc caattggctg
gccctgtgtc tcaagcactc tccagccagt ggggggccttg ggggtgcaca cagtgggttg
acaggcggtg ttgctgaaga aaaaaaagta cagctctgcc tgaggtagga ggatgtgatg NFAT
acgtggcccc tctcagcggg aattcgggcc ttaaaaagc tgaagaaaca gcctcagagt
aaacggtggt tccacgggag gaaa +1
B
NFAT site BOB.1/OBF.1 promoter
n=5
relative amount of precipitated genomic DNA
0
2
4
6
8
10
12
14
normal rabbit serum
mouse IgG
αp65
αNFATc1
H2O
C
NFAT/NF-κB sites BOB.1/OBF.1 promoter
n=5
relative amount of precipitated genomic DNA
0
2
4
6
8
10
normal rabbit serum
mouse IgG
αp65
αNFATc1
H2O
D
BOB.1/OBF.1 promoter control region
n=5
relative amount of precipitated genomic DNA
0
2
4
6
8
10
normal rabbit serum
mouse IgG
αp65
αNFATc1
H2O

3 Discussion and Future Plan

The computational framework, which was developed in this thesis, exploits phylogenetic conservation in terms of a complexity assessment of co-occurring TFBS to identify functional regulatory regions (as shown in this thesis for the *BOB.1/OBF.1* promoter: Brunner et al., 2013; Claussnitzer 2013-1). Furthermore, this work shows that its use in the integrative analysis of variants identified by GWAS and full-genome sequencing approaches may increase our ability to pinpoint genetic variants that mechanistically contribute to human traits, including disease, by dysregulation of gene expression (as shown in this thesis for risk SNPs associated with T2D and adiponectin level: Claussnitzer et al., Cell in press; Laumen et al., 2009; Claussnitzer 2013-2). While it has become increasingly clear that the majority of genotype-disease associations mapped so far in numerous GWAS and somatic mutations in cancer involve intergenic and intronic risk alleles (Trynka et al., 2013), so far no algorithms have been developed to harness the power of conserved TFBS patterns within CRMs, to predict regulatory variants involved in diseases and molecular phenotypes. Though I focused in this study on common T2D-associated SNPs that localize within predicted cross-species conserved CRMs, the approach may be applied to any non-coding genetic variant in regards to both heritable and somatic human variability, and may help to unveil causal gene regulatory mechanisms underlying complex diseases including cancer.

In the following, I briefly describe the differences of the PMCA methodology compared to existing functional genomics and computational-based analysis methods and proceed then with potential future directions on how to synergistically use the PMCA framework with other data modalities.

(1) Distinguishing functional conservation and sequence conservation

Before now, computational approaches that use evolutionary sequence conservation have been proposed to predict candidate sequence variants with functional effects in human disease (Cooper et al., 2005; Lindblad-Toh et al., 2011). However, although pure sequence alignment has been successful in predicting deleterious protein-altering variants in coding regions of the genome, the predictive power of these classical alignment-score approaches remains limited for causal variant discovery in non-coding regions from GWAS and sequencing studies. Given that TFBS turnover is characteristic for CRM evolution, it does not seem surprising that most functionally annotated regulatory elements are not constrained between species and that only a fraction of computationally inferred highly conserved non-coding sequences reveal features of transcriptional regulation (Ludwig et al., 2000; Dermitzakis and Clark, 2002; Fisher et al., 2006; The ENCODE Project Consortium, 2007; Attanasio et al., 2008; Blow et al., 2010, 2010; Schmidt et al., 2010; He et al., 2011; The ENCODE Project Consortium, 2012). The difference of the PMCA methodology is to discover co-occurrences of TFBS in a CRM, regardless of the cross-species conservation of the complete sequence. Indeed, using transgenic mouse embryos to identify active enhancers in subregions of the developing telencaephalon, Visel et al., 2013 have recently suggested that the combination of TFBS, rather than a single TFBS, via combinatorial TF binding governs spatial enhancer activity. Furtermore, it is of note that the majority of *cis*-regulatory SNP variants, which were identified and experimentally characterized in this work, lacked an overlap with constrained genomic regions. In essence, with instances for bounds of sequence conservation approaches, I show in this thesis that sequence alignment algorithms are often not informative and that accounting for sequence variability within CRMs is crucial to pinpoint *cis*-regulatory variants in the non-coding genome and to infer their functional upstream and downstream effects.

(2) Synergizing functional genomics and computational functional conservation approaches

Transcriptional regulation is controlled through multiple genomic regulatory layers. This includes chromatin accessibility, histone posttranslational modifications, TF binding and mRNA and nascent transcription. Projects such as ENCODE and GENCODE have begun epigenetic and non-coding RNA annotation of non-coding genomic regions and these databases have been used to associate GWAS findings en masse with putative function (ENCODE, 2012; Harrow et al., 2012). As a proof of concept, using ENCODE chromatin state and TF binding data (Neph et al., 2012b; The ENCODE Project Consortium, 2012), this work shows that examining cross-species TFBS co-occurrences is predicitive of regulatory functionality. It is important to note that non-computational approaches based on functional genomics data (e.g. DNase-, ChIP-seq, FAIRE-seq, RNA-seq) (The ENCODE Project Consortium, 2007; Gaulton et al., 2010; Stitzel et al., 2010; The ENCODE Project Consortium, 2012) have the disadvantage that they require access to appropriate human tissue, or to tissue from a particular developmental time stage (which frequently is impossible), and they can be hampered by the complexity of effects on gene regulation, such as environmental effects or epigenetic complexity. Another practical challenge for ChIP-seq experiments is the size of the TF repertoire which implies that more than 1,000 TFs (Vaquerizas et al., 2009) need to be potentially assayed in large numbers of different cell types and environmental states, until one could identify the disease-relevant TFs and their variant-dependent DNA sequence binding changes. Indeed, this thesis shows that computational analysis of cross-species conserved TFBS co-occurrences pinpoints cell-type and cell-stage-specific regulatory meaningful variants where annotation from large scale functional studies is still incomplete. PMCA might therefore serve as a more universally applicable approach than approaches relying on functional genomics data.

Interestingly, a study recently published in Science (Kilpinen et al., 2013) examined the allelic variability within families (trio study) for different molecular phenotypes, i.e.

chromatin states, histone modifications, TF binding and transcription assayed in lymphoblastoid cell lines. This study could show that allelic variability in gene expression within individuals is primarily caused by TF binding, rather than histone tail modifications, and that transmission of chromatin states from parents to children depends on the properties of the underlying DNA sequence. Apart from that, leveraging genome-wide occupancy maps for a large number of TF, recent studies have shown the crucial importance of combinatorial TF binding to TFBS clusters in regulatory genomic regions (Wilson et al., 2010; Ravasi et al., 2010). Thus, genetically driven changes in binding of transcription factors to regulatory genomic regions in place of nucleosomes appear to trigger chromatin remodelling, suggesting TF as major determinants of allele-specific regulatory interactions. In future studies, it will be exciting to integrate the context-independent computational framework PMCA with those molecular phenotypic data modalities to pinpoint common and rare *cis*-regulatory genetic variants, their underlying binding TFs and the effect on gene expression, reflected by histone modifications. In sum, these results stress the need of integrative frameworks, where systems biology and regulatory genomics data are synergistically combined with computational approaches to investigate disease underlying molecular mechanisms at disease susceptibility loci (Califano et al., 2012).

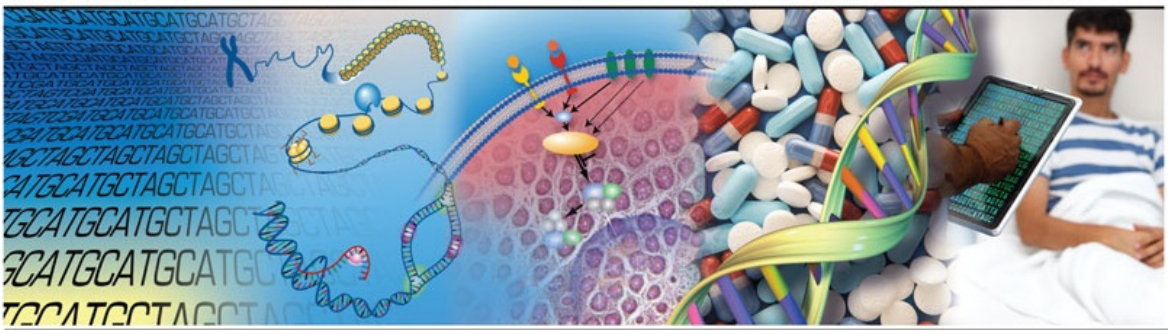


Figure 7: Eric Green from the NIH, US projects in his `Charting a course for genomic medicine from base pairs to bedside`(Nature Perspective, 2011) that the transition from genomics research towards genomics medicine consecutively depends on understanding of the structure and biology of genomes, understanding the biology and its perturbation in disease, advancing the science of medicine and improving the effectiveness of healthcare.

The rapid advances in genomics hold great promise for the study of human diseases and phenotypic traits. GWAS of common diseases have now identified thousands of genomic regions harboring disease-associated variants (Hindorff LA) and currently ongoing targeted sequencing and whole-genome sequencing studies will further lead the genomics community towards the discovery of millions of genetic variants not only for human common complex disorders but also for different types of cancer and rare Mendelian diseases. Before now, translation of these findings into an improved understanding of human disease has only been accomplished for a miniscule part of genomic risk loci. Identifying the disease-causing sequence variants is a prerequisite for the translation of GWAS- and sequencing-driven genomic risk loci towards personalized medicine, as the interpretation of a variants` functional effect is essential for developing therapeutic strategies and improving diagnostic means (Figure 7). This work may therefore represent a useful step toward translating genomics data into basic molecular mechanisms underlying human disease risk loci, and

ultimately to tailor medical treatment to the individual patient, i.e. personalized medicine (Figure 8).

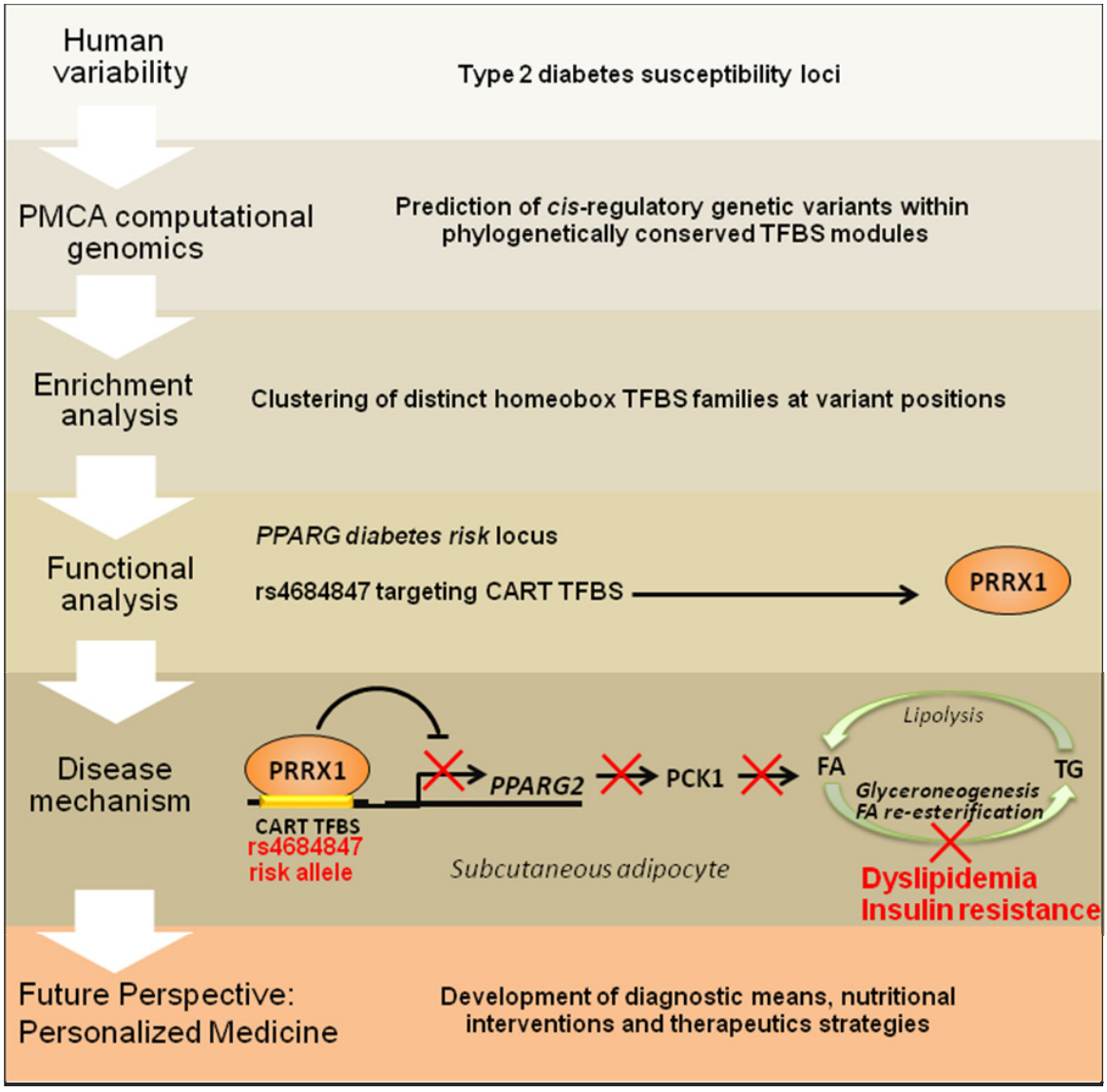


Figure 8: The interpretation of functional effects for the plethora of sequence variants is essential for understanding the genetic basis of variation in human diseases and traits.

4 References

(2007). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851-861.

(2010). On beyond GWAS. Nat Genet *42*, 551.

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

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

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Meth *7*, 248-249.

Alcamo, E., Mizgerd, J.P., Horwitz, B.H., Bronson, R., Beg, A.A., Scott, M., Doerschuk, C.M., Hynes, R.O. and Baltimore, D. (2001) Targeted mutation of TNF receptor I rescues the RelA-deficient mouse and reveals a critical role for NF-kappa B in leukocyte recruitment. *J Immunol*, **167**, 1592-1600.

Al Sarraj, J., Vinson, C., Han, J. and Thiel, G. (2005) Regulation of GTP cyclohydrolase I gene transcription by basic region leucine zipper transcription factors. *J Cell Biochem*, **96**, 1003-1020.

Annweiler, A., Zwilling, S., Hipskind, R.A. and Wirth, T. (1993) Analysis of transcriptional stimulation by recombinant Oct proteins in a cell free system. *J. Biol. Chem.*, **268**, 2525-2534.

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

Attanasio, C., Reymond, A., Humbert, R., Lyle, R., Kuehn, M., Neph, S., Sabo, P., Goldy, J., Weaver, M., and Haydock, A., et al. (2008). Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. Genome Biology *9*, R168.

Bamshad, M.J., Mummidi, S., Gonzalez, E., Ahuja, S.S., Dunn, D.M., Watkins, W.S., Wooding, S., Stone, A.C., Jorde, L.B., and Weiss, R.B., et al. (2002). A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc Natl Acad Sci U S A *99*, 10539-10544.

Beg, A.A., Sha, W.C., Bronson, R.T., Ghosh, S. and Baltimore, D. (1995) Embryonic lethality and liver degeneration in mice lacking the RelA component of NF-kappa B. *Nature*, **376**, 167-170.

Berg AH, Combs TP, Du X, Brownlee M, Scherer PE.: The adipocyte-secreted protein Acrp30 enhances hepatic insulin action. *Nat Med* 7: 947– 953, 2001.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. The American Journal of Human Genetics *74*, 1111-1120.

Bhargava, A.K., Li, Z. and Weissman, S.M. (1993) Differential expression of four members of the POU family of proteins in activated and phorbol 12-myristate 13-acetate-treated Jurkat T cells. *Proc Natl Acad Sci U S A*, **90**, 10260-10264.

Bhattacharyya, S., Deb, J., Patra, A.K., Thuy Pham, D.A., Chen, W., Vaeth, M., Berberich-Siebelt, F., Klein-Hessling, S., Lamperti, E.D., Reifenberg, K. *et al.* (2011) NFATc1 affects mouse splenic B cell function by controlling the calcineurin--NFAT signaling network. *J Exp Med*, **208**, 823-839.

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

Bonnefond, A., Froguel, P., and Vaxillaire, M. (2010). The emerging genetics of type 2 diabetes. Trends Mol.Med *16*, 407-416.

Bouatia-Naji N, Meyre D, Lobbens S, Seron K, Fumeron F, Balkau B, Heude B, Jouret B, Scherer PE, Dina C, Weill J, Froguel P.: ACDC/adiponectin polymorphisms are associated with severe childhood and adult obesity. *Diabetes* 55: 545– 550, 2006.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311-322.

Brunner, C., Marinkovic, D., Klein, J., Samardzic, T., Nitschke, L. and Wirth, T. (2003) B cell-specific transgenic expression of Bcl2 rescues early B lymphopoiesis but not B cell responses in BOB.1/OBF.1-deficient mice. *J Exp Med*, **197**, 1205-1211.

Brunner, C. and Wirth, T. (2006) Btk expression is controlled by Oct and BOB.1/OBF.1. *Nucleic Acids Res*, **34**, 1807-1815.

Brunner, C., Sindrilaru, A., Girkontaite, I., Fischer, K.D., Sunderkotter, C. and Wirth, T. (2007) BOB.1/OBF.1 controls the balance of TH1 and TH2 immune responses. *Embo J*, **26**, 3191-3202.

Brunvand, M.W., Schmidt, A. and Siebenlist, U. (1988) Nuclear factors interacting with the mitogen-responsive regulatory region of the interleukin-2 gene. *J Biol Chem*, **263**, 18904-18910.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., and Samani, N.J., et al. (2007a). Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet *39*, 1329-1337.

Burton, P.R., Clayton, D.G., Cardon, L.R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D.P., McCarthy, M.I., Ouwehand, W.H., and Samani, N.J., et al. (2007b). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661-678.

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

Cardon, L.R., and Abecasis, G.R. (2003). Using haplotype blocks to map human complex trait loci. Trends Genet *19*, 135-140.

Chuvpilo, S., Schomberg, C., Gerwig, R., Heinfling, A., Reeves, R., Grummt, F. and Serfling, E. (1993) Multiple closely-linked NFAT/octamer and HMG I(Y) binding sites are part of the interleukin-4 promoter. *Nucleic Acids Res*, **21**, 5694-5704.

Chuvpilo, S., Jankevics, E., Tyrsin, D., Akimzhanov, A., Moroz, D., Jha, M.K., Schulze-Luehrmann, J., Santner-Nanan, B., Feoktistova, E., Konig, T. *et al.* (2002) Autoregulation of NFATc1/A expression facilitates effector T cells to escape from rapid apoptosis. *Immunity*, **16**, 881-895.

Clipstone, N.A. and Crabtree, G.R. (1992) Identification of calcineurin as a key signalling enzyme in T-lymphocyte activation. *Nature*, **357**, 695-697.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet *12*, 628-640.

Corcoran, L.M., Karvelas, M., Nossal, G.J.V., Ye, Z.-S., Jacks, T. and Baltimore, D. (1993) Oct-2, although not required for early B-cell development, is critical for later B-cell maturation and for postnatal survival. *Genes Dev.*, **7**, 570-582.

CRICK, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561-563.

Daniel L. Hartl and Andrew G. Clark. Principles of Population Genetics (2007), Fourth Edition.

Denk, A., Goebeler, M., Schmid, S., Berberich, I., Ritz, O., Lindemann, D., Ludwig, S. and Wirth, T. (2001) Activation of NF-kappa B via the Ikappa B kinase complex is both essential and sufficient for proinflammatory gene expression in primary endothelial cells. *J Biol Chem*, **276**, 28451-28458.

Dermitzakis, E.T., and Clark, A.G. (2002). Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. Molecular Biology and Evolution *19*, 1114-1121.

de Grazia, U., Felli, M.P., Vacca, A., Farina, A.R., Maroder, M., Cappabianca, L., Meco, D., Farina, M., Screpanti, I., Frati, L. *et al.* (1994) Positive and negative regulation of the composite octamer motif of the Interleukin 2 enhancer by AP-1, OCT-2, and retinoic acid receptor. *J. Exp. Med.*, **180**, 1485-1497.

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez, A.M., and Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. Science *325*, 1246-1250.

Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., and Farrall, M., et al. (2007). A genome-wide association study of global gene expression. Nat.Genet *39*, 1202-1207.

Dolmetsch, R.E., Xu, K. and Lewis, R.S. (1998) Calcium oscillations increase the efficiency and specificity of gene expression. *Nature*, **392**, 933-936.

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

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

Eeles, R.A., Kote-Jarai, Z., Giles, G.G., Olama, A.A.A., Guy, M., Jugurnauth, S.K., Mulholland, S., Leongamornlert, D.A., Edwards, S.M., and Morrison, J., et al. (2008).

Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet *40*, 316-321.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L., and Jarvela, I. (2002). Identification of a variant associated with adult-type hypolactasia. Nat Genet *30*, 233-237.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., and Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Fanger, C.M., Neben, A.L. and Cahalan, M.D. (2000) Differential Ca2+ influx, KCa channel activity, and Ca2+ clearance distinguish Th1 and Th2 lymphocytes. *J Immunol*, **164**, 1153-1160.

Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science *312*, 276-279.

Frischbutter, S., Gabriel, C., Bendfeldt, H., Radbruch, A. and Baumgrass, R. (2011) Dephosphorylation of Bcl-10 by calcineurin is essential for canonical NF-kappaB activation in Th cells. *Eur J Immunol*, **41**, 2349-2357.

Fruebis J, Tsao TS, Javorschi S, Ebbets-Reed D, Erickson MR, Yen FT, Bihain BE, Lodish HF.: Proteolytic cleavage product of 30-kDa adipocyte complement-related protein increases fatty acid oxidation in muscle and causes weight loss in mice. *Proc Natl Acad Sci U S A* 98: 2005– 2010, 2001.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., and Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat.Genet *42*, 255-259.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.-K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., and Alexander, R., et al. (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91-100.

Giffin, M.J., Stroud, J.C., Bates, D.L., von Koenig, K.D., Hardin, J. and Chen, L. (2003) Structure of NFAT1 bound as a dimer to the HIV-1 LTR kappa B element. *Nat Struct Biol*, **10**, 800-806.

Green, E.D., and Guyer, M.S. (2011). Charting a course for genomic medicine from base pairs to bedside. Nature *470*, 204-213.

Greiner, A., Muller, K.B., Hess, J., Pfeffer, K., Muller-Hermelink, H.K. and Wirth, T. (2000) Up-regulation of BOB.1/OBF.1 expression in normal germinal center B cells and germinal center-derived lymphomas. *Am J Pathol*, **156**, 501-507.

Gstaiger, M., Georgiev, O., van Leeuwen, H., van der Vliet, P. and Schaffner, W. (1996) The B cell coactivator Bob1 shows DNA sequence-dependent complex formation with the Oct-1/Oct-2 factors, leading to differential promoter activation. *EMBO J.*, **15**, 2781-2790.

Gstaiger, M., Knoepfel, L., Georgiev, O., Schaffner, W. and Hovens, C.M. (1995) A B-cell coactivator of octamer-binding transcription factors. *Nature*, **373**, 360-362.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., and Chee, M.S. (2005). A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet *37*, 549-554.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., and Sakaguchi, A.Y. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. Nature *306*, 234-238.

Gustafson B, Jack MM, Cushman SW, Smith U.: Adiponectin gene activation by thiazolidinediones requires PPAR-gamma2, but not C/EBP-alpha: evidence for differential regulation of the aP2 and adiponectin genes. *Biochem Biophys Res Comm* 308: 933– 939, 2003.

Hakonarson, H., Grant, S.F.A., Bradfield, J.P., Marchand, L., Kim, C.E., Glessner, J.T., Grabs, R., Casalunovo, T., Taback, S.P., and Frackelton, E.C., et al. (2007). A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. Nature *448*, 591-594.

Hamblin, M.T., Thompson, E.E., and Di Rienzo, A. (2002). Complex Signatures of Natural Selection at the Duffy Blood Group Locus. The American Journal of Human Genetics *70*, 369-383.

Hardison, R., Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. Nature Reviews Genetics 13, 469-483.

Harhaj, E.W., Good, L., Xiao, G., Uhlik, M., Cvijic, M.E., Rivera-Walsh, I. and Sun, S.C. (2000) Somatic mutagenesis studies of NF-kappa B signaling in human T cells: evidence for an essential role of IKK gamma in NF-kappa B activation by T-cell costimulatory signals and HTLV-I Tax protein. *Oncogene*, **19**, 1448-1456.

Hauner H. (2002). The mode of action of thiazolidinediones. *Diabete Metab Res Rev* 18 (Suppl. 2): 10– 15.

Hauner H.: Secretory factors from human adipose tissue and their functional role. *Proc Nutr Soc* 64: 163– 169, 2005.

He, B.Z., Holloway, A.K., Maerkl, S.J., and Kreitman, M. (2011). Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. PLoS Genet *7*, e1002053EP -.

Heid IM, Wagner SA, Gohlke H, Iglseder B, Mueller JC, Cip P, Ladurner G, Reiter R, Stadlmayr A, Mackevics V, Illig T, Kronenberg F, Paulweber B.: Genetic architecture of the APM1 gene and its influence on adiponectin plasma levels and parameters of the metabolic syndrome in 1,727 healthy Caucasians. *Diabetes* 55: 375– 384, 2006.

Hentsch, B., Mouzaki, A., Pfeuffer, I., Rungger, D. and Serfling, E. (1992) The weak, fine-tuned binding of ubiquitous transcription factors to the Il-2 enhancer contributes to its T cell-restricted activity. *Nucleic Acids Res*, **20**, 2657-2665.

Herder C, Hauner H, Haastert B, Rohrig K, Koenig W, Kolb H, Muller-Scholze S, Thorand B, Holle R, Rathmann W.: Hypoadiponectinemia and proinflammatory state: two sides of the same coin? Results from the Cooperative Health Research in the Region of Augsburg Survey 4 (KORA S4). *Diabetes Care* 29: 1626– 1631, 2006.

Hess, J., Nielsen, P.J., Fischer, K.D., Bujard, H. and Wirth, T. (2001) The B lymphocyte-specific coactivator BOB.1/OBF.1 is required at multiple stages of B-cell development. *Mol Cell Biol*, **21**, 1531-1539.

Hill, W.G., and Robertson, A. (1968). The effects of inbreeding at loci with heterozygote advantage. Genetics *60*, 615-628.

Hindorff LA, M.J.W.A.J.H.H.P.K.A.a.M.T. A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed [July 2012].

Hogan, PG., Chen L., Nardone J., Rao A. (2003). Transcriptional regulation by calcium, calcineurin and NFAT. *Genes Dev.* 17(18):2205-32.

Hotta K, Funahashi T, Arita Y, Takahashi M, Matsuda M, Okamoto Y, Iwahashi H, Kuriyama H, Ouchi N, Maeda K, Nishida M, Kihara S, Sakai N, Nakajima T, Hasegawa K, Muraguchi M, Ohmoto Y, Nakamura T, Yamashita S, Hanafusa T, Matsuzawa Y.: Plasma concentrations of a novel, adipose-specific protein, adiponectin, in type 2 diabetic patients. *Arterioscler Thromb Vasc Biol* 20: 1595– 1599, 2000.

Human Genome Sequencing Consortium, I. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Iwaki M, Matsuda M, Maeda N, Funahashi T, Matsuzawa Y, Makishima M, Shimomura I.: Induction of adiponectin, a fat-derived antidiabetic and antiatherogenic factor, by nuclear receptors. *Diabetes* 52: 1655– 1663, 2003.

Iwashima Y, Katsuya T, Ishikawa K, Ouchi N, Ohishi M, Sugimoto K, Fu Y, Motone M, Yamamoto K, Matsuo A, Ohashi K, Kihara S, Funahashi T, Rakugi H, Matsuzawa Y, Ogihara T.: Hypoadiponectinemia is an independent risk factor for hypertension. *Hypertension* 43: 1318– 1323, 2004.

Jin, L., Sliz, P., Chen, L., Macian, F., Rao, A., Hogan, P.G. and Harrison, S.C. (2003) An asymmetric NFAT1 dimer on a pseudo-palindromic kappa B-like DNA site. *Nat Struct Biol*, 10, 807-811.

Jolma A., Yan J., Whitington T., Toivonen J., Nitta KR., Rastas P., Morgunova E., Enge M., Taipale M., Wei G., Palin K., Vaquerizas JM., Vincentelli R., Luscombe NM., Hughes TR., Lemaire P., Ukkonen E., Kivioja T., Taipale J. (2013). DNA-binding specificities of human transcription factors. Cell;152(1-2):327-39.

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

Kahn, C.R. (1994). Banting Lecture. Insulin action, diabetogenes, and the cause of type II diabetes. Diabetes *43*, 1066-1084.

Kadowaki T, Yamauchi T, Kubota N, Hara K, Ueki K, Tobe K. (2006) Adiponectin and adiponectin receptors in insulin resistance, diabetes, and the metabolic syndrome. *J Clin Invest* 116: 1784– 1792.

Kang, S.-M., Tsang, W., Doll, S., Scherle, P., Ko, H.-S., Tran, A.-C., Lenardo, M.J. and Staudt, L.M. (1992) Induction of the POU domain transcription factor Oct-2 during T-cell activation by cognate antigen. *Mol. Cell. Biol.*, **12**, 3149-3154.

Kantorovitz, M.R., Kazemian, M., Kinston, S., Miranda-Saavedra, D., Zhu, Q., Robinson, G.E., Göttgens, B., Halfon, M.S., and Sinha, S. (2009). Motif-Blind, Genome-Wide Discovery of cis-Regulatory Modules in Drosophila and Mouse. Developmental Cell *17*, 568-579.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., and Urban, A.E., et al. (2010). Variation in Transcription Factor Binding Among Humans. Science *328*, 232-235.

Kempe, S., Kestler, H., Lasar, A. and Wirth, T. (2005) NF-kappaB controls the global pro-inflammatory response in endothelial cells: evidence for the regulation of a pro-atherogenic program. *Nucleic Acids Res*, **33**, 5308-5319.

Kilpinen H, Waszak SM, Gschwind AR, Raghav SK, Witwicki RM, Orioli A, Migliavacca E, Wiederkehr M, Gutierrez-Arcelus M, Panousis NI, Yurovsky A, Lappalainen T, Romano-Palumbo L, Planchon A, Bielser D, Bryois J, Padioleau I, Udin G, Thurnheer S, Hacker D, Core LJ, Lis JT, Hernandez N, Reymond A, Deplancke B, Dermitzakis ET. Science 342: 744-7.

Kim, T.K., and Maniatis, T. (1997). The Mechanism of Transcriptional Synergy of an In Vitro Assembled Interferon-β Enhanceosome. Molecular Cell *1*, 119-129.

Kim HB, Kong M, Kim TM, Suh YH, Kim WH, Lim JH, Song JH, Jung MH.: NFATc4 and ATF3 negatively regulate adiponectin gene expression in 3T3–L1 adipocytes. *Diabetes* 55: 1342– 1352, 2006.

Kim, U., Qin, F.-F., Gong, S., Stevens, S., Luo, Y., Nussenzweig, M. and Roeder, R.G. (1996) The B-cell-specific transcription coactivator OCA-B/OBF-1/Bob-1 is essential for normal production of immunoglobulin isotypes. *Nature*, **383**, 542-547.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107-116.

Kita A, Yamasaki H, Kuwahara H, Moriuchi A, Fukushima K, Kobayashi M, Fukushima T, Takahashi R, Abiru N, Uotani S, Kawasaki E, Eguchi K.: Identification of the promoter region required for human adiponectin gene transcription: association with CCAAT/enhancer binding protein-beta and tumor necrosis factor-alpha. *Biochem Biophys Res Comm* 331: 484–490, 2005.

König, H., Pfisterer, P., Corcoran, L. and Wirth, T. (1995) Identification of CD36 as the first gene dependent on the B cell differentiation factor Oct2. *Genes Dev.*, **9**, 1598-1607.

Kumada M, Kihara S, Sumitsuji S, Kawamoto T, Matsumoto S, Ouchi N, Arita Y, Okamoto Y, Shimomura I, Hiraoka H, Nakamura T, Funahashi T, Matsuzawa Y.: Association of hypoadiponectinemia with coronary artery disease in men. *Arterioscler Thromb Vasc Biol* 23: 85– 89, 2003.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lander, E.S., and Schork, N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037-2048.

Laumen H, Skurk T, Hauner H. (2008). The HMG-CoA reductase inhibitor rosuvastatin inhibits plasminogen activator inhibitor-1 expression and secretion in human adipocytes. *Atherosclerosis* 196: 565– 573.

Lenhard, B., Sandelin, A., Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics 13, 233-245.

Li-Weber, M., Salgame, P., Hu, C., Davydov, I.V., Laur, O., Klevenz, S. and Krammer, P.H. (1998) Th2-specific protein/DNA interactions at the proximal Nuclear Factor-AT site contribute to the functional activity of the human IL-4 promoter. *J.Immunol.*, **161**, 1380-1389.

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.

Lins, K., Remenyi, A., Tomilin, A., Massa, S., Wilmanns, M., Matthias, P. and Scholer, H.R. (2003) OBF1 enhances transcriptional potential of Oct1. *EMBO J*, **22**, 2188-2198.

Liu, Q., Chen, Y., Auger-Messier, M., Molkentin, JD. (2012). Interaction between NFκB and NFAT coordinates cardiac hypertrophy and pathological remodeling. Circ Res. 110(8):1077-86.

Loewel H, Doering A, Schneider A, Heier M, Thorand B, Meisinger C.: The MONICA Augsburg surveys: basis for prospective cohort studies. *Gesundheitswesen* 67: 13– 18, 2005.

Ludwig, M.Z., Bergman, C., Patel, N.H., and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. Nature *403*, 564-567.

Luo, Y., Fujii, H., Gerster, T. and Roeder, R.G. (1992) A novel B cell-derived coactivator potentiates the activation of immunoglobulin promoters by octamer-binding transcription factors. *Cell*, **71**, 231-241.

Luo, Y. and Roeder, R.G. (1995) Cloning, functional characterization, and mechanism of action of the B-cell-specific transcriptional coactivator OCA-B. *Mol.Cell.Biol.*, **15**, 4115-4124.

Macian, F. and Rao, A. (1999) Reciprocal modulatory interaction between human immunodeficiency virus type 1 Tat and transcription factor NFAT1. *Mol Cell Biol*, **19**, 3645-3653.

Marson, A., Kretschmer, K., Frampton, G.M., Jacobsen, E.S., Polansky, J.K., MacIsaac, K.D., Levine, S.S., Fraenkel, E., von Boehmer, H. and Young, R.A. (2007) Foxp3 occupancy and regulation of key target genes during T-cell stimulation. *Nature*, **445**, 931-935.

Massa, S., Junker, S., Schubart, K., Matthias, G. and Matthias, P. (2003) The OBF-1 gene locus confers B cell-specific transcription by restricting the ubiquitous activity of its promoter. *Eur J Immunol*, **33**, 2864-2874.

Matsuzawa Y.: The metabolic syndrome and adipocytokines. *FEBS Lett* 580: 2917– 2921, 2006.

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

Maurano, M.T., Wang, H., Kutyavin, T., and Stamatoyannopoulos, J.A. (2012b). Widespread Site-Dependent Buffering of Human Regulatory Polymorphism. PLoS Genet *8*, e1002599EP.

McClellan, J., and King, M.-C. (2010). Genetic Heterogeneity in Human Disease. Cell *141*, 210-217.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature 464, 773–777.

Moriuchi, M. and Moriuchi, H. (2001) Octamer transcription factors up-regulate the expression of CCR5, a coreceptor for HIV-1 entry. *J Biol Chem*, **276**, 8639-8642.

Musri MM, Corominola H, Casamitjana R, Gomis R, Parrizas M.: Histone H3 lysine 4 dimethylation signals the transcriptional competence of the adiponectin promoter in preadipocytes. *J Biol Chem* 281: 17180– 17188, 2006.

Natoli, G., Saccani, S., Bosisio, D. and Marazzi, I. (2005) Interactions of NF-kappaB with chromatin: the art of being at the right place at the right time. *Nat Immunol*, **6**, 439-445.

Nayak, A., Glockner-Pagel, J., Vaeth, M., Schumann, J.E., Buttmann, M., Bopp, T., Schmitt, E., Serfling, E. and Berberich-Siebelt, F. (2009) Sumoylation of the transcription factor

NFATc1 leads to its subnuclear relocalization and interleukin-2 repression by histone deacetylase. *J Biol Chem*, **284**, 10935-10946.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814.

Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., and Small, K., et al. (2011). The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS Genet *7*, e1002003.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS. PLoS Genet *6*, e1000888.

Nielsen, P.J., Georgiev, O., Lorenz, B. and Schaffner, W. (1996) B lymphocytes are impaired in mice lacking the transcriptional co-activator Bob1/OCA-B/OBF1. *Eur. J. Immunol.*, **26**, 3214-3218.

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

Ouedraogo R, Gong Y, Berzins B, Wu X, Mahadev K, Hough K, Chan L, Goldstein BJ, Scalia R.: Adiponectin deficiency increases leukocyte-endothelium interactions via upregulation of endothelial cell adhesion molecules in vivo. *J Clin Invest* 117: 1718– 1726, 2007.

Paigen, K., and Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. Nat Rev Genet *11*, 221-233.

Palkowitsch, L., Marienfeld, U., Brunner, C., Eitelhuber, A., Krappmann, D. and Marienfeld, R.B. (2011) The Ca2+-dependent phosphatase calcineurin controls the formation of the Carma1-Bcl10-Malt1 complex during T cell receptor-induced NF-kappaB activation. *J Biol Chem*, **286**, 7522-7534.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669-680.

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

Pfeuffer, I., Klein-Heßling, S., Heinfling, A., Chuvpilo, S., Escher, C., Brabletz, T., Hentsch, B., Schwarzenbach, H., Matthias, P. and Serfling, E. (1994) Octamer factors exert a dual effect on the IL-2 and IL-4 promoters. *J.Immunol.*, **153**, 5572-5585.

Pfisterer, P., Zwilling, S., Hess, J. and Wirth, T. (1995) Functional characterization of the murine homolog of the B-cell-specific coactivator BOB.1/OBF.1. *J.Biol.Chem.*, **270**, 29870-29880.

Pierani, A., Heguy, A., Fujii, H. and Roeder, R.G. (1990) Activation of octamer-containing promoters by either octamer-binding transcription factor 1 (OTF-1) or OTF-2 and requirement of an additional B-cell-specific component for optimal transcription of immunoglobulin promoters. *Mol. Cell. Biol.*, **10**, 6204-6215.

Prokopenko, I., Langenberg, C., Florez, J.C., Saxena, R., Soranzo, N., Thorleifsson, G., Loos, R.J.F., Manning, A.K., Jackson, A.U., and Aulchenko, Y., et al. (2009). Variants in MTNR1B influence fasting glucose levels. Nat Genet *41*, 77-81.

Qiao L, MacLean PS, Schaack J, Orlicky DJ, Darimont C, Pagliassotti M, Friedman JE, Shao J.: C/EBP-alpha regulates human adiponectin gene transcription through an intronic enhancer. *Diabetes* 54: 1744– 1754, 2005.

Qin, X.F., Reichlin, A., Luo, Y., Roeder, R.G. and Nussenzweig, M.C. (1998) OCA-B integrates B cell antigen receptor-, CD40L- and IL 4-mediated signals for the germinal center pathway of B cell development. *EMBO J.*, **17**, 5066-5075.

Rathmann W, Haastert B, Herder C, Hauner H, Koenig W, Meisinger C, Holle R, Giani G. (2006): Differential association of adiponectin with cardiovascular risk markers in men and women? The KORA survey 2000. *Int J Obes* 31: 770– 776.

Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, et al. (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*;140:744-752.

Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., Richter, D.J., Lander, E.S., and Altshuler, D. (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat Genet 32*, 135-142.

Riordan, J.R., Rommens, J.M., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., and Chou, J.L. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science 245*, 1066-1073.

Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., and Datta, L.W., et al. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet 39*, 596-604.

Risch, N., and Merikangas, K. (1996). The Future of Genetic Studies of Complex Human Diseases. *Science 273*, 1516-1517.

Samardzic, T., Marinkovic, D., Nielsen, P.J., Nitschke, L. and Wirth, T. (2002) BOB.1/OBF.1 deficiency affects marginal-zone B-cell compartment. *Mol Cell Biol*, **22**, 8320-8331.

Sauter, P. and Matthias, P. (1997) The B cell-specific coactivator OBF-1 (OCA-B, Bob-1) is inducible in T cells and its expression is dispensable for IL-2 gene induction. *Immunobiology*, **198**, 207-216.

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.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., and Mackay, S., et al. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science 328*, 1036-1040.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA.: Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70: 425– 434, 2002.

Schorpp M, Mattei MG, Herr I, Gack S, Schaper J, Angel P. (1995): Structural organization and chromosomal localization of the mouse collagenase type I gene. *Biochem J* 308: 211– 217.

Schubart, D.B., Rolink, A., Kosco-Vilbois, M.H., Botteri, F. and Matthias, P. (1996) B-cell-specific coactivator OBF-1/OCA-B/Bob1 required for immune response and germinal centre formation. *Nature*, **383**, 538-542.

Schuh, K., Kneitz, B., Heyer, J., Bommhardt, U., Jankevics, E., Berberich-Siebelt, F., Pfeffer, K., Muller-Hermelink, H.K., Schimpl, A. and Serfling, E. (1998) Retarded thymic involution and massive germinal center formation in NF-ATp-deficient mice. *Eur J Immunol*, **28**, 2456-2466.

Schwarz PE, Govindarajalu S, Towers W, Schwanebeck U, Fischer S, Vasseur F, Bornstein SR, Schulze J.: Haplotypes in the promoter region of the ADIPOQ gene are associated with increased diabetes risk in a German Caucasian population. *Horm Metab Res* 38: 447– 451, 2006.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., and Jackson, A.U., et al. (2007). A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. Science *316*, 1341-1345.

Serfling, E., Berberich-Siebelt, F., Avots, A., Chuvpilo, S., Klein-Hessling, S., Jha, M.K., Kondo, E., Pagel, P., Schulze-Luehrmann, J. and Palmetshofer, A. (2004) NFAT and NF-kappaB factors-the distant relatives. *Int J Biochem Cell Biol*, **36**, 1166-1170.

Serfling, E., Chuvpilo, S., Liu, J., Hofer, T. and Palmetshofer, A. (2006) NFATc1 autoregulation: a crucial step for cell-fate determination. *Trends Immunol*, **27**, 461-469.

Shaw, J., Sicree, R., and Zimmet, P. (2010). Global estimates of the prevalence of diabetes for 2010 and 2030. Diabetes Research and Clinical Practice *87*, 4-14.

Shen, Y. and Hendershot, L.M. (2007) Identification of ERdj3 and OBF-1/BOB-1/OCA-B as direct targets of XBP-1 during plasma cell differentiation. *J Immunol*, **179**, 2969-2978.

Shibuya, H. and Taniguchi, T. (1989) Identification of multiple cis-elements and trans-acting factors involved in the induced expression of human IL-2 gene. *Nucleic Acids Res*, **17**, 9173-9184.

Shore, P., Dietrich, W. and Corcoran, L.M. (2002) Oct-2 regulates CD36 gene expression via a consensus octamer, which excludes the co-activator OBF-1. *Nucleic Acids Res*, **30**, 1767-1773.

Sosinsky, A., Honig, B., Mann, R.S., and Califano, A. (2007). Discovering transcriptional regulatory regions in Drosophila by a nonalignment method for phylogenetic footprinting. Proceedings of the National Academy of Sciences *104*, 6305-6310.

Staudt, L.M., Singh, H., Sen, R., Wirth, T., Sharp, P.A. and Baltimore, D. (1986) A lymphoid-specific protein binding to the octamer motif of immunoglobulin genes. *Nature*, **323**, 640-643.

Stefan N, Stumvoll M, Vozarova B, Weyer C, Funahashi T, Matsuzawa Y, Bogardus C, Tataranni PA.: Plasma adiponectin and endogenous glucose production in humans. *Diabetes Care* 26: 3315– 3319, 2003.

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

Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., and Koller, D., et al. (2007). Population genomics of human gene expression. Nat.Genet *39*, 1217-1224.

Strubin, M., Newell, J.W. and Matthias, P. (1995) OBF-1, a novel B cell-specific coactivator that stimulates immunoglobulin promoter activity through association with octamer proteins. *Cell*, **80**, 497-506.

Sukiennicki, T.L. and Fowell, D.J. (2006) Distinct molecular program imposed on CD4+ T cell targets by CD4+CD25+ regulatory T cells. *J Immunol*, **177**, 6952-6961.

Taher, L., McGaughey, D.M., Maragh, S., Aneas, I., Bessling, S.L., Miller, W., Nobrega, M.A., McCallion, A.S., and Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. Genome Research *21*, 1139-1149.

Tan, Y., Rouse, J., Zhang, A., Cariati, S., Cohen, P. and Comb, M.J. (1996) FGF and stress regulate CREB and ATF-1 via a pathway involving p38 MAP kinase and MAPKAP kinase-2. *EMBO J*, **15**, 4629-4642.

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

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

The International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Thorand B, Kolb H, Baumert J, Koenig W, Chambless L, Meisinger C, Illig T, Martin S, Herder C.: Elevated levels of interleukin-18 predict the development of type 2 diabetes: results from the MONICA/KORA Augsburg Study, 1984–2002. *Diabetes* 54: 2932– 2938, 2005.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., and Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet *39*, 31-40.

Tomilin, A., Remenyi, A., Lins, K., Bak, H., Leidel, S., Vriend, G., Wilmanns, M. and Scholer, H.R. (2000) Synergism with the coactivator OBF-1 (OCA-B, BOB-1) is mediated by a specific POU dimer configuration. *Cell*, **103**, 853-864.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. Nat Genet *45*, 124-130.

Tschritter O, Fritsche A, Thamer C, Haap M, Shirkavand F, Rahe S, Staiger H, Maerker E, Haring H, Stumvoll M.: Plasma adiponectin concentrations predict insulin sensitivity of both glucose and lipid metabolism. *Diabetes* 52: 239– 243, 2003.

Vasseur F, Helbecque N, Dina C, Lobbens S, Delannoy V, Gaget S, Boutin P, Vaxillaire M, Lepretre F, Dupont S, Hara K, Clement K, Bihain B, Kadowaki T, Froguel P.: Single-nucleotide polymorphism haplotypes in the both proximal promoter and exon 3 of the APM1 gene modulate adipocyte-secreted adiponectin hormone levels and contribute to the genetic risk for type 2 diabetes in French Caucasians. *Hum Mol Genet* 11: 2607– 2614, 2002.

Vasseur F, Helbecque N, Lobbens S, Vasseur-Delannoy V, Dina C, Clement K, Boutin P, Kadowaki T, Scherer PE, Froguel P.: Hypoadiponectinaemia and high risk of type 2 diabetes

are associated with adiponectin-encoding (ACDC) gene promoter variants in morbid obesity: evidence for a role of ACDC in diabesity. *Diabetologia* 48: 892– 899, 2005.

Venter, J.C. (2001). The Sequence of the Human Genome. Science *291*, 1304-1351.

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

Vionnet N, Hani EH, Dupont S, Gallina S, Francke S, Dotte S, De Matos F, Durand E, Lepretre F, Lecoeur C, Gallina P, Zekiri L, Dina C, Froguel P.: Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21–q24. *Am J Hum Genet* 67: 1470– 1480, 2000.

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

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

Wagner, G.P., and Zhang, J. (2011). The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. Nat Rev Genet *12*, 204-213.

Wang, V.E., Schmidt, T., Chen, J., Sharp, P.A. and Tantin, D. (2004) Embryonic lethality, decreased erythropoiesis, and defective octamer-dependent promoter activation in Oct-1-deficient mice. *Mol Cell Biol*, **24**, 1022-1032.

Wang, V.E., Tantin, D., Chen, J. and Sharp, P.A. (2004) B cell development and immunoglobulin transcription in Oct-1-deficient mice. *Proc Natl Acad Sci U S A*, **101**, 2005-2010.

Ward, L.D., and Kellis, M. (2011). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Research.

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

Weber, K.S., Miller, M.J. and Allen, P.M. (2008) Th17 cells exhibit a distinct calcium profile from Th1 and Th2 cells and have Th1-like motility and NF-AT nuclear localization. *J Immunol*, **180**, 1442-1450.

Weidinger S, Klopp N, Wagenpfeil S, Rummler L, Schedel M, Kabesch M, Schafer T, Darsow U, Jakob T, Behrendt H, Wichmann HE, Ring J, Illig T.: Association of a STAT 6 haplotype with elevated serum IgE levels in a population based cohort of white adults. *J Med Genet* 41: 658– 663, 2004.

Wilson NK, Foster SD, Wang X, Knezevic K, Schutte J, Kaimakis P, Chilarska PM, Kinston S, Ouwehand WH, Dzierzak E, et al. (2010) Combinatorial transcriptional control in blood stem/progenitor cells: genome-wide analysis of ten major transcriptional regulators. *Cell Stem Cell*;7:532-544.

Wolfe, S.A., Zhou, P., Dotsch, V., Chen, L., You, A., Ho, S.N., Crabtree, G.R., Wagner, G. and Verdine, G.L. (1997) Unusual Rel-like architecture in the DNA-binding domain of the transcription factor NFATc. *Nature*, **385**, 172-176.

Wu, C.C., Hsu, S.C., Shih, H.M. and Lai, M.Z. (2003) Nuclear factor of activated T cells c is a target of p38 mitogen-activated protein kinase in T cells. *Mol Cell Biol*, **23**, 6442-6454.

Yamamoto Y, Hirose H, Saito I, Tomita M, Taniyama M, Matsubara K, Okazaki Y, Ishii T, Nishikai K, Saruta T.: Correlation of the adipocyte-derived protein adiponectin with insulin resistance index and serum high-density lipoprotein-cholesterol, independent of body mass index, in the Japanese population. *Clin Sci* (*Lond*) 103: 137– 142, 2002.

Yamauchi T, Kamon J, Waki H, Terauchi Y, Kubota N, Hara K, Mori Y, Ide T, Murakami K, Tsuboyama-Kasaoka N, Ezaki O, Akanuma Y, Gavrilova O, Vinson C, Reitman ML, Kagechika H, Shudo K, Yoda M, Nakano Y, Tobe K, Nagai R, Kimura S, Tomita M, Froguel P, Kadowaki T.: The fat-derived hormone adiponectin reverses insulin resistance associated with both lipoatrophy and obesity. *Nat Med* 7: 941– 946, 2001.

Yanez-Cuna, J.O., Dinh, H.Q., Kvon, E.Z., Shlyueva, D., and Stark, A. (2012). Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. Genome Research *22*, 2018-2030.

Yáñez-Cuna, J.O., Kvon, E.Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. Trends in Genetics *29*, 11-22.

Yang WS, Jeng CY, Wu TJ, Tanaka S, Funahashi T, Matsuzawa Y, Wang JP, Chen CL, Tai TY, Chuang LM.: Synthetic peroxisome proliferator-activated receptor-gamma agonist, rosiglitazone, increases plasma levels of adiponectin in type 2 diabetic patients. *Diabetes Care* 25: 376– 380, 2002.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., Bakker, P.I.W. de, Abecasis, G.R., Almgren, P., and Andersen, G., et al. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet *40*, 638-645.

Zabel, U., Schreck, R. and Baeuerle, P.A. (1991) DNA binding of purified transcription factor NF-kappa B. Affinity, specificity, Zn2+ dependence, and differential half-site recognition. *J Biol Chem*, **266**, 252-260.

Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R.B., Rayner, N.W., and Freathy, R.M., et al. (2007). Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. Science *316*, 1336-1341.

Zelivianski, S., Glowacki, R. and Lin, M.F. (2004) Transcriptional activation of the human prostatic acid phosphatase gene by NF-kappaB via a novel hexanucleotide-binding site. *Nucleic Acids Res*, **32**, 3566-3580.

Zhang D, Ma J, Brismar K, Efendic S, Gu HF.: A single nucleotide polymorphism alters the sequence of SP1 binding site in the adiponectin promoter region and is associated with diabetic nephropathy among type 1 diabetic patients in the Genetics of Kidneys in Diabetes Study. *J Diabetes Complications* Epub ahead: doi:10.1016/j.jdiacomp 2008.

Zhang, P., Zhang, X., Brown, J., Vistisen, D., Sicree, R., Shaw, J., and Nichols, G. (2010). Global healthcare expenditure on diabetes for 2010 and 2030. Diabetes Research and Clinical Practice *87*, 293-301.

Zheng, L., Roeder, R.G. and Luo, Y. (2003) S phase activation of the histone H2B promoter by OCA-S, a coactivator complex that contains GAPDH as a key component. *Cell*, **114**, 255-266.

Zuckerkandl, E., and Pauling, L. (1965). Molecules as documents of evolutionary history. J Theor Biol *8*, 357-366.

Zwilling, S., Dieckmann, A., Pfisterer, P., Angel, P. and Wirth, T. (1997) Inducible expression and phosphorylation of coactivator BOB.1/OBF.1 in T cells. *Science*, **277**, 221-22.

5 Publications

1. **Claussnitzer M,** Dankel SN , Klocke B, Grallert H, Glunk V, Berulava T, Lee H, Oskolkov N, Fadista J, Ehlers K, Wahl S, Hoffmann C, Qian K, Rönn T, Riess L, Müller-Nurasyid M, Skurk T, Bretschneider N, Schroeder T, Skurk T, Horsthemke B, DIAGRAM+ Consortium, Spieler D, Klingenspor M, Seifert M, Kern MJ, Mejhert N, Dahlman I, Hansson O, Hauck S, Blüher, Arner P, Groop L, Illig T, Suhre K, Hsu YH, Mellgren G, Hauner H, Laumen H. **Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms.** In Press, Cell.
2. **Claussnitzer M**, Laumen H, Hauner H., **COMPUTER IMPLEMENTED METHOD FOR IDENTIFYING REGULATORY REGIONS OR REGULATORY VARIATIONS,** published patent. Patent file at Claussnitzer 2013-1.
3. **Claussnitzer M**, Laumen H, Hauner H., '**DIAGNOSTIC MEANS AND METHODS FOR TYPE 2 DIABETES**'. The patent file at Claussnitzer 2013-2.
4. **Claussnitzer M**, Skurk T, Hauner H, Daniel H, Rist MJ. Effect of flavonoids on basal and insulin-stimulated -deoxyglucose uptake in adipocytes. Mol Nutr Food Res. 2011 Suppl 1:S26-34.. View in: PubMed
5. Laumen H, Saningong AD, Heid IM, Hess J, Herder C, **Claussnitzer M**, Baumert J, Lamina C, Rathmann W, Sedlmeier EM, Klopp N, Thorand B, Wichmann HE, Illig T, Hauner H. **Functional characterization of promoter variants of the adiponectin gene complemented by epidemiological data**. Diabetes. 2009 Apr; 58(4):984-9. View in: PubMed

6. Mueller K, Quandt J, Marienfeld RB, Weihrich P, Fiedler K, **Claussnitzer M**, Laumen H, Vaeth M, Berberich-Siebelt F, Serfling E, Wirth T, Brunner C. **Octamer-dependent transcription in T cells is mediated by NFAT and NF-κB**. Nucleic Acids Res. 2013 Feb ; 4(4):238-54. View in: PubMed

7. Then C, Wahl S, Kirchhofer A, Grallert H, Krug S, Kastenmüller G, Römisch-Margl W, **Claussnitzer M**, Illig T, Heier M, Meisinger C, Adamski J, Thorand B, Huth C, Peters A, Prehn C, Heukamp I, Laumen H, Lechner A, Hauner H, Seissler J. **Plasma Metabolomics Reveal Alterations of Sphingo- and Glycerophospholipid Levels in Non-Diabetic Carriers of the Transcription Factor 7-Like 2 Polymorphism rs7903146.** PLoS One. 2013; 8(10): e78430. doi: 10.1371/journal.pone.0078430.

8. Simon N. Dankel, Dag J. Fadnes, Isin Dalkilic-Liddle, Andrea Christoforou, Iain Mathieson, Vivian L. Veum, Oddrun A. Gudbrandsen, Christine Haugen, Margit H. Solsvik, Villy Våge, Hans Jørgen Nielsen, Yong-Jun Liu, Yu-Fang Pei, Hong-Wen Deng, André Scherag, Anke Hinney, Johannes Hebebrand, Philippe Froguel, David Meyre, GIANT Consortium, Brian Seed, Hans Hauner, Bernward Klocke, Helmut Laumen, Cecilia M. Lindgren, **Melina Claussnitzer**, Jørn V. Sagen, Vidar M. Steen, Gunnar Mellgren. **Integrative functional genomics links homeobox transcription factors to visceral adiposity.** Under Review.

6 APPENDIX

6.1 INVENTORY of Supplemental Information

Claussnitzer et al. *Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms*

SUPPLEMENTAL FIGURES

Figure S1. Linkage disequilibrium (LD) block structure at eight T2D susceptibility loci included into the proof-of-concept analysis and cell-type specific *cis*-regulatory effects of SNPs in complex regions at T2D risk loci. Related to Figure 1.

Figure S2. Distance of predicted *cis*-regulatory SNPs to transcriptional start site; frequency distribution of complex regions; and correlations of *cis*-regulatory predictions from PMCA at Crohn's disease susceptibility loci with evolutionary constraint elements and functionally annotated genomic regions. Related to Figure 2.

Figure S3. Performance of PMCA at T2D, asthma and Crohn's disease susceptibility loci; positional bias analysis of TFBS matrices for Crohn`s-associated loci; and association of homeobox TFs – inferred from combinatorial framework analysis of PMCA and RNAseq - with metabolic processes and impaired glucose stimulated insulin secretion. Related to Figure 3.

Figure S4. Figure S4. Variants in complex regions at the *PPARG* T2D risk locus and the homeobox factor PRRX1 as a rs4684847-dependent repressor of endogenous *PPARG2* expression. Related to Figures 4-5 and Table 2.

SUPPLEMENTAL TABLES

Table S1. PMCA measures for candidate SNPs at eight T2D susceptibility loci included into the proof of concept analysis. Related to Figure 1.

Table S2. PMCA measures for candidate SNPs at eight T2D susceptibility loci included into the proof of concept analysis (upper 25% of complex region ranking). Related to Figure 1.

Table S3. PMCA measures and experimental validation of *cis*-regulatory predictions for candidate SNP regions. Related to Figure 1.

Table S4. Association of tagSNPs and predicted *cis*-regulatory SNPs with T2D in DIAGRAM v2 data and glycemic traits in MAGIC consortium meta-analysis data. Related to Figure 1.

Table S5. Reported GWAS loci for 19 diseases traits (A), PMCA measures of 2,045 candidate SNPs at the selected loci (B) and matched random variants (C).

Table S6. PMCA measures for known *cis*-regulatory SNPs, associated to different traits. Related to Figure 1.

Table S7. PMCA measures for candidate SNPs at 47 autosomal T2D susceptibility loci compromising 1,465 SNPs. Related to Figure 2 and 3.

Table S8. Overlap of complex regions and non-complex regions with evolutionary constraint elements and localization to next TSS. Related to Figure 2.

Table S9. Enrichment of functional annotation from DHSseq and ChIPseq peaks overlaps with complex regions. Related to Figure 2.

Table S10. Association of *cis*-regulatory predictions at PMCA selected complex regions with functional RegulomeDB annotations. Related to Figure 2.

Table S11. Transcription factors, TFBS matrices and TFBS matrix families. Related to Figure 3.

Table S12. Positional bias analysis of TFBS matrix families in complex regions and non-complex regions. Related to Figure 3.

Table S13. PMCA measures for candidate SNPs at asthma susceptibility loci and for candidate SNPs at Crohn's disease susceptibility loci. Related to Figure 3.

Table S14. T2D-related homeobox TFs identified by a combinatorial analysis of PMCA and RNAseq expression data in islets from normoglycemic versus T2D subjects. Related to Figure 3.

Table S15. Co-expression of T2D-related homeobox TFs with all transcripts from RNAseq in pancreatic islets from 51 healthy subjects. Related to Figure 3.

Table S16. Co-expression of T2D-related homeobox TFs with all transcripts from RNAseq in pancreatic islets from 26 subjects at risk of diabetes. Related to Figure 3.

Table S17. Experimental validation of PMCA predicted *cis*-regulatory variants at the PPARG T2D risk locus, 3p25.3. Related to Figure 4.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

13. Electrophoretic mobility shift assay (EMSA)
14. DNA-Protein affinity chromatography, LC-MS/MS and label free quantification.
15. Genome editing of SGBS preadipocytes
16. Analysis of human tissue samples
17. Analysis of RNAseq data from primary human islets
18. eQTL analysis
19. Isolation, culture and differentiation of primary human adipose stromal cells (hASC)
20. Genotyping
21. Gene knock-down by siRNA
22. Quantitative RT-PCR and allele-specific primer extension analysis
23. Genome-wide expression analysis in primary human hASC
24. Assessment of lipid accumulation after PRRX1 overexpression
25. Glyceroneogenesis and 2-deoxyglucose uptake measurements in primary hASC
26. Statistical analysis

SUPPLEMENTAL NOTES

Authors from DIAGRAM+

SUPPLEMENTAL REFERENCES

6.2 SUPPLEMENTAL FIGURES

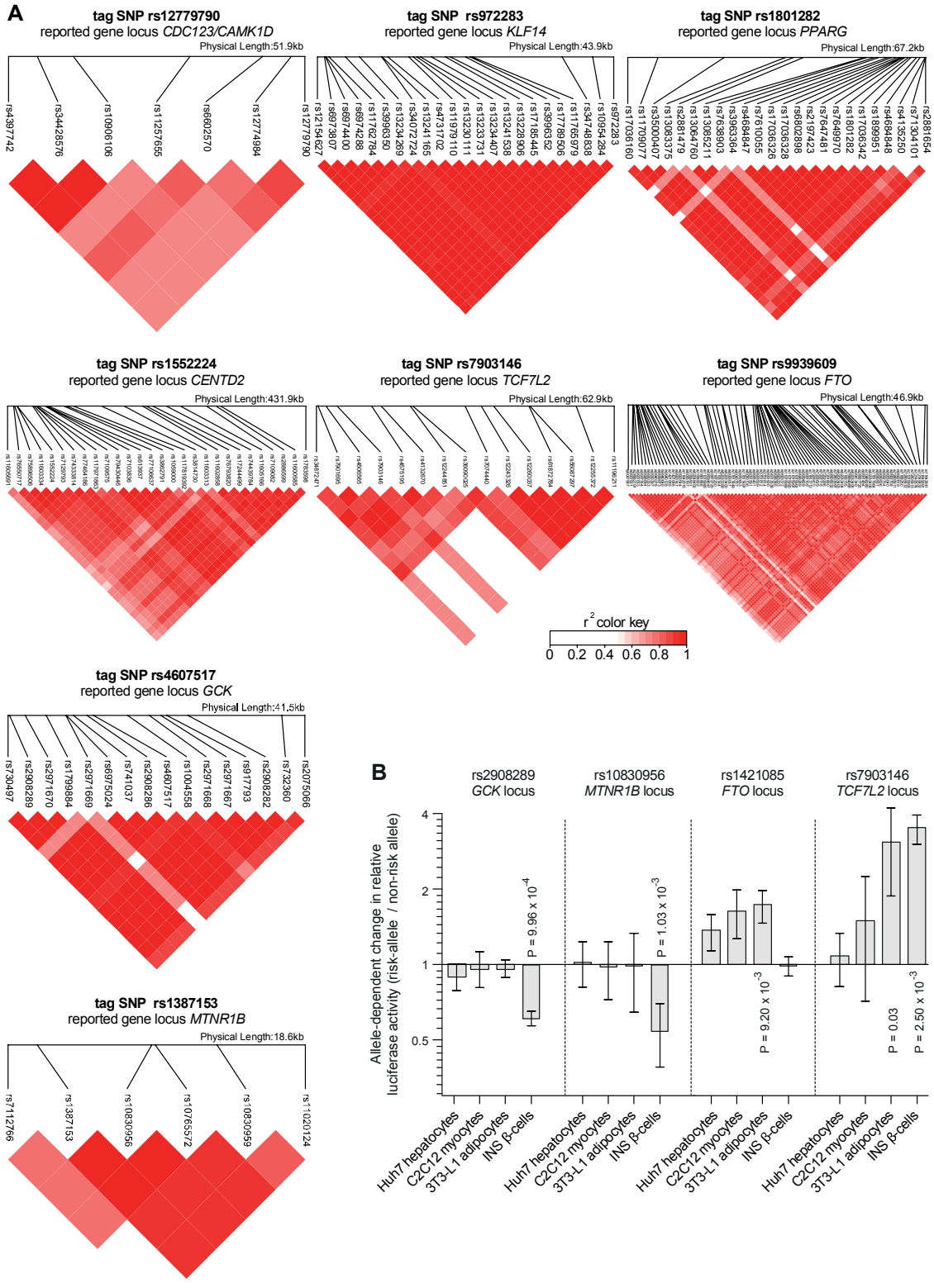

A
tag SNP rs12779790
reported gene locus CDC123/CAMK1D
Physical Length:51.9kb
tag SNP rs972283
reported gene locus KLF14
Physical Length:43.9kb
tag SNP rs1801282
reported gene locus PPARG
Physical Length:67.2kb
tag SNP rs1552224
reported gene locus CENTD2
Physical Length:431.9kb
tag SNP rs7903146
reported gene locus TCF7L2
Physical Length:62.9kb
tag SNP rs9939609
reported gene locus FTO
Physical Length:46.9kb
r2 color key
0 0.2 0.4 0.6 0.8 1
tag SNP rs4607517
reported gene locus GCK
Physical Length:41.5kb
tag SNP rs1387153
reported gene locus MTNR1B
Physical Length:18.6kb
B
rs2908289
GCK locus
rs10830956
MTNR1B locus
rs1421085
FTO locus
rs7903146
TCF7L2 locus
Allele-dependent change in relative luciferase activity (risk-allele / non-risk allele)
P = 9.96 x 10-4
P = 1.03 x 10-3
P = 9.20 x 10-3
P = 0.03
P = 2.50 x 10-3
Huh7 hepatocytes
C2C12 myocytes
3T3-L1 adipocytes
INS β-cells

Figure S1. Linkage disequilibrium (LD) block structure at eight T2D susceptibility loci included into the proof of concept analysis and cell-type specific *cis*-regulatory effects of complex regions at T2D loci. Related to Figure 1.

(A) LD blocks derived from eight tagSNPs 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.

(B) Cell type-specific *cis*-regulatory effects of SNPs located in complex regions. Luciferase constructs of the respective complex regions were transfected into INS-1 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 ± SD (n = 9). p-values were calculated by paired t-test.

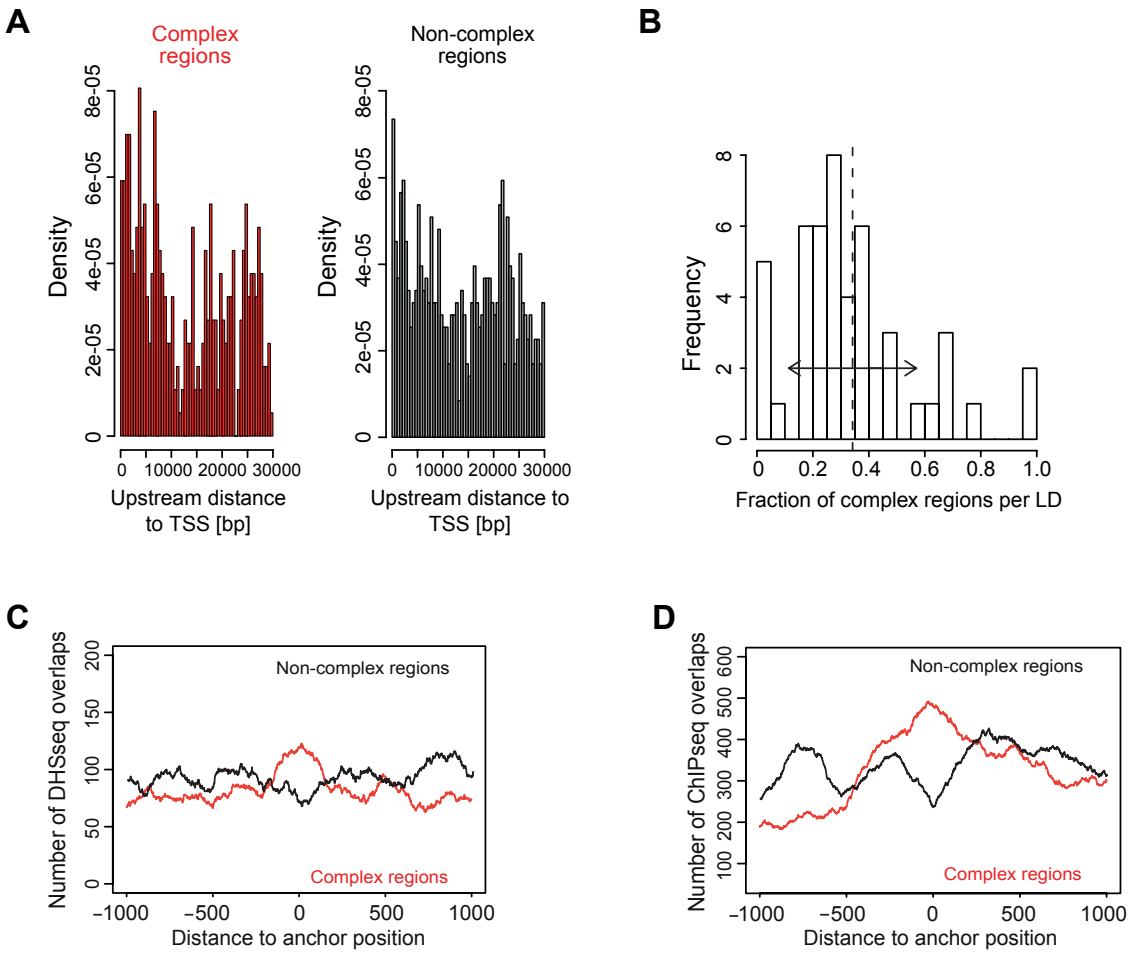


Figure S2. Distance to transcriptional start site, LD block frequency distribution and correlations with evolutionary constrained elements and functionally annotated genomic regions for PMCA-inferred *cis*-regulatory predictions. Related to Figure 2.

(A) 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 TSSs (TSS annotated within 30,000 bp downstream of SNP positions). The distance distribution is shown for 487 complex regions (left) and 978 non-complex regions (right) identified by PMCA within the set of 47 T2D loci (for detailed information see Table S7). The histogram shapes of the two diagrams illustrate the equal positioning of PMCA categories (complex and non-complex regions) relative to downstream TSSs.

(B) 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)).

(D-E) Correlations of PMCA results with DHSseq (A) and ChIPseq (B) data for 1,218 SNPs associated with Crohn's diseases. For PMCA classification of SNP-surrounding genomic regions in complex and non-complex regions see Table S13B. The occurrences of DHSseq and ChIPseq DNA peaks in vicinity of complex and non-complex Crohn's-associated SNP regions are shown (each position ± 500 bp from the SNP positions of complex and non-complex regions was scanned for overlaps with DHSseq or ChIPseq peaks, see Supplemental Experimental Procedures). The number of complex and non-complex regions that directly overlap DHSseq and ChIPseq regions was determined by a comparison of their genomic positions. Complex regions were significantly enriched for overlaps with DHSseq and ChIPseq 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 S9).

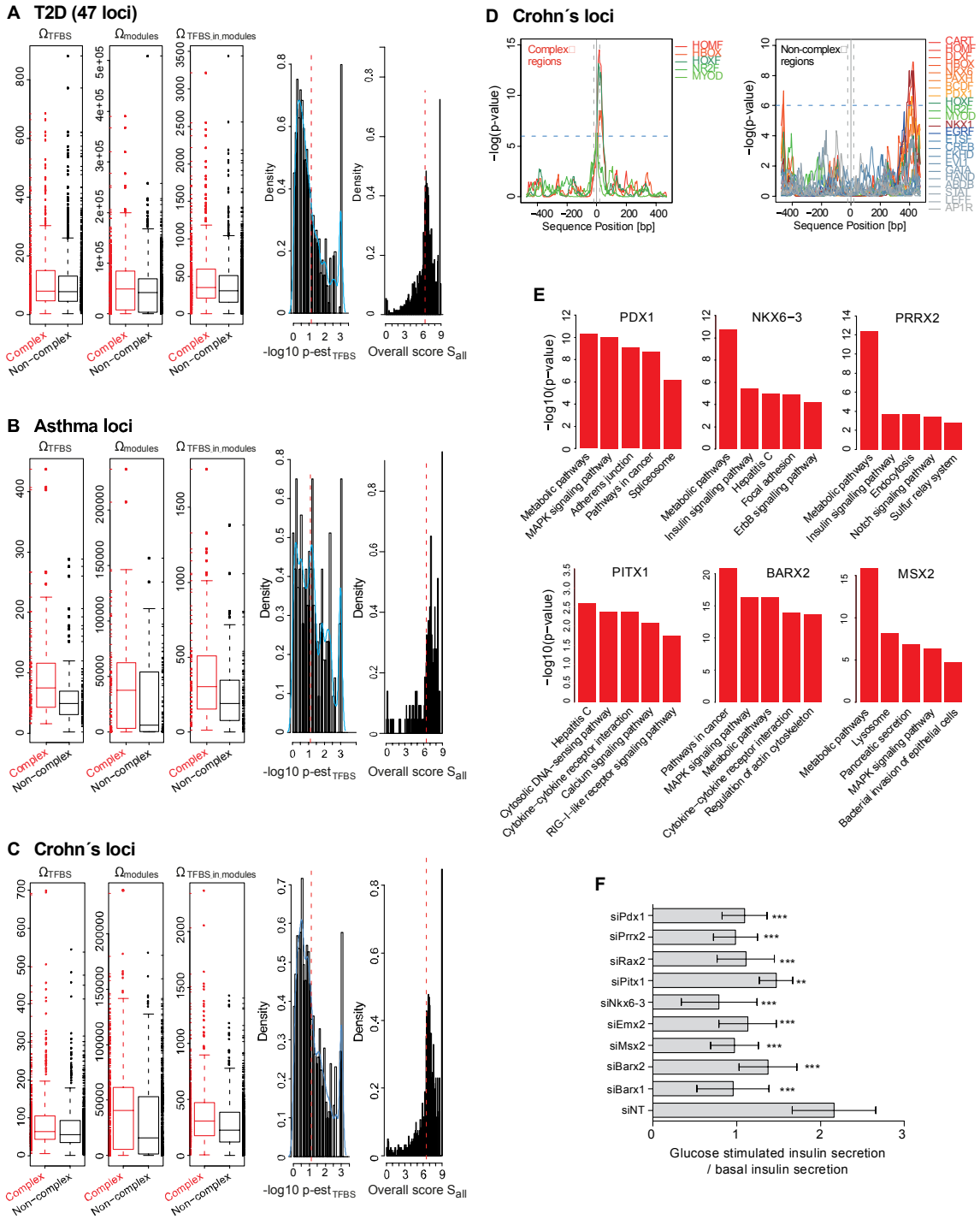


Figure S3. Performance of PMCA at T2D, asthma and Crohn's disease susceptibility loci; positional bias analysis of TFBS matrices for Crohn`s-associated loci; and association of homeobox TFs – inferred from combinatorial framework analysis of

PMCA and RNAseq - with metabolic processes and impaired glucose stimulated insulin secretion. Related to Figure 3.

(A-C) PMCA results are shown for 47 T2D (A), eight asthma (B) and eight Crohn's disease (C) susceptibility loci ($r^2 \geq 0.7$, Table S7). Box-whisker plots show the numbers obtained for each classification strategy (analysis for the complete set of input sequences). Plots show the distributions for Ω_{TFBS}, $\Omega_{modules}$ and $\Omega_{TFBS\_in\_modules}$, 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. Histograms illustrate the PMCA measures p-est$_{TFBS}$ and overall score S_{all} used for PMCA scoring. The histograms show the distribution of –log10 of the estimated probability p-est to randomly observe an equal or higher Ω_{TFBS} and the distribution for an equal or higher *overall score* S_{all} from all three criteria, 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.

(D) Positional bias analysis of TFBS matrices for a set of eight Crohn's disease susceptibility loci (1,218 candidate SNPs). Distribution of TFBS matrices relative to SNP positions (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 1,000 bp SNP regions for 192 TFBS matrix families (Genomatix Matrix Library version 8.4) within sliding 50 bp windows under a binomial distribution model (detailed in Supplemental 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 positions and –log10 (P) > 6 in the complex regions. The positional bias analysis within complex regions reveals specific clustering at SNP positions ± 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) Genes co-expressed with homeobox TFs - identified by a combinatorial framework analysis of PMCA and RNAseq expression data in islets - associate with metabolic pathways. PMCA applied on 47 T2D loci identified 487 complex regions (Table S7) and a distinct clustering of five homeobox TFBS matrix families (Figure 3B; see also Table S11 for TFBS matrices) at SNP positions in complex regions. Analysis of mRNA levels in RNAseq data of primary human islets, comparing donors with and without T2D, implicated nine homeobox TFs *RAX, PRRX2, BARX1, PITX1, EMX2, NKX6-3, BARX2, MSX2* and *PDX1* as candidate TFs in T2D pathophysiology (Table S14). Pathway analysis was performed for gene sets co-expressed with the identified nine homeobox TF in pancreatic islets from 51 donors without T2D. Gene sets were defined by correlating expression levels of all transcripts identified by RNAseq with the expression levels of the nine identified homeobox TFs (significantly co-expressed genes defined by FDR 5%, Table S15). The top five significantly enriched pathways (hypergeometric test, FDR 5%), inferred from WEBGESTALT analysis using the KEGG database, are shown.

(F) Glucose-stimulated insulin secretion in rat INS-1 β-cells transfected with non-targeting (NT) control siRNA and siRNAs targeting the nine homeobox TFs *BARX1, BARX2, MSX2, EMX2, NKX6-3, PITX1, RAX2, PRRX2* or *PDX1* that were identified by a combinatorial framework analysis integrating PMCA findings at 47 T2D-risk loci and RNAseq expression data in islets from normoglycemic *versus* T2D subjects (Table S14). Insulin levels in the medium after 1h incubation with high glucose or low glucose (basal) were measured by ELISA (Supplemental Experimental Procedures). The ratio of (glucose-stimulated insulin levels) / (basal insulin levels) was calculated for siNT control and for each homeobox TF siRNA. The experiments were performed in triplicate, and are presented as mean ± SD (n = 5). p-values were calculated by paired t-test.

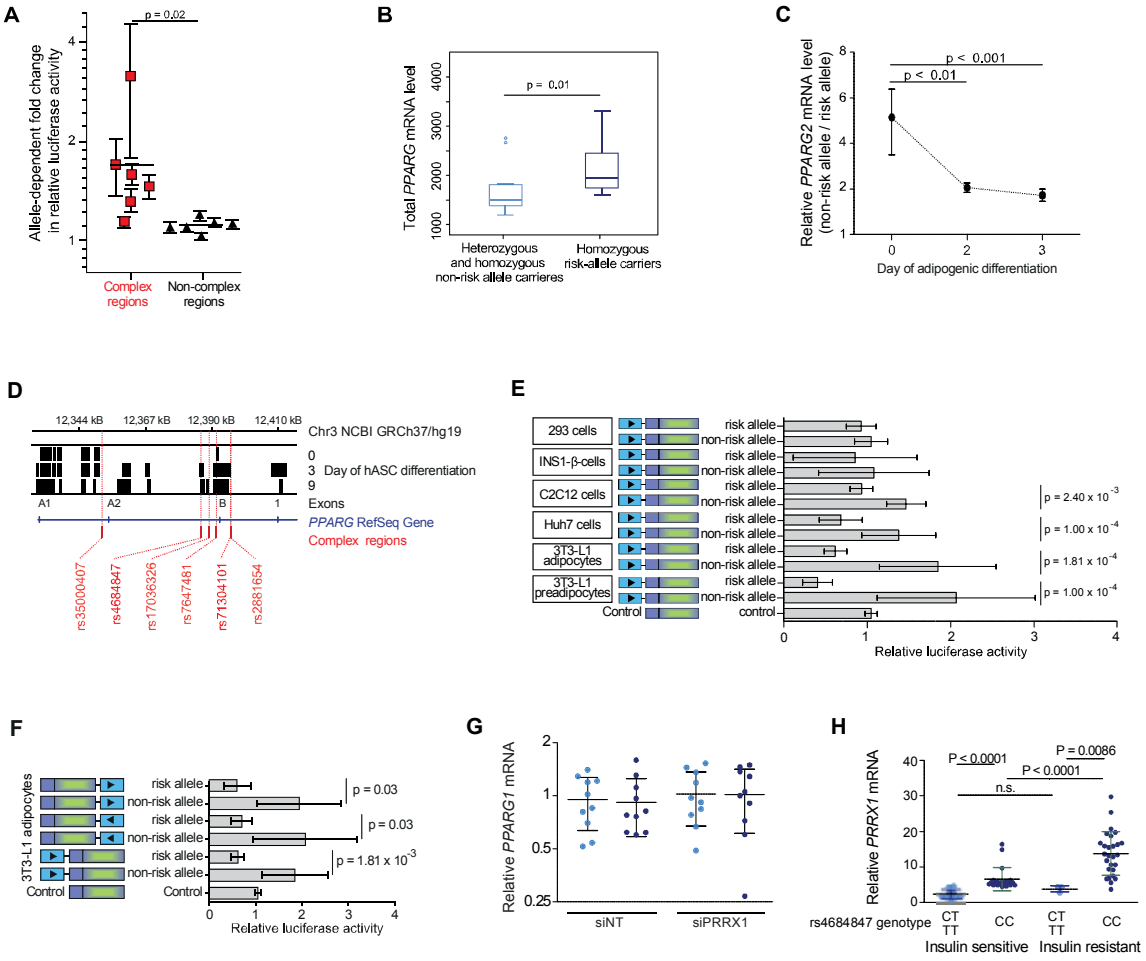


Figure S4. Variants in complex regions at the *PPARG* T2D risk locus and the homeobox factor PRRX1 as a rs4684847-dependent repressor of endogenous *PPARG2* expression. Related to Figures 4-5 and Table 2.

(A) Validation of *cis*-regulatory predictions at the *PPARG* T2D risk locus. *Cis*-regulatory predictions for variants in complex regions (red dots) were validated at the level of luciferase transcriptional activity (a random selection of variants in 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 change in luciferase expression comparing the risk/non-risk or non-risk/risk allele (change ≥ 1) is shown for each SNP as mean ± SD (n = 3-13). P-value was calculated by linear mixed-effects model. Details on the

analyzed SNPs are given in Table S17. Complex regions significantly differed from non-complex regions at the transcriptional level.

(B) Genotype-dependent increase in mRNA expression of total *PPARG* in human subcutaneous adipose tissue (n = 36). Box plots of the total *PPARG* 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) are shown. 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 (Affymetrix) and statistics were calculated by Wilcoxon signed rank test.

(C) Allelic imbalance of *PPARG2* mRNA expression levels during early stages of adipocyte differentiation measured in 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 ± SD. p-values were calculated by 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 and hASC three days and 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 localization of SNPs at complex regions at the *PPARG* locus are indicated, together with the *PPARG* exons A1, A2, the *PPARG2* specific exon B and the first exon of *PPARG1* and *PPARG2*. The complex region surrounding rs4684847 reveals cell stage-dependent H3K27ac marks (also H3K4me1, H3K4me2 and H3K36me3).

(E) Allele-dependent repression of reporter gene activity in 3T3-L1 adipocytes, Huh7 hepatocytes, C2C12 myocytes, INS-1 β-cells and 293 cells. Luciferase assays in 3T3-L1 adipocytes, Huh7 hepatoma cells, C2C12 muscle cells, INS-1 pancreatic β-cells and 293T cells reveal cell type-specific *cis*-regulatory activity of the complex region SNP rs4684847. 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 ± SD. p-values were calculated by paired t-test.

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

(G) Regulation of *PPARG1* mRNA expression in SGBS adipocytes with homozygous risk or non-risk allele introduced by the CRISPR/Cas9 genome editing approach. Cells were transfected with siPRRX1 and siNT concurrent with induction of differentiation. *PPARG2* mRNA was assessed by qPCR, standardized to *HPRT* mRNA. The data are presented as mean ± SD (n = 12). p-values were calculated by 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 adipose tissue was measured by qPCR, standardized to *HPRT* mRNA. Insulin sensitivity was measured by euglycemic hyperinsulinemic clamp. Data are presented as mean ± SD (n = 30 per group). p-values were calculated by unpaired t-test.

6.3 SUPPLEMENTAL TABLES

6.4 SUPPLEMENTAL EXPERIMENTAL PROCEDURES

6.4.1 Definition of LD blocks

TagSNPs were derived from reported GWAS loci (corresponding references are listed in Tables S1, S7 and S13A). For each tagSNP, LD blocks were defined based on 1000G Pilot 1 CEU data (1000 Genomes Project Consortium, 2010) ($r^2 \geq 0.7$, NCBI GRCh37/hg19) using the SNAP viewer tool (Johnson et al., 2008), Broad Institute. For Crohn's diseases susceptibility loci, a previously published SNP set (Schaub et al., 2012) was chosen for PMCA analysis of candidate SNPs at (Table S13B).

6.4.2 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 finch (Taeniopygia guttata)

6.4.3 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.**

Motivation

Bioinformatics approaches that reliably assess the regulatory role of specific genetic variants would be highly desirable (Cooper and Shendure, 2011). 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 TFBS 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 TFBS 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 detect and classify genetic regions that contain evolutionary conserved TFBS modules. In the following, we describe such a method, called phylogenetic module complexity analysis (PMCA).

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 tagSNP 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, $r^2 \geq 0.7$) with the tag SNP (for the analysis performed in this manuscript, see chapter *2. Definition of LD blocks* in the Supplemental Experimental Procedures; note that any set of sequence 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.

10. The flanking region (±60nt) of the non-coding SNP is extracted from the human genome;
11. 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;
12. TFBS are identified in the set of ortholog sequences using position weight matrices from the Genomatix library;
13. TFBS modules are identified in each ortholog sequence; TFBS modules are specifically defined as all two or more TFBS that occur in the same order and in a certain distance range in all or a subset of the input sequences.
14. Phylogenetically conserved TFBS (Ω_{TFBS}), TFBS modules ($\Omega_{modules}$), and occurrence of TFBS in TFBS modules ($\Omega_{TFBS\_in\_modules}$) are counted.
15. 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-linage specific TFBS). To account for this possibility, we also determine phylogenetically conserved TFBS with more restricted parameters ($\Omega_{restr\text{-}TFBS}$, details see below).
16. Steps 3-5 are repeated one thousand times using randomized input sequences to estimate the probability of observing a given Ω_{TFBS}, $\Omega_{restr\text{-}TFBS}$, $\Omega_{modules}$, and $\Omega_{TFBS\_in\_modules}$. 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 TFBS 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.

17. Based on the four weighed counts Ω_{TFBS}, $\Omega_{restr\text{-}TFBS}$, $\Omega_{modules}$, and $\Omega_{TFBS\_in\_modules}$ and the estimated background probability of observing these counts by chance, we determine an overall classification criterion S_{all}.
18. The overall classification criterion labels the input region as *complex* or *non-complex*. (*Note***:** steps (1-9) are detailed in the **pseudo code** on page 21-23 of the Supplemental Experimental Procedures).

10. To further select the variant with a function in disease, the overall disease-distinct clustering of TFBS at complex regions is assessed using positional bias analysis. (*Note*: the calculation of positional bias in step (10) is detailed in chapter 3 of the Extended Experiment Procedures, page 31-32).

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.

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 et al., 1998) for the identification of homologous regions, TRANSFAC (Matys et al., 2006) 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:

$\zeta = i$ */ 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 TFBS 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 $\zeta >= 0.5$ *sequence set to compute* $\Omega_{restr\text{-}TFBS}$

2.4 Repeat the following 1,000 times

Randomly shuffle the sequence set S; use a sliding window of 10 bp and permutate 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}, $\Omega_{restr\text{-}TFBS}^{rnd}$, $\Omega_{modules}^{rnd}$, *and* $\Omega_{TFBS\_in\_modules}^{rnd}$.

3. Scoring and classification

3.1 Estimate the probability $p\text{-}est_i = f(\Omega_i^{rnd} > \Omega_i)$ *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* $p\text{-}est_i$ *is set to 1/1001 if this never occurs:*

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

3.2 Compute an Overall-score $S_{all} = -log(p\text{-}est_{TFBS} * p\text{-}est_{modules} * p\text{-}est_{TFBS\_in\_modules})$

3.3 Classify a non-coding SNP as being located in a complex region if and only if:
$(S_{all} > 6.5)$ *and* $(p\text{-}est_{restr\text{-}TFBS} < 0.15)$ *and* $(p\text{-}est_{TFBS} < 0.075)$
(Scoring criteria for classification)

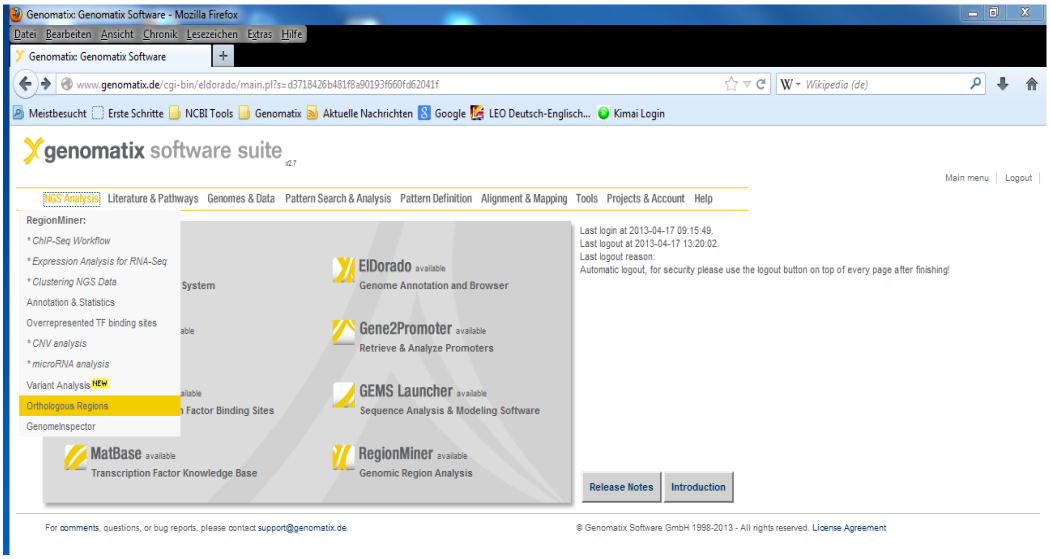
Step-by-step example for running PMCA manually using the graphical user interface

➔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.

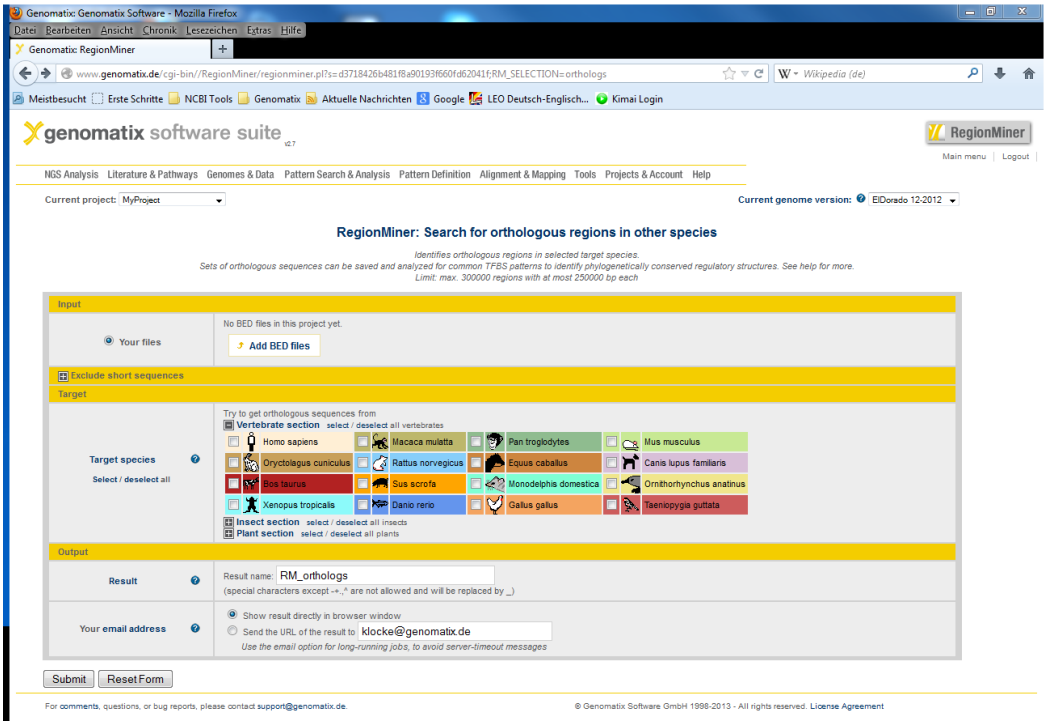
Below is an example for such a BED-file:

chr3 12386277 12386397 rs4684847

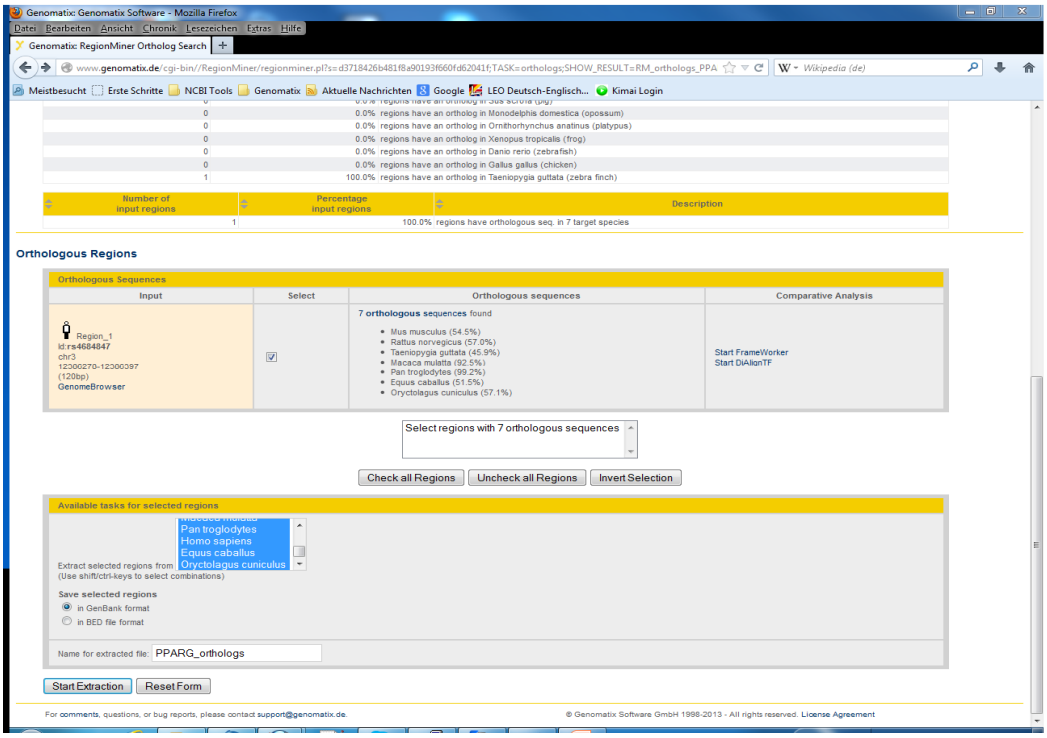
➔Upload the bed file to the Genomatix genome analyzer (GGA) software.



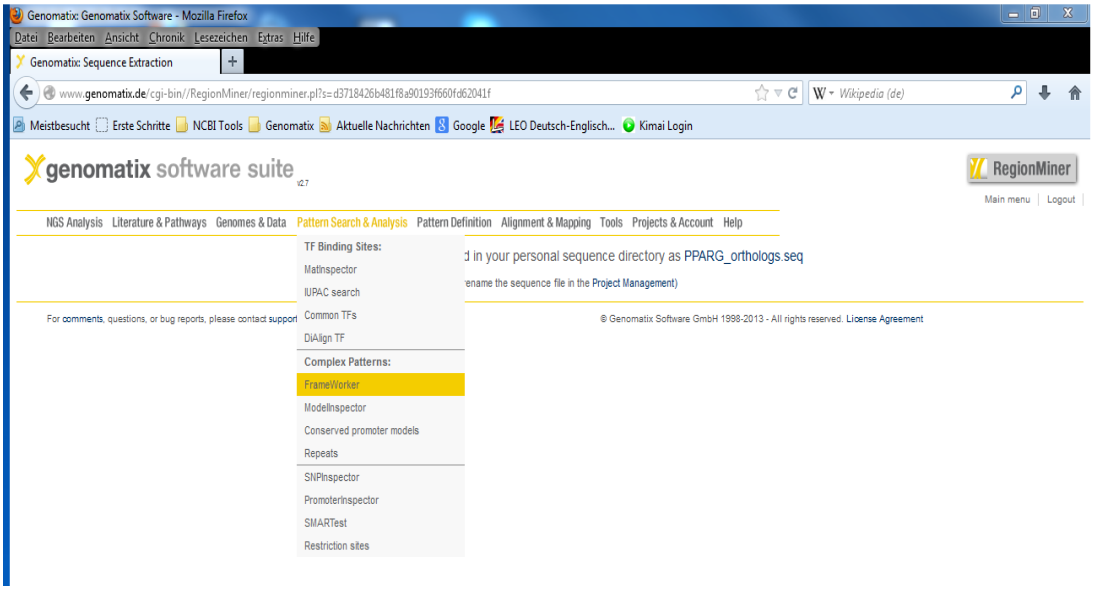
➔Search for orthologous regions by clicking on ‘Orthologous regions’



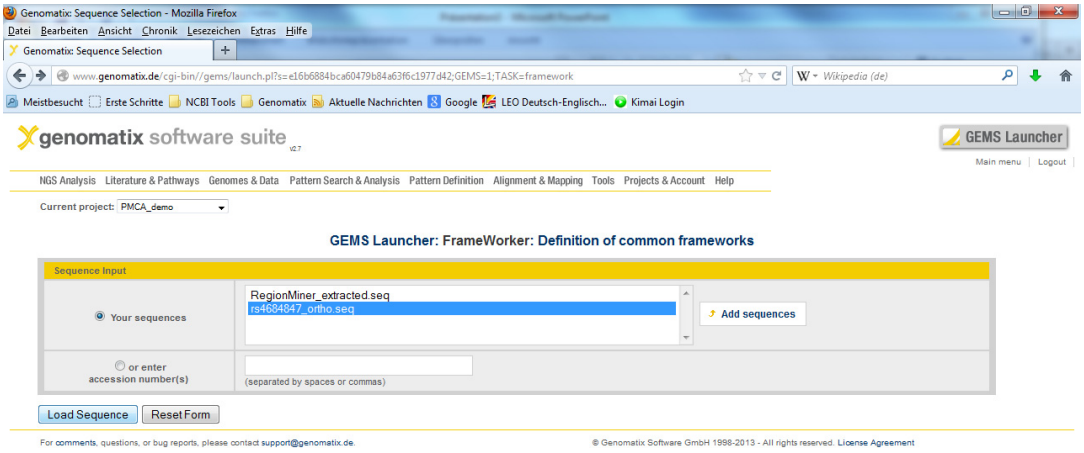
➔Extract the sequences



➔Start FrameWorker

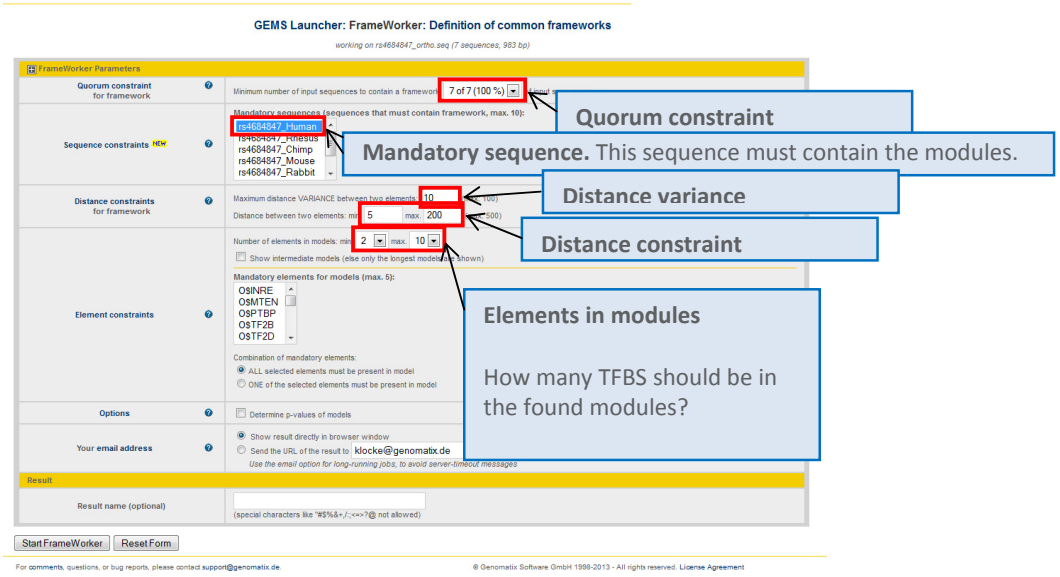


➔ Load the ortholog set that has been extracted in the previous step



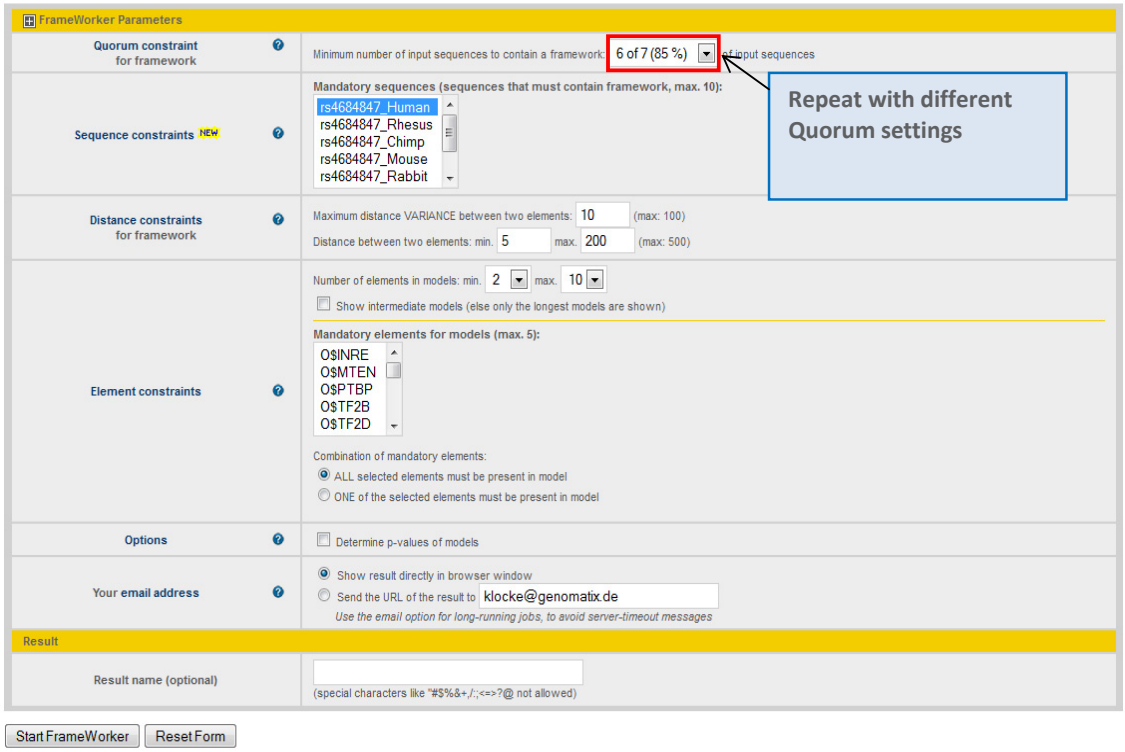
➔Select parameter as described below and click on ‘Start FrameWorker’

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



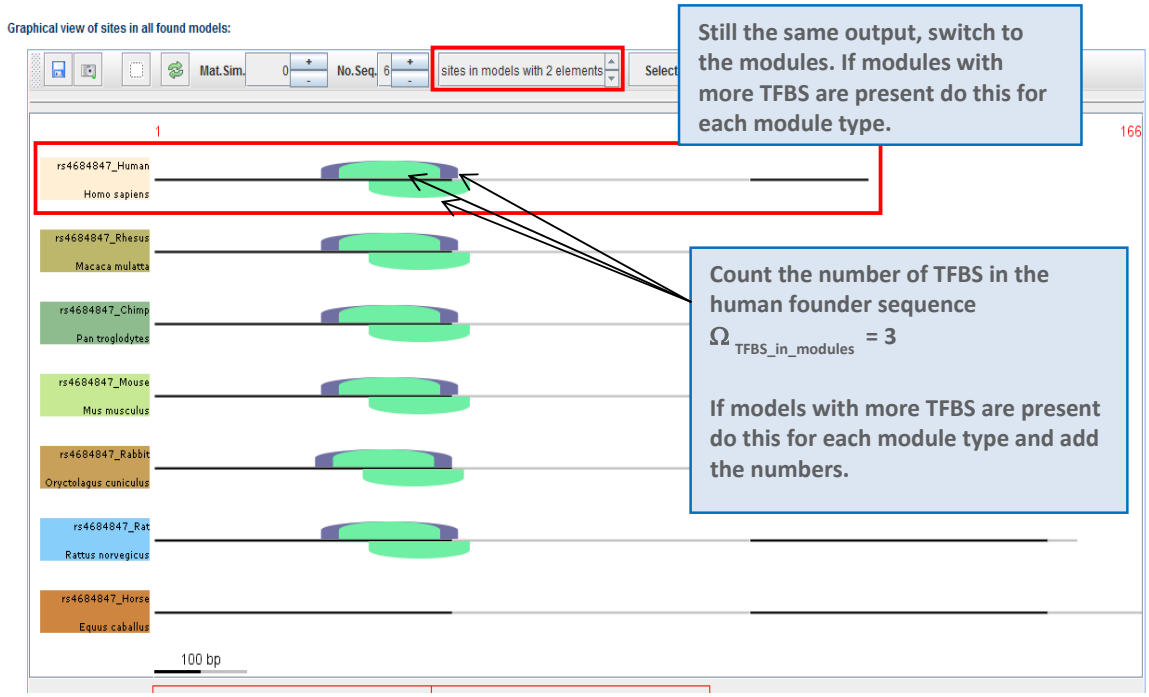
➔Count TFBS in the graphical output according to counting scheme described in the algorithm

➔ Repeat the step with the next quorum setting



➔ Count again


Overview: Models common to at least 6 sequences (85%) and in mandatory sequences
Models consisting of
single element
2 elements
# of different models
12 common elements found
2 models found
12 common elements:
Element
Strand
Matrix sim.
Common
O$PTBP
V$HOXC
V$OCT1
V$PARF
V$STEM
V$CREB
V$HOMF
V$PARF
V$HBOX
V$LHXF
V$LHXF
O$VTBP
Optimized
raphical view of sites in all found models:
Mat.Sim.
No.Seq.
all common sites
Select all
Deselect all
166
rs4684847_Human
Homo sapiens
rs4684847_Rhesus
Macaca mulatta
rs4684847_Chimp
Pan troglodytes
rs4684847_Mouse
Mus musculus
Two modules with 2 TFBS were found (Ω modules = 2).
If modules with more TFBS are found the numbers are added.
ΩTFBS: count the conserved TFBS sites in the human founder sequence = 12


Graphical view of sites in all found models:
Mat.Sim.
No.Seq.
sites in models with 2 elements
Select
1
166
rs4684847_Human
Homo sapiens
rs4684847_Rhesus
Macaca mulatta
rs4684847_Chimp
Pan troglodytes
rs4684847_Mouse
Mus musculus
rs4684847_Rabbit
Oryctolagus cuniculus
rs4684847_Rat
Rattus norvegicus
rs4684847_Horse
Equus caballus
100 bp
Still the same output, switch to the modules. If modules with more TFBS are present do this for each module type.
Count the number of TFBS in the human founder sequence
Ω TFBS_in_modules = 3
If models with more TFBS are present do this for each module type and add the numbers.

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

$\mathbf{\Omega}_{\text{TFBS}} = 6 + 12 = 18$

$\mathbf{\Omega}_{\text{modules}} = 2$

$\mathbf{\Omega}_{\text{TFBS in modules}} = 3$

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

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

1. $\mathbf{\Omega}_{\text{TFBS}}$ over all Quorum constraints
2. $\mathbf{\Omega}_{\text{modules}}$ over all Quorum constraints
3. $\mathbf{\Omega}_{\text{TFBS in modules}}$ over all Quorum constraints
4. $\Omega_{\text{restr-TFBS}}$ over Quorum restricted to $\zeta >= 50\%$

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}, $\Omega_{restr\text{-}TFBS}$ $\Omega_{modules}$, $\Omega_{TFBS\_in\_modules}$ counts as shown in the manual examples.

Finally each of the counts Ω_{TFBS}, $\Omega_{restr\text{-}TFBS}$, $\Omega_{modules}$, $\Omega_{TFBS\_in\_modules}$ 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.

5.4.4 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 et al., 2000). 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 positions. 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 regions and non-complex regions. The sequences were scanned by MatInspector (Cartharius et al.; Quandt et al., 1995) (Genomatix, Munich, Germany) for the presence of TFBS matrix family matches with respect to SNP positions (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 TFBS 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 et al. 2005). 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-value was calculated for each matrix family and each window (Table S6). For graphical visualization, –log10(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.

5.4.5 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., 2011). 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 SNPs in complex or non-complex regions 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 2 and Table S8.

5.4.6 Correlation of SNP regions to DHSseq regions and ChIPseq regions

Genomic regions surrounding a candidate SNP were classified as complex and non-complex and were correlated to DNase hypersensitive regions (referred to as DHSseq peaks; summary of Encode data wgEncodeRegDnaseClustered.bed from UCSC regulation super-track) and regions of transcription factor binding (referred to as ChIPseq peaks; summary of ENCODE ChIPseq 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 smoothening 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 2, Figure S2 and Table S9.

5.4.7 Enrichment of complex regions in diseases loci

We chose 19 disease traits of major importance (Table S5) to test the PMCA method. For each trait a maximum of six SNPs and with lowest p-values were selected from the GWAS catalogue (Hindorff LA, accessed July 7th 2013), if their GWAS associations were reported in individuals of European descent, the SNP was not selected for another trait before, and they were part of the 1000 genomes data (1000G Pilot 1 CEU data, 1000 Genomes Project Consortium, 2010). Matched random variants were drawn from the 1000G data (1000G Pilot 1 CEU data, 1000 Genomes Project Consortium, 2010). 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 ElDorado 2012 annotation (Genomatix, Munich, Germany), and for the distance to the nearest TSS within ± 10% of the disease-associated SNP. The process of random drawing was done using a Pearl Script.

5.4.8 GWAS enrichment analysis

GWAS results of insulin resistance (HOMA-IR) and impaired insulin secretion (HOMA-B) were extracted from the MAGIC consortium data repository (Dupuis et al., 2010) (ftp://ftp.sanger.ac.uk/pub/magic/). We identified 713 SNPs in 47 previously reported T2D susceptibility loci (SNPs in LD with the most significant SNP in each locus with $r^2 \geq 0.7$, Supplemental Table 7). TFBS-targeting SNPs were defined as the localization of SNPs in close proximity (SNP ± 20 bp) to at least one of the TFBS matrix clusters with p-value < 10^{-6} (positional bias analysis, Supplemental Table 12). We performed an enrichment analysis by the hypergeometric test to assess an over-representation of six TF families around 713 T2D SNP targets of interest. To estimate empirical p-value of the enrichment, we created 1,000

datasets each with ~700 SNPs in ~45 to 50 genomic loci randomly selecting from the same GWAS results. Each simulated dataset has approximately similar total length of DNA sequences compared to the total length of 47 T2D loci. The same enrichment analyses were performed in each simulated dataset to create the null distribution, and the empirical p-value of T2D SNPs was estimated from this null distribution.

5.4.9 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. For visualization we used all distances where a TSS was annotated within 30,000 bp downstream of a SNP. To directly compare theses distances for SNPs located in complex and non-complex regions we used density histograms with a bin size of 500 bp.

5.4.10 Culture of cell lines, Luciferase expression assays, EMSA, DNA-protein affinity chromatography

Culture of cell lines and Luciferase expression assays

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 µM 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 et al., 2008) in DMEM/Ham's F12 (1:1) medium (supplemented with 10% FCS, 17 µM biotin,

33 µM 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 µM biotin, 22 µM pantothenic acid, 1% penicillin/streptomycin, 10 µg/ml human transferrin, 66 nM insulin, 100 nM cortisol, 1 nM triiodothyronine, 20 nM dexamethasone, 500 µM 3-isobutyl-1-methyl-xanthine (Serva, Germany) and 2 µM 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).

Luciferase expression constructs

To characterize the SNP-surrounding 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-surrounding 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).

Luciferase expression assays

Huh7 cells (96-well plate, 1.1 x 10^4 / well) were transfected one day after plating with approximately 90% confluence, INS-1 cells (12-well plate, 8 x 10^4 / well) were transfected three days after plating with approximately 70% confluence, 3T3-L1 cells (12-well plate, 8 x 10^4 / well) were transfected at day eight after the induction of differentiation with approximately 80% confluence and C2C12 cells (12-well plate, 2 x 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 et al., 2009) 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 $\geq$ 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.

Electrophoretic mobility shift assay (EMSA)

EMSA was performed with Cy5-labelled oligonucleotide probes. Respective SNP-surrounding 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 hours 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 $MgCl_2$, 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.

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 $MgCl_2$, 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 et al., 2010). 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-surrounding region ($p = 0.034$ from Oneway ANOVA comparing the allelic difference of three independent experiments).

5.4.11 Genome editing in 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 (Ding et al., 2013; Wang et al., 2013a). 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'*CACCG*AAACTCACAACAATGCTGGG-3' and 5'*AAAC*CCCAGCATTGTTGTGAGTTT*C*-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 et al., 2013 (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 +1,200 bp 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'-CTTCCTTTTCTGCCCAGCTTCAAA-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'-CATCTCTAATTCTTACAACTCCGAAAAGATAAGAAAACAGAG-3' and 5'-CTCTGTTTTCTTATCTTTTCGGAGTTGTAAGAATTAGAGATG-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 would 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 Supplemental Experimental Procedure chapter *10. Culture of cell lines*. *PPARG1* and *PPARG2* mRNA expression levels were determined as described in the chapter *22. Quantitative RT-PCR and allele-specific primer extension analysis*. We assessed the genotype of the rs4684847 variant after HDR-mediated genome editing by sequencing 200 bp surrounding the SNP and confirmed homozygous C-allele and T-allele in the cells transfected with the respective genomic DNA targeting-vectors.

5.4.12 Analysis of human adipose tissue samples

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

PRRX1 mRNA was measured by qPCR (see chapter 21) in subcutaneous adipose tissue samples obtained from severely obese subjects matched for BMI (mean ± SD 43.2 ± 3.1 kg/m^2, n=67), body fat, age and sex, as described previously (Klöting et al., 2010). 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), *PRRX1* mRNA was also measured in another cohort comprising 30 obese (BMI>30 kg/m^2) otherwise healthy and 26 non-obese (BMI<30 kg/m2) healthy women (Arner et al., 2012). 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 fP-Glucose (mmol/L) x (fS-Insulin (microU/ml)/ 22.5) (Bonora et al., 2000). After the blood sampling an abdominal subcutaneous adipose tissue biopsy was obtained by needle aspiration. Adipose microarray analysis was performed exactly as described (Arner et al., 2012) using the Affymetrix GeneChip miRNA Array protocol with 1μg 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.nim.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.

5.4.13 Analysis of RNAseq data from primary human islets

Written informed consent was obtained from all patients who donated biological samples. The study was approved by the local ethics committee of Lund University, Sweden. RNAseq 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 101 bp length output reads were aligned to the human reference genome (NCBI 37.1/hg19) with TopHat (Trapnell et al., 2009). Gene expression was measured as the

normalized sum of expression of all its exons. The dexseq_count python script (Anders et al., 2012) was used by counting uniquely mapped reads in each exon. Gene expression normalization was done with the TMM method (Robinson and Oshlack, 2010). 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 et al., 2009), and significance was defined as FDR < 1% (Table S14). Further, Pearson's correlation and linear regression analyses were run using the R statistical computing environment for mRNA of the nine homeobox TFs, i.e. *BARX1, BARX2, EMX2, MSX2, NKX6-3, PDX1, PITX1, PRRX2* and *RAX*, separately for the normoglycaemic group encompassing 51 donors (Table S14) and a group at high risk of T2D (HbA1C > 6) encompassing 26 donors (Table S15) 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 nine TFs, pathway analysis was performed by WEBGESTALT (Wang et al., 2013b) with KEGG (Kanehisa et al., 2011) 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 six

significantly co-expressed genes which could not be unified to any pathway. In summary, six out of the nine TFs had pathway analysis information for the donors with HbA1c < 6.

5.4.14 eQTL analysis

Written 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 *PPARG* mRNA expression levels of carriers and non-carriers of the protective allele of the rs7638903 variant (perfect LD (r^2 = 1.0) to the tagSNP 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 exer*cis*e intervention (Elgzyri et al., 2012). Only baseline (before exer*cis*e) 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.

5.4.15 Isolation, culture, differentiation and genotyping of primary human adipose stromal cells (hASC)

Written informed consent was obtained from all patients who donated biological samples. The studies were approved by the local ethics committee of the Faculty of Medicine of the Technical University of Munich, Germany or the Regional Committee for Medical Research Ethics (REK) of Haukeland University Hospital, Bergen, Norway. Primary human adipocyte

progenitor cells for allele-specific primer extension analysis were obtained by lipoaspiration or surgical ex*cis*ion of subcutaneous adipose tissue, and were isolated and cultured as previously described (Hauner et al., 2001) 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, 17µM biotin, 33µM 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 50µL insulin (10mg/ml), 100µL cortisol (0.1mM), 1ml transferrin (1mg/ml), 50µL T3 (1nM/L), 50µL rosiglitazone (2mM), 100µL dexamethasone (25µM) and 1.25ml IBMX (20mM). The cells were harvested in TRIzol reagent (Invitrogen, Germany) (qPCR) or buffer RLT (Qiagen, Germany) (microarrays).

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 et al., 2008). 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.

5.4.16 Gene knock-down by siRNA

SGBS cells grown to confluence in 6-well plates (day 0) were treated to induce adipocyte differentiation (see chapter 10) and simultaneously transfected using the same protocol and siRNA as for primary hASCs (see below). 72 hours after induction of differentiation, the cells were harvested in TRIzol reagent (Invitrogen, Germany) and frozen at -80ºC. Primary hASCs were grown as described above (chapter 19), and on the same day of inducing adipogenic differentiation, cells were transfected with 25nM non-targeting siRNA (siNT) control or 25nM siRNA targeting PRRX1 (ON-TARGETplus human siRNA SMARTpool, Dharmacon, USA) for 72 hours, using HiPerFect (Qiagen, Germany) according to the manufacturer's protocol. Knock-down efficiency was 70-80%. 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 an insulin-concentrations were measured using a commercially available insulin-ELISA (Mercodia, Sweden). The cells were harvested in buffer RLT (Qiagen, Germany) and frozen at -80ºC for extraction of RNA and determination of knockdown efficiency.

5.4.17 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 *PPARG1* and *PPARG2* 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 ΔΔCT 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 *PPARG2* 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 *PPARG2* 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

Gene	Forward primer	Reverse primer
Human		
PCK1/PEPCKC	GCTCTGAGGAGGAGAATGG	TGCTCTTGGGTGACGATAAC
PDK4	TGCCAATTTCTCGTCTGTATG	AAAAACAGATGGAAAACTGAGG
LIPE	AGAAGATGTCGGAGCCCATA	GGTCAGGTTCTTGAGGGAATC
BBOX1	TTTCCAAGCAGGCCAGAG	CTGAACCCCAGGTGGATG
ADIPOQ	CATGACCAGGAAACCACGACT	TGAATGCTGAGCGGTAT
OPG	TTATGAGCATCTGGGACGGTGCTGT	AAGGAAGGTACAGTTGGTCCAGGGT
GLUT4	CTGTGCCATCCTGATGACTG	CCAGGGCCAATCTCAAAA
TIMP3	CTGACAGGTCGCGTCTATGA	AGTCACAAAGCAAGGCAGGT
THRSP	CGAGAAAGCCCAGGAGGTGA	AGCATCCCGGAGAACTGAGC

PPARG1	CGTGGCCGCAGATTTGA	AGTGGGAGTGGTCTTCCATTAC	
PPARG2	GAAAGCGATTCCTTCACTGAT	TCAAAGGAGTGGGAGTGGTC	
PRRX1	GTGGAGCAGCCCATCGTA	TGGGAGGGACGAGGATCT	
HPRT	TGAAAAGGACCCCACGAAG	AAGCAGATGGCCACAGAACTAG	
Rat			Knockdown efficiency*
Pdx1	TCCCGAATGGAACCGAGA	GTCAAGTTGAGCATCACTGCC	39 %
Barx2	AGTACCTCTCTACCCCAGACAG	CGTCTTCACCTGTAACTGGCT	12%
Pitx1	ACTCAGCCAGCGAGTCATCC	TTCTTCTTGGCTGGGTCTTCC	41%
Rax2	AGCGGGACCTTCAGTTTGG	CTTGGTCTTCGTGCCGTACTC	65%
Msx2	AAGGCAAAAAGACTGCAGGA	GGATGGGAAGCACAGGTCTA	24%
Emx2	GTCCCAGCTTTTAAGGCTAGA	CTTTTGCCTTTTGAATTTCGTTC	42%
Nkx6-3	ATGCAGCAACACCCCAGA	CCAGTGAATAAGCCAGCCTC	54%
Prrx2	ACTTCTCGGTGAGCCACCT	GCTGCTTCTTCTTCCGTTTG	38%
Barx1	CCTAGCCGTGGTCGCAT	GCCAGTGGGAACTTGAACA	52%
Gapdh	TGGGAAGCTGGTCATCAAC	GCATCACCCCATTTGATGTT	-

***Knockdown efficiency in rat INS-1 cells:** Efficiency of siRNA knockdown in INS-1 cells determined by qPCR is shown as the ratio of (mRNA level in siNT transfected cells)/(mRNA level in siTF transfected cells) with mRNA level=mRNA levels of the indicated Genes (normalized to *Gapdh* expression levels); siNT=non-targeting control siRNA.

Allele-specific primer extension analysis *PPARG2* mRNA		
	Forward primer	**Reverse primer**
genomic DNA	TCCATGCTGTTATGGGTGAA	GGAGCCATGCACAGAGATAA
cDNA	TCCATGCTGTTATGGGTGAA	GATGCAGGCTCCCATTTGAT
Snapshot	CTCTGGGAGATTCTCCTATTGAC	TATCAGTGAAGGAATCGCTTTCTG

5.4.18 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. Tissue was digested for 2 hours at 37ºC 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 200*g* 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).

Cells were seeded in 12-well plates in DMEM GlutaMax (Gibco) supplemented with 10% FCS and 1% penicillin/streptomycin, and induced to differentiate the day after plating ("day 0") by adding cortisol (100nM/L), insulin (66nM/L), transferrin (10μg/ml), biotin (33μM), pantothenate (17μM/L), T3 (1nM/L) and rosiglitazone (10μM). On the same or following day (day 0 or 1), new differentiation medium was added and cells were transfected 25nM siPRRX1 and 10nM non-targeting siRNA or 25nM siPRRX1 and 10nM siPPARG (ON-TARGETplus human siRNA SMARTpool, Dharmacon) using HiPerFect (Qiagen). After 72 hours, the cells were harvested in buffer RLT (Qiagen, Germany) and frozen at -80ºC.

RNA was extracted from siRNA-transfected lysates using the RNeasy Lipid Tissue Mini Kit (Qiagen, Germany), and total RNA quality was controlled by the Agilent 2100 Bioanalyzer (RIN > 9). 240ng of total RNA from each sample was biotin-labelled using the Illumina TotalPrep RNA Amplification Kit. 750ng 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: E-MTAB-1906).

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 and Jonassen, 2001). One of the non-targeting control samples was excluded because there were no expression signals from this sample, leaving a total of 9 sample pairs transfected with PRRX1 siRNA, four of which were co-transfected with PPARG siRNA. 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 PPARγ2, 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 PPARγ2 and not PPARγ1, 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 et al., 2005) 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 PPARγ2. 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 PPARγ2, 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).

5.4.19 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. The *Prrx1* isoform lacking the OAR-domain was previously described to be responsible for an inhibitory function of Prrx1 protein (Norris et al., 2001) and designated as mouse *Prrx1b*. We synthesized (Eurofins, Germany) the human *PRRX1* open reading frame lacking the OAR-domain (NCBI NM_006902.3, designated as *PRRX1a*) with an additional flag-tag sequence (5'-GACTACAAGGACGACGACGATAAG-3') inserted immediately after the *PRRX1* start codon. This DNA followed by an internal ribosomal entry site and a reading frame for the fluorescent protein VENUS targeted to the nuclear membrane (Okita et al., 2004) was cloned into the lentiviral expression backbone pRRL.PPT.SFFV.EGFP.pre (Schambach et al., 2006) to replace its EGFP reading frame. VSV-G pseudotyped lenti virus was produced by transient cotransfection of this plasmid with 3 plasmids containing reading frames for viral genes Rev, gag, pol and the VSV-G protein into 293HEK cells (Schambach, 2006). Virus supernatant was enriched by centrifugation and SGBS cells were infected at an MOI of 10. Cells were differentiated into mature adipocytes as described above (section 9). 14 days after induction of differentiation, medium was removed, cells were washed twice with PBS, followed by fixing in 3.7% formaldehyde for 60 min. The fixation solution was removed and replaced by Oil-Red-O stain solution (0.3% Oil-Red-O in 60/40 isopropanol/H2O, filtered through a

0.2µm mesh) for 60 min, before carefully washing twice with PBS, adding 1ml PBS, and photography under a Nikon TE2000 microscope.

5.4.20 Glyceroneogenesis and 2-deoxyglucose uptake measurements in primary hASC

For metabolic studies, genotyped primary hASCs from BMI-matched subjects were induced to differentiate and treated with NT control or PRRX1 siRNA as described above (chapter 21) and treated or not with 10 µM rosiglitazone. After 72 hours, cells were fasted for 3 hours in serum-free, 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 µM [1-^{14}C]-pyruvate (0.5 µCi) 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 and Dyer, 1959). The subsequent [1-^{14}C]-pyruvate incorporation was estimated by counting the radioactivity associated with the lipid fraction. The incubation medium (2 hours) 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 established in the lab during my Diploma thesis and were performed as previously described (Claussnitzer et al., 2011). Briefly, hASCs were induced to differentiate for three days, transfected with or without siRNA for 72 hours, and 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 1µM 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 µM 2-

DG and [^{3}H]-2-DG [2 µCi/ml]. Uptake was terminated by addition of ice-cold KRH containing 150 µM phloretin and 15 µM 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).

5.4.21 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 *PPARG2* mRNA expression during adipogenesis, p-values were calculated using Kruskal-Wallis Oneway ANOVA followed by Dunn's Multiple Comparison post-test. For qPCR analysis of siRNA experiments, p-values were calculated using the Wilcoxon rank-sum test (INS-1 cells, n = 9) or paired t-test (hASC, n = 16/32). Unpaired t-test was used for qPCR 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 *PRRX1* mRNA with *PPARG2* mRNA, pyruvate incorporation, free fatty acid release and ratio insulin-stimulated/basal 2-deoxyglucose uptake were calculated by Pearson's correlation. For correlation analysis of adipose tissue PRRX1 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 ratio levels (homozygous n=20, heterozygous n=18) we performed linear regression with log-transformed residuals (adjusted for age, sex and BMI). In addition, based on these residuals an interaction model was used to calculate the interaction p-value for *PRRX1* mRNA, rs4684847 genotype and HOMA-IR (adjusted for age, sex and BMI, n=38): adjusted phenotype (1) ~

adjusted mRNA (2) + SNP + adjusted mRNA * SNP. Statistical analyses were done using the Graph Pad Prism software version 5.02 or the Statistical Software R, version 2.14.2.

ACKNOWLEDGEMENTS

This work was funded by the Else Kröner-Fresenius Foundation, Bad Homburg v. d. H, Germany; the grant Virtual Institute 'Molecular basis of glucose regulation and type 2 diabetes' received from the Helmholtz Zentrum München, München-Neuherberg, Germany; the grant Clinical Cooperation Group 'Nutrigenomics and type 2 diabetes' received from the Helmholtz Zentrum München, München-Neuherberg, Germany, and the Technische Universität München; the Helmholtz Graduate School for Environmental Health, HELENA; a grant from the German Federal Ministry of Education and Research to the German Centre for Diabetes Research (DZD e.V.); the Competence Network Obesity (German Obesity Biomaterial Bank; FKZ 01GI1128), and the University Duisburg-Essen (01KU1216E); the KG Jebsen Center for Diabetes Research, University of Bergen, Norway and the Western Norway Regional Health Authority, Norway; by grants from the Swedish Research Council, including strategic research area grant EXODIAB (2009-1039), Linnaeus grant (349-2006-237), Collaborative Grant (2011-3315) and Project Grant (521-2010-3490) and by an ERC Advanced Researcher Grant GENETARGET-T2D (GA 269045); and by 'Biomedical Research Program' funds at Weill Cornell Medical College in Qatar, a program funded by the Qatar Foundation. We thank Karl-Fredrik Eriksson and Targ Elgzyri for providing human fat biopsy material and Charlotte Ling for supporting generation of microarray data (Lund University). We are grateful to Andrea Califano (Columbia University, New York, US) for critical review and constructive scientific comments on the manuscript. We further thank Bernd Baumann, Vidar M. Steen and Jørn V. Sagen for expert advice, and Elisabeth Hofmair, Manuela Hubersberger, Margit Solsvik, Linn Jeanette Waagbø, Tone Nygaard Fløtlo, Jan-Inge Bjune, Zina Fandalyuk, Vivian Veum and Christine Haugen for excellent technical assistance. We thank Ralf Kühn (Helmholtz Zentrum München, München-Neuherberg) for providing CRISPR/Cas vectors. We thank Christine Stansberg, Rita Holdhus and Kjell Petersen at the Norwegian Microarray Consortium (NMC) (Bergen node, Norway), Douglas P. Kiel and David Karasik (Hebrew Senior Life Institute for Aging Research, Harvard Medical School, Boston, US), and Michael Molla (Joslin Diabetes Center, Harvard Medical School, Boston) for expert assistance. We thank surgeon Hans Jørgen Nielsen and colleagues (Voss Hospital, Norway), surgeons Barbara Auras Jaatun and Inge Glambæk and colleagues (Haraldsplass Deaconess Hospital, Bergen), surgeon Christian Busch (Klinikk Bergen, Nesttun), and all patients for providing human adipose tissue. We thank Reiner Schroeer and Dorothy Dankel for critical reading of the manuscript. The authors declare competing financial interests: details accompany the full-text HTML version of the paper. Correspondence should be addressed to M.C. (MelinaClaussnitzer@hsl.harvard.edu) or H.L. (helmut.laumen@wzw.tum.de).

5.5 SUPPLEMENTAL NOTES

– Authors from DIAGRAM+:

Benjamin F Voight[1,2,3], Laura J Scott[4], Valgerdur Steinthorsdottir[5], Andrew P Morris[6], Christian Dina[7,8], Ryan P Welch[9], Eleftheria Zeggini[6,10], Cornelia Huth[11,12], Yurii S Aulchenko[13], Gudmar Thorleifsson[5], Laura J McCulloch[14], Teresa Ferreira[6], Harald Grallert[11,12], Najaf Amin[13], Guanming Wu[15], Cristen J Willer[4], Soumya Raychaudhuri[1,2,16], Steve A McCarroll[1,17], Claudia Langenberg[18], Oliver M Hofmann[19], Josée Dupuis[20,21], Lu Qi[22–24], Ayellet V Segrè[1,2,17], Mandy van Hoek[25], Pau Navarro[26], Kristin Ardlie[1], Beverley Balkau[27,28], Rafn Benediktsson[29,30], Amanda J Bennett[14], Roza Blagieva[31], Eric Boerwinkle[32], Lori L Bonnycastle[33], Kristina Bengtsson Boström[34], Bert Bravenboer[35], Suzannah Bumpstead[10], Noël P Burtt[1], Guillaume Charpentier[36], Peter S Chines[33], Marilyn Cornelis[24], David J Couper[37], Gabe Crawford[1], Alex SF Doney[38,39], Katherine S Elliott[6], Amanda L Elliott[1,17,40], Michael R Erdos[33], Caroline S Fox[21,41], Christopher S Franklin[42], Martha Ganser[4], Christian Gieger[11], Niels Grarup[43], Todd Green[1,2], Simon Griffin[18], Christopher J Groves[14], Candace Guiducci[1], Samy Hadjadj[44], Neelam Hassanali[14], Christian Herder[45], Bo Isomaa[46,47], Anne U Jackson[4], Paul RV Johnson[48], Torben Jørgensen[49,50], Wen HL Kao[51,52], Norman Klopp[11], Augustine Kong[5], Peter Kraft[22,23], Johanna Kuusisto[53], Torsten Lauritzen[54], Man Li[51], Aloysius Lieverse[55], Cecilia M Lindgren[6], Valeriya Lyssenko[56], Michel Marre[57,58], Thomas Meitinger[59,60], Kristian Midthjell[61], Mario A Morken[33], Narisu Narisu[33], Peter Nilsson[56], Katharine R Owen[14], Felicity Payne[10], John RB Perry[62,63], Ann-Kristin Petersen[11], Carl Platou[61], Christine Proença[7], Inga Prokopenko[6,14], Wolfgang Rathmann[64], N William Rayner[6,14], Neil R Robertson[6,14], Ghislain Rocheleau[65–67], Michael Roden[45,68], Michael J Sampson[69], Richa Saxena[1,2,40], Beverley M Shields[62,63], Peter Shrader[3,70], Gunnar Sigurdsson[29,30], Thomas Sparsø[43], Klaus Strassburger[64], Heather M Stringham[4], Qi Sun[22,23], Amy J Swift[33], Barbara Thorand[11], Jean Tichet[71], Tiinamaija Tuomi[46,72], Rob M van Dam[24], Timon W van Haeften[73], Thijs van Herpt[25,55], Jana V van Vliet-Ostaptchouk[74], G Bragi Walters[5], Michael N Weedon[62,63], *Cis*ca Wijmenga[75], Jacqueline Witteman[13], Richard N Bergman[76], Stephane Cauchi[7], Fran*cis* S Collins[77], Anna L Gloyn[14], Ulf Gyllensten[78], Torben Hansen[43,79], Winston A Hide[19], Graham A Hitman[80], Albert Hofman[13], David J Hunter[22,23], Kristian Hveem[61,81], Markku Laakso[53], Karen L Mohlke[82], Andrew D Morris[38,39], Colin NA Palmer[38,39], Peter P Pramstaller[83], Igor Rudan[42,84,85], Eric Sijbrands[25], Lincoln D Stein[15], Jaakko Tuomilehto[86], Andre Uitterlinden[25], Mark Walker[87], Nicholas J Wareham[18], Richard M Watanabe[76,88], Goncalo R Abecasis[4], Bernhard O Boehm[31], Harry Campbell[42], Mark J Daly[1,2], Andrew T Hattersley[62,63], Frank B Hu[22–24], James B Meigs[3,70], James S Pankow[89], Oluf Pedersen[43,90,91], H.-Erich Wichmann[11,12,92], Inês Barroso[10], Jose C Florez[1,2,3,93], Timothy M Frayling[62,63], Leif Groop[56,72], Rob Sladek[65–67], Unnur Thorsteinsdottir[5,94], James F Wilson[42], Thomas Illig[11], Philippe Froguel[7,95], Cornelia M van Duijn[13], Kari Stefansson[5,94], David Altshuler[1,2,3,17,40,93], Michael Boehnke[4], 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. Landspitali 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. Malmska Municipal 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 Crohnic 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 Preventitive 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

5.6 SUPPLEMENTAL REFERENCES

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

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Research *22*, 2008-2017.

Arner, E., Mejhert, N., Kulyte, A., Balwierz, P.J., Pachkov, M., Cormont, M., Lorente-Cebrian, S., Ehrlund, A., Laurencikiene, J., and Heden, P., et al. (2012). Adipose Tissue MicroRNAs as Regulators of CCL2 Production in Human Obesity. Diabetes *61*, 1986-1993.

Bligh, E.G., and Dyer, W.J. (1959). A RAPID METHOD OF TOTAL LIPID EXTRACTION AND PURIFICATION. Can. J. Biochem. Physiol. *37*, 911-917.

Bonora, E., Targher, G., Alberiche, M., Bonadonna, R.C., Saggiani, F., Zenere, M.B., Monauni, T., and Muggeo, M. (2000). Homeostasis model assessment closely mirrors the glucose clamp technique in the assessment of insulin sensitivity: studies in subjects with various degrees of glucose tolerance and insulin sensitivity. Diabetes Care *23*, 57-63.

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.

Ding, Q., Regan, S.N., Xia, Y., Oostrom, L.A., Cowan, C.A., and Musunuru, K. (2013). Enhanced Efficiency of Human Pluripotent Stem Cell Genome Editing through Replacing TALENs with CRISPRs. Cell Stem Cell *12*, 393-394.

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

Dysvik, B., and Jonassen, I. (2001). J-Express: exploring gene expression data using Java. Bioinformatics *17*, 369-370.

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.

Hauner, H., Skurk, T., and Wabitsch, M. (2001). Cultures of Human Adipose Precursor Cells. Adipose Tissue Protocols. In , G. Ailhaud, ed. (Springer New York), pp. 239–247.

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

Holzapfel, 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.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., and Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol.

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.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2011). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research *40*, D109.

Klöting, N., Fasshauer, M., Dietrich, A., Kovacs, P., Schön, M.R., Kern, M., Stumvoll, M., and Blüher, M. (2010). Insulin-sensitive obesity. American Journal of Physiology - Endocrinology And Metabolism *299*, E506-E515.

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.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., and Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res *34*, D108-10.

Norris, R.A., Kern, M.J. (2001). The Identification of Prx1 Transcription Regulatory Domains Provides a Mechanism for Unequal Compensation by the Prx1 and Prx2 Loci. J. Biol. Chem. *276*, 26829-26837.

Okita, C., Sato, M., Schroeder, T. (2004). Generation of optimized yellow and red fluorescent proteins with distinct subcellular localization. Biotechniques 36, 418-22.

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

Claussnitzer, M., Skurk, T., Hauner, H., Daniel, H., and Rist, M.J. (2011). Effect of flavonoids on basal and insulin-stimulated 2-deoxyglucose uptake in adipocytes. Mol. Nutr. Food Res. *55*, S26.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol *11*, R25.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol *16*, 939-945.

Schambach, A., M. Galla, et al. (2006). Lentiviral vectors pseudotyped with murine ecotropic envelope: increased biosafety and convenience in preclinical research. Exp. Hematol., *34*, 588-592.

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.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., and Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

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

Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013b). WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. Nucleic Acids Research *41*, W77.