Wissenschaftszentrum Weihenstephan
für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Ernährungsmedizin

Understanding the molecular mechanism underlying the effect of *cis*-regulatory variants on gene expression at T2D associated loci

Heekyoung Lee

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. H. Daniel
Prüfer der Dissertation:	1. Univ.-Prof. Dr. J.J. Hauner
	2. Univ.-Prof. Dr. M. Hrabě de Angelis

Die Dissertation wurde am 23.12.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 15.04.2016 angenommen.

Table of contents

List of figures

List of tables

List of publications

1. Claussnitzer M, Dankel SN, Klocke B, Grallert H, Glunk V, Berulava T, **Lee H**, Oskolkov N, Fadista J, Ehlers K, et al. 2014. Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms. *Cell* 156: 343–358

2. Spieler D, Kaffe M, Knauf F, Bessa J, Tena JJ, Giesert F, Schormair B, Tilch E, **Lee H**, Horsch M, et al. 2014. Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. Genome Research 24: 592–603

3. Kretschmer A, Möller G, **Lee H**, Laumen H, Toerne C, Schramm K, Prokisch H, Eyerich S, Wahl S, Baurecht H, et al. 2014. A common atopy-associated variant in the Th2 cytokine locus control region impacts transcriptional regulation and alters SMAD3 and SP1 binding. Allergy 69: 632–642.

4. **Lee H**, von Toerne C, Claussnitzer M, Hoffmann C, Glunk V, Wahl S, Breier M, Molnos S, Grallert H, Dahlmann I, Arner P, Hauner H, Hauck SM, Laumen H. 2014. Unbiased allele-specific quantitative proteomics unravels molecular mechanisms modulated by *cis*-regulatory variation at the *PPARG* locus. Manuscript in preparation.

1. Abbreviations

293T	Human embryonic kidney cell line
3T3-L1	Mouse white pre-adipocyte cell line
AC	Affinity chromatography
$AgNO_3$	Silver nitrate
AKT	Protein kinase B
AMPk	AMP-activated protein kinase
APS	Ammonium-persulphate
ATCC	American Type Culture Collection
B&W	Binding and Wash buffer
BAT	Brown adipose tissue
BB	Binding buffer
BMI	Body mass index
bp	Base pair
BSA	Bovine serum albumin
°C	Degree centigrade
C2C12	Mouse myoblast cell line
CALR	Calreticulin
CAMK1D	Calcium/calmodulin-dependent protein kinase 1D
cAMP	Cyclic adenosine monophosphate
CBX3	Chromobox homolog 3
cDNA	Complementary DNA
CDXA	Caudal-type homeodomain protein A
C/ebpβ	CCAAT/enhancer-binding protein β
CENTD2	Centaurin, delta 2
CEU	Northern and Western European ancestry
Chaps	3-[(3-Cholamidopropyl)-dimethyl-ammonio]-propansulfonat
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation-sequencing
CNBr	Cyanogen bromide
CO_2	Carbon dioxide
Conc.	Concentration
COX	Cyclooxygenase
CRMs	*Cis*-regulatory modules
Ct	Control
CTCF	CCCTC-binding factor
CUX1	Cut-Like Homeobox 1
CVD	Cardiovascular disease
Cy5	Cyanine 5
Da	Dalton
DDX5	DEAD (Asp-Glu-Ala-Asp) box helicase 5
DDX17	DEAD (Asp-Glu-Ala-Asp) box helicase 17
DEK	DEK oncogene
Dest.H_2O	Destilled water
Dexa	Dexamethason
DHX9	DEAH (Asp-Glu-Ala-His) box helicase 9
DIDO1	Death inducer-obliterator 1
DMEM	Dulbecco's Modified Eagle Medium
DMEM:F12	Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12
DMSO	Dimethyl sulphoxide
DNA	Deoxy-ribonucleic acid
DNase-Seq	DNase I hypersensitive sites-sequencing
dNTPs	Deoxynucleotide triphosphates
ds	Double-stranded
DTT	Dithiothreitol
ECL	Enhanced chemiluminescence
EDTA	Ethylendiamine-tetra-acetic acid
E2F	Elongation factor 2
EGTA	Ethylene glycol tetraacetic acid
EMSA	Electrophoretic mobility shift assay
ENCODE	ENCyclopedia Of DNA Elements
eQTL	Expression quantitative trait loci
ESRRA	Estrogen-related receptor alpha
EWSR1	EWS RNA-binding protein 1
FAIRE-Seq	Formaldehyde-Assisted Isolation of Regulatory Elements-sequencing
FASP	Filter-aided sample preparation
FCS G	Fetal calf serum gold
FFAs	Free fatty acids
FoxC2	Forkhead box C2
fM	Femto mole
FTO	Fat mass and obesity
fwd	Forward
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GBB	Gel binding buffer
GCK	Glucokinase
GEO	Gene Expression Omnibus
GePS	Genomatix Pathway System
GO term	Gene ontology term

GWAS	Genome-wide association studies
h	Hour
HapMap	International Haplotype Map project
HCFC1	Host cell factor C1
HDL	High-density lipoprotein
HEPES	4-(2-hydroxyethyl)-1-piperazineethane-sulfonic acid
HMGB2	High mobility group box 2
HMGB3	High mobility group box 3
HIB 1B	Mouse brown pre-adipocyte cell line
H3K27ac	Histone H3-lysine 27 acetylation
H3K4me1	Histone H3-lysine 4 mono-methylation
H3K4me2	Histone H3-lysine 4 di-methylation
H3K4me3	Histone H3-lysine 4 tri-methylation
HMGB2	High mobility group box 2
HMGN1	High mobility group nucleosome binding domain 1
HNRNPA2B1	Heterogeneous nuclear ribonucleoprotein A2/B1
HOMA-IR	Homeostatic model assessment-insulin resistance
HPRT	Hypoxanthin phosphoribosyltransferase
HSB	High salt buffer
Huh7	Human hepatocyte cell line
IBMX	3-isobutyl-1-methylxanthine
IL2RA	Interleukin 2 (IL2) receptor alpha
INS-1	Rat pancreatic beta cell line
IRX3	Iroquois homeobox protein 3
KCL	Kalium chloride
kDa	Kilo dalton
kg/m^2	Kilogram/square meter
KHDRBS1	KH domain containing, RNA binding, signal transduction associated 1
KH_2PO_4	Potassium dihydrogen phosphate
KLF14	Krüppel-like factor 14
LADA	Latent Autoimmune Diabetes in Adults
LB	Lysogeny broth
LC	Liquid chromatography
LD	Linkage disequilibrium
LSB	Low salt buffer
MAF	Minor allele frequency
mg	Mili gram
$MgCl_2$	Magnesium chloride
MHO	Metabolically healthy obese
min	Minute
miRNA	Micro ribonucleic acid
ml	Mili liter
mM	Milli mole
mRNA	Messenger ribonucleic acid
MS	Mass Spectrometry
MTA2	Metastasis associated 1 family, member 2
MTDH	Metadherin
MTR1B	Melatonin receptor 1B
MyoD	Myogenic differentiation 1
MS	Microsoft
NaCl	Sodium chloride
NaOH	Sodium hydroxide
$NaHCO_3$	Sodium bicarbonate
Na_2HPO_4	Sodium phosphate dibasic
NF	Sodium fluoride
NFATC4	Nuclear factor of activated T-cells, cytoplasmic, calcineurin-dependent 4
ng	Nano gram
NGS	Next-generation sequencing
nM	Nano mole
nm	Nano meter
No.	Number
NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)
Np-40	Nonidet P-40
NT	Non-targeting
Oct-1	Octamer-binding transcription factor-1
OD	Optical density
o/n	Over night
OR	Odds ratio
PAGE	Polyacrylamide gel electrophoresis
PA2G4	Proliferation-associated 2G4, 38kDa
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PFN1	Profilin 1
PGC-1α	Peroxisome proliferator activated receptor gamma coactivator 1α
PHB	Prohibitin
pM	Pico mole
PMCA	Phylogenetic module complexity analysis
PMSF	Phenylmethylsulfonyl fluoride

Poly (dI-dC)	Poly (deoxyinosinic-deoxycytidylic) acid sodium salt
PPARA	Peroxisome proliferator-activated receptor, alpha
PPARB	Peroxisome proliferator-activated receptor, beta
PPARG	Peroxisome proliferator-activated receptor, gamma
PRDM16	PRD1-BF-1-RIZ1 homologous domain containing protein-16
PPI	Protein-protein interaction
PPREs	PPAR-response elements
PRRX1	High mobility group box 2
P/S	Penicillin and streptomycin
PUF60	Poly-U binding splicing factor, 60KDa
PVDF	Polyvinylidenfluorid
qPCR	Quantitative polymerase chain reaction
RBBP4	Retinoblastoma binding protein 4
RBM14	RNA binding motif protein 14
RBM39	RNA binding motif protein 39
RBP4	Retinol-binding protein 4
rev	Reverse
RING1	Ring finger protein 1
RNA	Ribonucleic acid
RP53	Retinol dehydrogenase 12
rpm	Round per minute
RPMI	Roswell Park Memorial Institute
RSLC	Rapid separation liquid chromatography
RT	Room temperature
RYBP	RING1 and YY1 binding protein
sc	Santa cruz biotechnology
SCX	Strong cation exchange
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulfate- polyacrylamide gel electrophoresis
SFPQ	Splicing factor proline/glutamine-rich
SGBS	Simpsom golabi behmel syndrome cell line
SH-SY5Y	Human neuroblastoma cell line
SILAC	Stable isotope labeling by amino acids in cell culture
siRNA	Small interfering RNA
sn	Supernatant
SNP	Single nucleotide polymorphism
Sp1	Sp/KLF family of transcription factor
SPSS	Statistical Package for the Social Sciences
SSBP1	Single-stranded DNA binding protein 1, mitochondrial
SUB1	SUB1 homolog (S. cerevisiae)
TARDBP	TAR DNA binding protein
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TBE	Tris/Borate/EDTA
TBS	Tris/Buffer/Saline
TBST	Tris/Buffer/Saline/Tween
TCF7L2	Transcription factor 7 like 2
TE	Tris-EDTA
TEMED	N, N, N´, N´-Tetramethylendiamine
TFBS	Transcription factor binding site
TG	Triglycerides
TK	Thymidine kinase
Trex	Transcriptional regulatory element X
TRIM28	Tripartite motif containing 28
TSSs	Transcription start sites
TZDs	Thiazolidinediones
UCP	Uncoupling protein
UHRF1	Ubiquitin-like with PHD and ring finger domains 1
UV	Ultraviolet
V	Volt
v/v	Volume/volume
WAT	White adipose tissue
WHO	World Health Organization
w/o	Without
w/v	Weight/Volume
WT	Wild type
YAF2	YY1 associated factor 2
YMHAH	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta
YWHAQ	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta
ZNF35	Zinc finger protein 35
µg	Micro gram
µM	Micro mole
µl	Micro liter

2. Summary

Genome-wide association studies (GWAS) identified numerous risk loci associated with diverse diseases including type 2 diabetes (T2D) and obesity. Interestingly, most variants are located in noncoding regions of the genome, thus may modulate gene transcription. Recent technological advances, such as ChIP-seq, DNase-seq and FAIRE-seq based profiling of epigenetic marks of regulatory regions, fine mapping and novel bioinformatics approaches, enabled identification of *cis*-regulatory, potentially disease-causing variants within complex loci on a large-scale. Analysis of epigenetic marks of regulatory regions revealed that active *cis*-regulatory regions are associated with allele-specific transcription factor (TF) binding, which triggers regulatory cascades. However, only a few studies identified causal *cis*-regulatory variants and additionally reported the *trans*-regulatory proteins such as TFs, which modulate gene expression by allele-specific DNA binding at *cis*-regulatory variants.

Allele-specific TF binding analysis from ChIP-seq provides mechanistic evidence for how regulatory variants function. However, the ChIP-seq assay is mainly restricted to known biological targets. Proteomics can be a complementary approach to ChIP-seq for defining allele-specific binding TFs. Diverse studies have focused on the identification of allele-specific DNA-binding proteins using quantitative proteomics. Most studies apply stable isotope labeling which faces limitations such as applicability to disease-relevant human tissues, time-consuming procedure, high-cost or inefficient labeling. Therefore, this study focused first on establishment of a highly sensitive label-free quantitative DNA protein interaction approach for identification of allele-specific binding proteins at disease-causing noncoding *cis*-regulatory variants and additionally unraveling the affected molecular mechanisms. For prediction of *cis*-regulatory variants, a combination of different approaches, such as bioinformatics Phylogenetic transcription factor Module Complexity Analysis (PMCA) in the respective SNP-surrounding genomic regions, public data on epigenetic marks of regulatory regions and published fine-mapping studies were used. Regulatory variants were selected at T2D or obesity risk-loci, i.e. *PPARG* rs4684847 and rs7647481, *FTO* rs1421085 and *TCF7L2* rs7903146. The associations of *PPARG*, *FTO* and *TCF7L2* loci with T2D, obesity and related diseases have been well established in various studies. However, the precise mechanisms by which those variants affect T2D or obesity remain elusive yet. Detection of allele-specific protein-DNA interactions was optimized under various conditions in Electrophoretic Mobility Shift Assay (EMSA) experiments, and the

proteins were isolated using magnetic bead-based affinity chromatography, followed by an unbiased label-free quantitative proteomics. After selection of the candidate proteins based on the Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) data and biological relevance, the allele-specific binding of candidate proteins was confirmed by competition/supershift EMSA.

Second, this study focused on the *in depth*-analysis of variants at the T2D associated *PPARG* locus. Using PMCA analysis and epigenetic marks data the *cis*-regulatory variant rs4684847 at the *PPARG* locus was previously predicted. Here, the highly efficient label free proteomics methodology identified the TF PRRX1 regulating PPARγ2 expression in adipocytes. It has been suggested previously that multiple causal variants within a given locus cooperatively may modulate gene expression and confer susceptibility to common traits. Indeed, an integrative framework combining PMCA analysis and epigenetic marks data with the highly sensitive label free proteomics enabled identification of *cis*-regulatory variant rs7647481 at the *PPARG* locus, and a nonrisk allele-specific binding of the *YY1 (Ying Yang 1)* TF and its coregulator *RYBP (RING1 and YY1 binding protein).* Pathophysiological relevance of these findings was supported by a correlation of expression levels in primary adipose tissue with improved insulin sensitivity in subjects carrying the rs7647481 nonrisk allele. Moreover, this study provides experimental evidence for a significantly increased binding of TFs and related proteins to predicted *cis*-regulatory vs. non *cis*-regulatory variants.

Finally, the findings at the *PPARG* locus prove the high efficiency of the unbiased proteomics approach and support the power of the PMCA analysis to predict causal variants. Moreover, these findings further support the recently proposed "multiple enhancer hypothesis". The successful identification of both, TFs and their cofactors enables to infer allele-specific protein-DNA interaction networks, which will serve as a base for a prediction of how genetic variants in regulatory mechanisms change gene expression profiles and human phenotype. Moreover, the presented unbiased proteomics approach will enable analysis in primary human tissue, can be applied to any kind of variability, including somatic mutations in cancer, and thus can help to clarify the role of both inherited and somatic variability.

2. Zusammenfassung

Genomweite Assoziationsstudien (GWAS) identifizierten zahlreiche Risiko-Loci für diverse komplexe Krankheiten wie Typ 2 Diabetes (T2D) und Adipositas. Interessanterweise finden sich die meisten Varianten in nicht-kodierenden Regionen des Genoms, welche die Gentranskription modulieren könnten. Neuste technologische Fortschritte wie ChIP-seq, DNase-seq und FAIRE-seq zur Identifizierung von epigenetischen Markierungen regulatorischer Regionen, genetische Feinkartierung oder neuartiger Bioinformatik-Ansätze ermöglichen die Identifizierung von *cis*-regulatorischen, potenziell krankheitsrelevanten Varianten in komplexen Loci in großem Maßstab. Die Analyse von epigenetischen Markierungen regulatorischer Regionen ergab weiterhin, dass aktive *cis*-regulatorische Regionen mit allel-spezifischen Transkriptionsfaktoren (TF) assoziiert sind, die Regulationskaskade auslösen können. Bislang identifizierten nur wenige Studien kausale *cis*-regulatorische Varianten und zusätzlich die *trans*-regulatorischen Proteine wie TFs, welche Genexpression durch allel-spezifische DNA-Bindung an *cis*-regulatorischen Varianten modulieren.

Die allele-spezifische TF Bindungsanalyse durch ChIP-seq liefert mechanistische Beweise für die Funktion regulatorischer Varianten, wobei sie gezielt bekannte biologische Targets analysieren. Eine nicht-gezielte Proteomanalyse wäre eine den ChIP-seq-Ansatz ergänzende Methode zur Definition von allel-spezifisch bindenden TFs. Viele Studien haben sich auf die Identifizierung von allel-spezifischen DNA-bindenden Proteinen mit Hilfe der quantitativen Proteomik fokussiert. Die meisten Studien verwenden eine Markierung mit stabilen Isotopen, die mit hohem Zeitaufwand und mit hohen Kosten sowie begrenzter Anwendbarkeit in krankheitsrelevanten menschlichen Geweben einhergehen. Deshalb konzentrierte sich diese Studie zunächst auf die Etablierung eines hochempfindlichen, markierungsfreien quantitativen DNA-Protein-Wechselwirkungs-Ansatzes zur Identifizierung allel-spezifisch bindender Proteine an krankheits-relevanten, nicht-kodierenden *cis*-regulatorischen Varianten und die Aufklärung der molekularen Mechanismen. Für die Vorhersage von *cis*-regulatorischen Varianten wurde eine Kombination von unterschiedlichen Ansätzen, wie die bioinformatische „Phylogenetic Transcriptionsfactor Module Complexity Analysis“ (PMCA) in den jeweiligen SNP-umgebenden genomischen Regionen, die Analyse öffentlicher Daten über epigenetische Marker regulatorischer Regionen, sowie veröffentlichte Feinkartierungsstudien verwendet. Es wurden regulatorische Varianten in den T2D- und

Adipositas-Risiko-Loci *PPARG* rs4684847 und rs7647481, *FTO* rs1421085 oder *TCF7L2* rs7903146 ausgewählt. Die Assoziationen der *PPARG*, *FTO* und *TCF7L2* Loci mit T2D, Adipositas und assoziierten Erkrankungen wurden in verschiedenen Studien festgestellt. Die genauen Mechanismen, über welche diese Varianten T2D oder Adipositas beeinflussen, sind jedoch noch unklar. Die experimentelle Analyse allel-spezifischer Protein-DNA-Wechselwirkungen wurde unter verschiedenen Bedingungen in „Electrophoretic Mobility Shift Assay“ (EMSA)-Experimenten optimiert und die Proteine wurden unter Verwendung von Magnetkügelchen-basierter Affinitätschromatographie isoliert, gefolgt von einer markierungsfreien, quantitativen Proteomik. Nach der Auswahl der Kandidatenproteine auf Grundlage von „Liquid Chromatography–tandem Mass Spectrometry“ (LC-MS/MS)-Daten und der biologischen Relevanz wurde die allel-spezifische Bindung der Kandidatenproteine durch Kompetitions- und Supershift- EMSAs bestätigt.

Des Weiteren fokussierte sich diese Studie auf eine Detailanalyse der T2D-assoziierten Varianten im *PPARG* Locus. Mithilfe von PMCA, Daten über epigenetische Marker sowie die in dieser Studie durchgeführten LC-MS/MS Messungen wurden die *cis*-regulatorische Variante rs4684847 im *PPARG* Locus sowie der TF PRRX1, welcher die PPARγ2-Expression in Adipozyten reguliert, gefunden. Es wurde beschrieben, dass mehrere kausale Varianten innerhalb eines gegebenen Locus die Genexpression kooperativ beeinflussen können. Eine integrative Analyse aus PMCA sowie Daten über epigenetische Marker, gekoppelt mit einer hochempfindlichen, markierungsfreien Proteomik ermöglichte die Identifizierung der *cis*-regulatorischen Variante rs7647481 im *PPARG* Locus, und des nicht-Risiko allel-spezifisch bindenden TFs YY1 (Ying Yang 1) sowie dessen Koregulator RYBP (RING1 und YY1 bindendes Protein). Die pathophysiologische Relevanz dieser Ergebnisse wurde durch eine Korrelation des Expressionsniveaus in primärem Fettgewebe mit einer verbesserten Insulinempfindlichkeit bei Patienten, die das rs7647481 nicht-Risiko Allel tragen, unterstützt. Darüber hinaus weisen weitere experimentelle Studien auf eine deutlich erhöhte Bindung von TFs und verwandten Proteinen zu vorhergesagten *cis*-regulatorischen versus nicht *cis*-regulatorischen Varianten hin.

Schlussendlich zeigen die Ergebnisse im *PPARG* Locus die hohe Effizienz des nicht-gezielten Proteomik-Ansatzes und unterstützen die Vorhersage kausaler Varianten durch eine integrative Kombination verschiedener Methoden. Darüber hinaus unterstützen diese Erkenntnisse die kürzlich vorgeschlagene "Multiple Enhancer-Hypothese". Die erfolgreiche

Identifizierung sowohl von TFs und deren Kofaktoren ermöglichen es allel-spezifische Protein-DNA-Interaktionsnetzwerke abzuleiten, die als Basis für die Vorhersage, wie genetische Varianten in Regulationsmechanismen Genexpressionsprofile und menschliche Phänotypen verändern, dienen können. Des Weiteren ermöglicht die beschriebene ungezielte Proteomik-Methodik die Analyse in primären humanen Geweben, und sollte für verschiedene Formen genetischer Variabilität, wie z.B. somatische Mutationen bei Krebs, angewandt werden können und kann somit helfen, die Rolle sowohl von vererbten als auch somatischen Varianten zu verstehen.

3. Introduction

3.1 Type 2 diabetes and obesity

Type 2 diabetes (T2D) and obesity are serious health risk worldwide with increasing prevalence. Over the last decade, there has been a dramatic increase in the number of global prevalence of T2D due to population growth, aging, urbanization, and increasing prevalence of obesity and physical inactivity [1]. In 2013, 383 million individuals were diagnosed with T2D in the world and the number is expected to increase up to 592 million by 2035 [2]. The World Health Organization (WHO) estimated that in 2008, there were more than 1.4 billion overweight adults and of these, over 500 million were obese adults. Overweight and obesity result in adverse metabolic effects on blood pressure, cholesterol, triglycerides and insulin resistance. Yearly, at least 2.8 million people die from overweight or obesity-associated diseases [3]. T2D and obesity are complex diseases, and they are closely related to each other since obesity stands out as the largest risk factor for developing T2D [4](reviewed in Guilherme et al. 2008 [5]). However, it was also reported that some lean T2D subjects are probably with Latent Autoimmune Diabetes in Adults (LADA) [6]. In addition, during a majority of patients with T2D are obese or overweight in developed countries, those are often non-obese or lean in India [7,8]. Conversely, many obese subjects do not develop diabetes via a compensatory increase in insulin secretion [9]. In this regard, a new phenotype, "metabolically healthy obese" (MHO) has been identified within 30% of the obese population without demonstrable obesity-related metabolic abnormalities such as dyslipidemia or impaired glucose tolerance (reviewed in Karelis et al. 2011 [10])[11]. The MHO type was rather seen to be related to younger age and a more peripheral fat distribution [12]. Nevertheless, abundant evidence suggests that along with environmental and multiple genetic factors, obesity is one of the main risk factors for developing T2D [13,14]. While environmental influences on development of T2D have been well studied in a variety of clinical trials [15-17], a large part of research on the genetic factor of T2D remains to be fully understood. The recent progress of genome wide association studies (GWAS) yielded increasing evidence demonstrating the impact of genetic factors on T2D susceptibility [18], and may provide more detailed insights into the molecular mechanisms underlying T2D and related other human diseases.

3.1.1 Type 2 diabetes (T2D)

Diabetes mellitus refers to a group of metabolic disorders of heterogeneous etiology characterized by abnormally high levels of blood glucose (hyperglycemia) with disturbances of carbohydrate, fat and protein metabolism as a result of defects in insulin secretion, impaired effectiveness of insulin action or both [19,20]. The diabetes mellitus is classified by underlying causes: type 1 diabetes (T1D), type 2 diabetes (T2D), gestational diabetes and monogenic diabetes [19]. Type 1 and type 2 diabetes are a vast majority of diabetes mellitus. T1D is insulin-dependent diabetes mellitus and usually developed in early childhood [21], which is caused by an auto-immune system attack on the insulin-producing β-cells in the pancreas, leading to defect in insulin secretion [22]. In contrast, T2D is independent on exogenous insulin and can stay undetected for many years, and occur normally in adulthood. T2D is often associated with obesity (reviewed in Guilherme et al. 2008 [5], Barrett-Connor et al. 1989 [23]) contributing to development of insulin resistance and relative insulin deficiency in multiple organs and tissues including digestive system, pancreas, brain, liver, muscle and adipose tissue [20,24,25]. Insulin is a hormone produced by the pancreatic β-cells and stimulates glucose uptake in muscle and adipose tissue. Insulin-stimulated glucose uptake occurs mainly in skeletal muscle where glucose is converted to glycogen, and about 10 % of which occurs in adipose tissue where energy is stored as triglycerides. Triglycerides are released from adipocytes in form of free fatty acids (FFAs) and are utilized as an energy source by other tissues. In addition, insulin can inhibit hepatic glucose production by inhibiting gluconeogenesis and glycogenolysis. The molecular basis for control of glucose homeostasis and energy balance by adipocytes is not completely understood, but is partially known to be modulated / regulated by actions of adipokines such as leptin, adiponectin, resistin and retinol-binding protein 4 (RBP4). Adipose tissue is well-established to have an additional endocrine function by secreting adipokines and lipids which communicate with other organs including the liver, muscle and brain (reviewed in Leto et al. 2012 [24]). Finally, hypothalamus, a small brain area located under the anterior commissure plays a critical role in the regulation of energy and glucose homeostasis via hormones such as insulin and leptin. These two hormones are released in proportion to body fat mass and act in the brain to maintain energy balance by circulating in the body (reviewed in Morton et al. 2007 [26]) (Fig. 1).

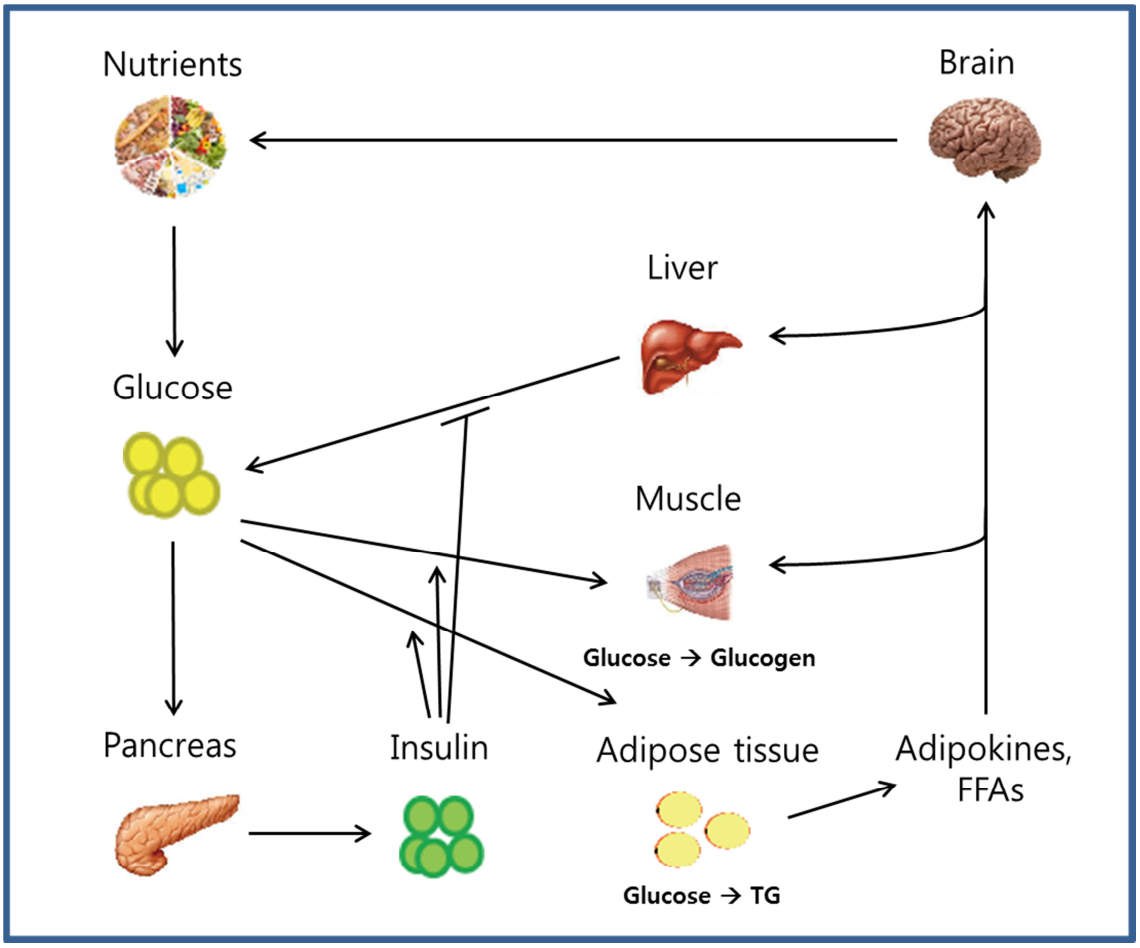


Figure 1. Regulation of glucose and insulin action in multiple organs and tissues (Redrawn and modified after Leto and Saltiel 2012 {Leto 2012 #1493})**.** Energy homeostasis depends on the concerted regulation of glucose and insulin action in various organs such as pancreas, brain, liver, muscle and adipose tissue. Defects in the insulin-dependent glucose metabolism result in metabolic diseases such as T2D. FFAs: free fatty acids; TG: triglycerides

Disruption of this energy balance and inability of tissues to respond to insulin can result in metabolic diseases such as T2D. A long-term exposure to T2D can affect major organs including heart, blood vessels, nerves, eyes and kidneys. The reduced quality of life, and increased morbidity and mortality of T2D patients are generally caused by the associated complications of those diseases. Overall, severe complications accompanied by T2D are mostly divided by two groups: microvascular (e.g. retinopathy (eye), neuropathy (nerve) and nephropathy (kidney)) and macrovascular (e.g. cardiovascular disease (heart)) diseases. Diabetic retinopathy might be the most common microvascular complication of T2D

(reviewed in Leto et al. 2014 [24]). In follow-up studies 9 years after diabetes diagnosis in the United Kingdom, 28% of T2D and 24% of T1D patients developed diabetic retinopathy [27]. It was also reported that 11% and 52 % of deaths in T2D patients (non-insulin-dependent) are caused by renal and cardiovascular disease (CVD) throughout the world, respectively [28]. As with other complications, risk of developing diabetic neuropathy [29] and nephropathy [30] is associated to both magnitude and duration of hyperglycemia and hyperlipidemia.

3.1.2 Obesity

Changes in lifestyle resulted in dominant increase of the number of overweight or obesity individuals [3]. In turn, the increasing number of obese in the population is highly correlated with the prevalence of T2D, cardiovascular disease and cancer [31]. Being overweight or obese is defined by using weight and height to calculate a number called the body mass index (BMI). If BMI is in range to 25-29.9 kg/m^2, it´s considered as overweight, above 30 kg/m^2 as obesity. Obesity is a consequence of energy imbalance caused from increased food intake, non-healthy food and reduced physical activity. In addition, obesity is a cause of 14-20 % of all cancer deaths (reviewed in Aggarwal et al. 2010 [32]). Food intake is regulated via circuits of neural signals (such as leptin and insulin) located in the hypothalamus. In the presence of leptin and insulin, food intake is repressed via anabolic neural circuits, and energy expenditure is increased via an interaction with specific leptin receptors located in the hypothalamus. Massive obesity is closely associated with absence of circulating, functionally active, leptin and insulin (reviewed in Schwartz et al. 2000 [33])[34]. However, some individuals developed obesity despite of the tight control of hypothalamus energy balance [35], supporting a role of genetic factors in obesity (see chapter 3.2). In addition, obese individuals often show increased serum FFA concentration which causes defective glucose metabolism through insulin resistance development. Insulin resistance is clinically important due to its close association with several diseases including T2D, hypertension, dyslipidemia and abnormalities in blood coagulation and fibrinolysis [36,37]. In obesity, glucose uptake is reduced and fatty acid uptake is elevated by the liver, skeletal muscle and pancreatic β-cells. Reduced glucose uptake elevates in turn glucose in blood, stimulating further insulin secretion. Continuous secretion of high amount of insulin leads to metabolic stress in pancreatic β-cells mitochondria via release of reactive oxygen species being able to damage mitochondria, contributing to apoptosis of β-cell and irreversibly reducing insulin secretion potential (reviewed in Westley et al. 2013 [38])[39].

3.1.3 Risk factors for T2D and obesity

The etiology of T2D is multifactorial influencing several different defects of glucose homeostasis, primarily in muscle, β-cells, liver and adipose tissue [40]. In the last decade, researchers have given effort to unravel the causes of T2D development. Several risk factors have been associated with T2D and include: biological factors (age, gender, family history, genetic, ethnicity), environmental factors and lifestyle (obesity, physical inactivity, excessive caloric intake and smoking) (Fig. 2) (reviewed in Noble et al. 2011 [41]). Previously, the prevalence of T2D was shown to increase with increasing age due to insulin resistance [42–46]. Additionally, the development of T2D tends to be related to the sex-differences [44,47]. Both T1D and T2D are strongly linked to family history, however T2D is also dependent on environmental factors such as lifestyle due to children learning habits [48]. Studies of twins demonstrated that a substantial genetic component plays a crucial role in T2D development [49] with heritability estimates of 75–85% for *in vivo* insulin secretion, ~50% for peripheral insulin sensitivity, and ~50% for glucose metabolism [50]. For offspring with a single diabetic parent, risk for T2D was 3.5-fold higher and for those with two diabetic parents was 6-fold higher than with offspring without parental diabetes [51]. Thus, such evidence indicates that family history is a more powerful T2D predictor that likely captures the genetic and environmental determinants of T2D risk [52]. Although risk factors like age, sex, family history explain development of T2D to some degree, there are large differences in the individual susceptibility to T2D when other risk factors are similar. It was also shown that not all obese people develop T2D [9,53] or often non-obese people develop T2D [7,8], which are more likely due to genetic factors [54]. Recently, linkage analysis, candidate gene approach, large-scale association studies and genome-wide association studies (GWAS) have successfully identified multiple genes that contribute to T2D susceptibility [55]. This great advance in the technologies of analysis provides us the insights into the pathogenesis of the T2D and obesity. The influences of ethnic differences on the prevalence of T2D have been well demonstrated previously [56–59], showing that certain ethnic groups have more or less prevalence of T2D [56,57]. Although environmental factors such as food intake, physical activity and obesity appear to differ clearly in various ethnic groups, genetic factors may play a more determinant role. To elucidate the pathophysiologic mechanisms responsible for the heterogeneous relationship between obesity and T2D in various ethnicities may give some clues to better understand the complex mechanisms involved in the development of T2D [59]. Although there are several

causes contributing to T2D development, T2D is primarily caused by obesity [14]. About 80 percent of all T2D patients are also diagnosed as obese, providing an interesting clue to the link between diabetes and obesity (reviewed in Astrup et al. 2000 [60]). Obesity itself is caused by a combination of genetic, lifestyle and environment factors [61–63]. The correlations between BMI and other phenotypic indices of obesity (skinfold thickness, waist circumference (WC) and waist-to-hip ratio (WHR)) have been reported with high statistical evidence for some loci relevant to human obesity and causal genes [64](reviewed in Herrera et al. 2011 [65]). Environmental factors and lifestyle are also important risk factors for the T2D development such as dietary pattern, physical activity [66–68] and smoking [69] whereas consumption of coffee [70] and of alcohol [69,71] have been reported to be inversely associated with the risk of T2D in a dose-dependent manner.

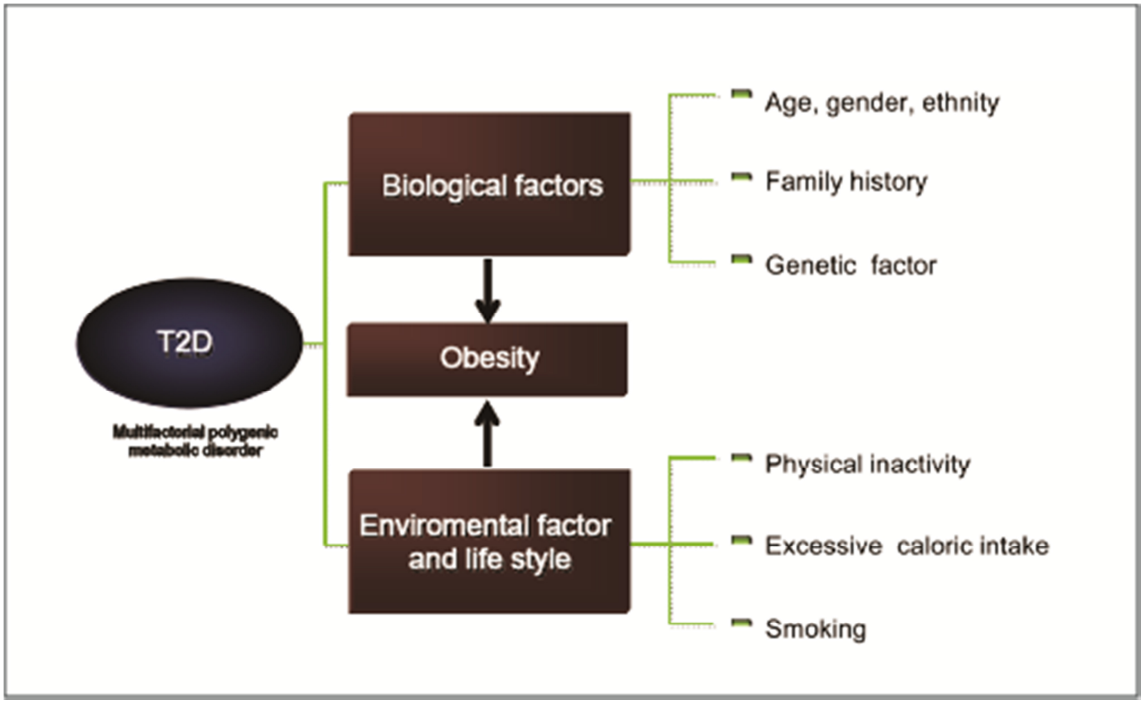


Figure 2. Type 2 diabetic risk factors. A number of factors are known to contribute to the development of T2D. The impact of risk factors is amplified by biological factors (age, gender, family history, genetic, ethnicity) as well as environmental factors and lifestyle (physical inactivity, excessive caloric intake and smoking). Obesity is the most important predictor and cause of T2D, and is also influenced by biological, environmental factors and lifestyle.

3.2 Genetics of T2D and obesity

3.2.1. Genetic mapping in human diseases

Genetic mapping is a powerful approach to identify genes and biological processes that increase susceptibility to human diseases or affect physiological traits. The methodology is based on the localization of genes underlying phenotypes on the basis of correlation with DNA variation, without the need for prior hypotheses about biological function [72]. Such approach was conceived by Sturtevant for fruit flies in simplest form, called linkage analysis in 1913 [73]. By the early 1900s, geneticists understood that some traits are inherited according to Mendel's ratios in case, as a result of alterations in single genes. They also recognized that even if some traits are inherited according to Mendel's laws, most naturally occurring trait variation involves the action of multiple genes and nongenetic factors while it indicates strong correlation among relatives [72]. The classical method of mapping disease genes is to use the long stretches of linkage disequilibrium (LD) in affected families by performing linkage analysis (reviewed in Ahlqvist et al. 2011 [74]). In genetics, LD is a measure of the co-occurrence of particular alleles at two loci in a population.

Various statistic methods have been used to measure the amount of LD between two alleles, one of the most useful being the coefficient of correlation r^2. When $r^2 = 1$, the two alleles are in complete LD, whereas values of $r^2 < 1$ indicate that the ancestral complete LD is eroded (reviewed in Rahim et al. 2008 [75]). Disease loci can be mapped on a genome-wide level by genotyping about 400–500 genetic markers. The affected family members share a certain marker more often than expected by chance that might be inherited from the same parent, indicating that a disease causing variant is in LD with the genotyped marker. Although geneticists applied genetic mapping successfully to common diseases inheritance mode, it has been less useful for complex diseases such as T2D. Very few genes with large phenotypic effect such as calpain 10 (*CAPN10*) and transcription factor 7-like 2 (*TCF7L2*) [76] were identified by linkage in common complex diseases (reviewed in Ahlqvist et al. 2011 [74]). This suggests that, for many common disorders the predominant pattern is that of multiple loci, individually with small effects on phenotype during complex traits differ in their underlying genetic architectures. For most human disorders, the sum of the identified genetic effects comprises only in part, generally less than half, of the estimated trait heritability [77]. The enormous investment in human genomics has been driven with the expectation that increasing genetic knowledge would translate into improved tools for the treatment and prevention of

disease. Indeed, in the case of obesity [78] and diabetes [79] genetic knowledge has improved the health and well-being of some people (reviewed in O'Rahilly et al. 2009 [80]). In the future, increasing knowledge of the genetic architecture of metabolic disease will give us benefits to human health, first, by discovering and validating key points for metabolic homeostasis, and human genetics will give some hints regarding the selection of molecular targets for novel therapeutics. Second, the reliable dissection of genetic and pathophysiological heterogeneity would be predictable within metabolic diseases, providing improvements in personalized diagnosis, prognostication, therapy and prevention (reviewed in O'Rahilly et al. 2009 [80]). Therefore, the identification of the causal genes or variants susceptible to human diseases remains a main challenge in human genetics.

3.2.2 Genome-wide association studies (GWAS)

Over the last decade, the marked advances in the field of genomics allowed to understand more about human phenotypic diversity and susceptibility to human diseases. Thus, human genetic variation was claimed as a "breakthrough of the year" by Science in 2007. Using single nucleotide polymorphisms (SNPs)-arrays based typing technologies and comparison of the frequency of SNP alleles between cases and controls in the population, the GWA approaches allowed the detection of numerous genetic loci associated with complex diseases with modest phenotypic effects in a systematic and unbiased manner [81,82]. These advances enabled the characterization of over 3.1 million human SNPs genotyped in 270 individuals from four geographically diverse populations by HapMap Project "Phase II haplotype map" in 2007 [83]. Using the collective data created by HGP (http://www.genome.gov/11006943), the SNP Consortium (http://snp.cshl.org) and the International HapMap Project (http://hapmap.ncbi.nlm.nih.gov), scientists were able to create rational design of GWAS [84]. GWAS have dramatically advanced by assessing from fewer than 100,000 SNPs to more than one million [55]. By 2013, over 1,700 GWAS have been published, reporting associations for over 11,000 SNPs with significant trait associations [85]. Genetic variants are classified in terms of their frequency within the population. In the population, alleles with a frequency greater than 5% are common variants while those with a frequency of 1%–5% are low frequent variants, and those less than 1% are rare variants. The majority of common genetic variants occurred once in human history and are shared by many individuals today through descent from common ancient ancestors (reviewed in Rahim et al. 2008 [75]).

The recent application of linkage analysis, candidate gene approach, large-scale association studies and GWAS in the genetics of human disease has resulted in the identification over 70 common risk loci conferring susceptibility to T2D and obesity. Among them, 45 loci were identified in European populations, and the other 29 loci were identified in East and South Asians populations [86]. Of the 70 loci identified by GWAS, several loci including *TCF7L2*, peroxisome proliferator-activated receptor gamma (*PPARG*), fat mass- and obesity-associated gene (*FTO*) have been confirmed through numerous GWAS analyses, which will be introduced in more detail in next chapters. It was previously reported that several diabetes and obesity- associated SNPs may affect the protein structure or gene expression [87–92]. However, most of the identified variants are located in non-coding DNA regions that might affect transcriptional regulation [93–99]. Recent studies indicated that the physical location of those variants in the genome give some clue to their ultimate biological effect, not only with the closely co-located functional gene [80,100], but also with genes at long distance [101,102]. A large number of diabetes SNPs are also close to highly expressed genes in the adult or developing pancreas, and many SNPs have been shown to be associated with reduced β-cell dysfunction in non-diabetic individuals [103–105]. Furthermore, SNPs associated with expression in disease-relevant tissues such as liver and adipose tissues are enriched for associating with T2D in humans [106].

Obesity is the most important predictor, and cause of both T2D and cardiovascular disease. Thus, genes closely associated with obesity are important candidates for T2D risk as well. Obesity is a highly heritable trait in ranging from 40%–70% for BMI [64]. A variety of GWAS showed strong associations between common variants at the *FTO* locus and BMI, often consequently causing T2D [107,108]. SNPs in the first intron of the *FTO* locus were the first identified variants showing the strongest associations with human obesity [107,109] (see chapter 3.4 for more details). The *FTO* gene is highly expressed in hypothalamus, and its expression is regulated by feeding and fasting [110]. Carriers for *FTO* risk variants tend to show an increased appetite or measured food intake [111], and thus the mechanism underlying the impact of the common genetic variants on obesity seems to be through energy intake. However, there are still many questions to be answered about the *FTO* gene function (reviewed in O'Rahilly et al. 2009 [80]). Subsequently, several other loci associated with BMI have been reported, such as melanocortin 4 receptor (*MC4R*); transmembrane protein 18 (*TMEM18*), glucosamine-6-phosphate deaminase 2 (*GNPDA2*), SH2B adaptor protein 1 (*SH2B1*), neuronal growth regulator (*NEGR1*); Potassium channel tetramerisation domain containing 15 (*KCTD15*), SEC16 Homolog B (*S. Cerevisiae*) (*SEC16B*), serologically defined colon cancer antigen 8

(*SDCCAG8*) and TRF1-interacting ankyrin-related ADP-ribose polymerase/Peptide methionine sulfoxide reductase (TNKS/MSRA) [112–117]. However, in most cases the precise functional roles of those loci in T2D, obesity and related traits are not yet fully understood.

3.3 *PPARG*

3.3.1 PPAR subfamily

Peroxisome proliferator activated receptors (PPARs) are transcription factors and belong to the nuclear receptors (NRs) protein family found in a variety of species. In the early 1990s, PPARs were discovered in rodents liver tissue, [118], in *Xenopus laevis* [119] and next in humans as the NRs for inducing peroxisome proliferation (reviewed in Michalik et al. 2004 [120]). This specific family of receptors constitutes of three subtypes including PPAR-α (NR1C1), PPAR-β/δ (NR1C2) and PPAR-γ (NR1C3) encoded by three distinct genes. These PPAR proteins consist of a common conserved domain structure which contains a highly conserved DNA-binding domain consisting of two zinc fingers at the N-terminal region, and a ligand (hormone)-binding domain (LBD) at the C-terminal region connected by a short hinge region. The A/B domain harbours the activation function-1 region (AF1) at the extreme N-terminal region, which is responsible for differences in the biological function between the three PPAR subtypes. The ligand binding domain (LBD) at the C-terminal region exhibits significant variation in amino acid residues, which results in each subtype being pharmacologically distinct. The activation function -2 region (AF-2) is located at the C-terminus of the LBD. The transactivation of AF-2 domains is generally ligand dependent, whereas AF-1 functions in a ligand- independent fashion. The N-terminus of each receptor is responsible for selective gene expression and function, in part, to limit receptor activity. Deletion of the N-terminus results in non-selective activation of target genes. The binding of ligands to the receptor triggers a conformational change within the LBD and the release of co-repressor proteins. It leads, in turn, to the association of co-activator proteins which mediate the transcriptional activation of target genes (reviewed in Savkur et al. 2006 [121]). However, the activation of genomic target genes by PPARs involves an intricate interplay between the properties of the subtype- and cell-type-specific settings at the individual target loci [122]. PPARs exert their effect in a form of heterodimers with 9-*cis*-retinoid X receptor α (RXRα) and bind to the specific DNA sequence termed as PPRE (peroxisome proliferator response elements) that are present in the promoter region of PPAR target genes. The sequence of PPRE is composed of a direct repeat of

AGGNCA interfered with a single nucleotide (DR-1). The heterodimerization of PPARs with RXRα has been experimentally demonstrated as ligand independent. In living cells, PPARs heterodimerize efficiently with RXRα in the absence and presence of ligand (**Fig. 3**) [123](reviewed in Savkur et al. 2006 [121]).

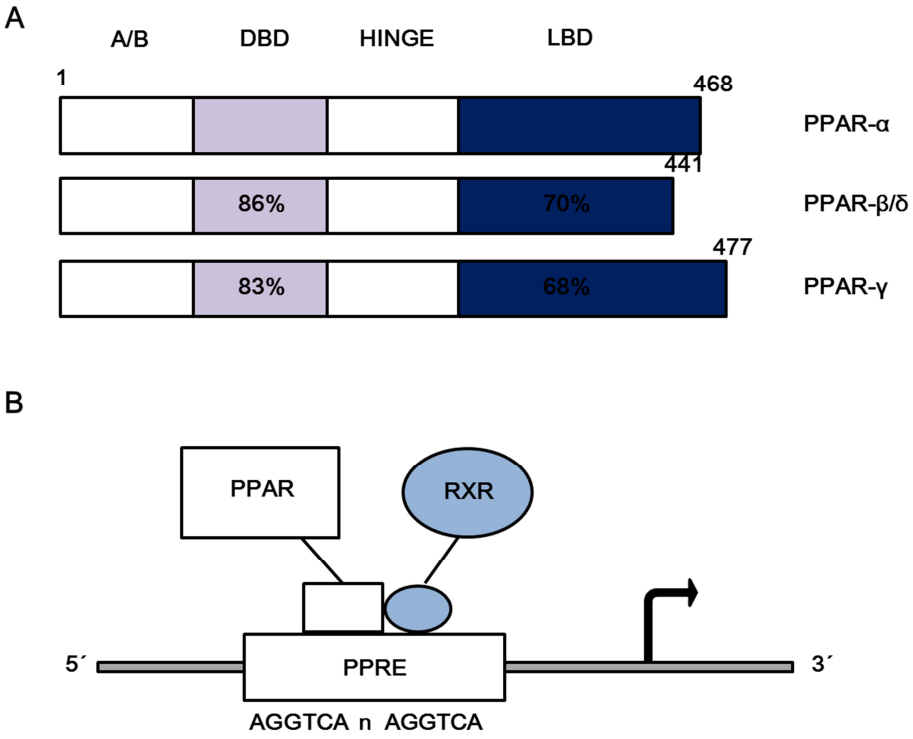


Figure 3. Structure of human PPAR subfamily (Redrawn and modified after Savkur et al. 2006 [121])**.** A. Domain structure of the human PPAR subtypes. The PPARs contain a common conserved domain structure including the N-terminal A/B domain, the central region harbouring the DBD, a hinge region and the C-terminal LBD. The two domains, DBD and LBD show the amino acid identity (%) with PPAR-α. B. The PPARs heterodimerize with the 9-*cis*-RXR and activate transcription of their target genes on binding to PPREs which comprise of a direct repeat of the element half site spaced by a single nucleotide (DR-1). DBD: DNA-binding domain; LBD: Ligand-binding domain; PPAR: Peroxisome proliferator-activated receptor; PPRE: PPAR response element; RXR: Retinoid X receptor.

PPARs are activated by endogenous ligands including fatty acids and fatty acid derivatives which are mainly derived from the lipoxygenase and cyclooxygenase (COX) pathways, and play a crucial role in the regulation of transcription involved in lipid and glucose metabolism (reviewed in Michalik et al. 2004 [120]). PPARs exhibit a high structural similarity, but they are

different from each other in their expression pattern and function. The first identified PPAR, PPARα is expressed in liver, kidney, heart, skeletal muscle, brown adipose, macrophages, vascular endothelial and vascular smooth muscle cells. PPARα plays an important role in intracellular lipid metabolism where it regulates the transcriptome of genes involved in reverse cholesterol transport and β-oxidation of FFAs. PPARα regulates transcriptionally several genes including ATP-binding cassette, sub-family A1 (*ABCA1*), fatty acid transport protein, fatty acid transporter (*FAT/CD36*), acyl-CoA oxidase, carnitine palmitoyltransferase I and the mitochondrial uncoupling proteins (*UCPs*). PPARα agonists were demonstrated to decrease lipid content in tissues, to minimize lipotoxicity and thereby to increase insulin sensitivity. Hence, PPARα is regarded as an ideal target for pharmaceutical agents to treat metabolic syndrome and T2D (reviewed in Savkur et al. 2006 [121]. PPAR β/δ is the latest identified PPAR subfamily and is expressed in several tissues including brown adipose tissues and skeletal muscle. PPAR β/δ serves as a regulator of reverse cholesterol transport, fatty acid catabolism and energy metabolism. In addition, PPAR β/δ is involved in the control of cell proliferation, cell differentiation and apoptosis. PPAR β/δ agonists lead to direct activation of glucose transport in primary human myotubes and improvement of insulin resistance and T2D in animal models, which are interesting targets for the development of agents to treat T2D (reviewed in Savkur et al. 2006 [121] and Michalik et al. 2004 [120]). The adipogenic potential of members of the CCAAT/enhancer-binding protein (C/EBP) family (α, β and δ) is well known to bind and *trans*-activate the promoters of a number of adipocyte genes. Notably, only PPARγ showed marked synergy with C/EBPα to activate adipocytes differentiation [124], suggesting the role of PPARγ in the adipogenesis. Besides the role of PPARs in adipocytes, PPARs have also been shown to have anti-inflammatory effects and to reduce the progression of atherosclerosis in animals and humans [125,126].

3.3.2 PPAR gamma (PPARγ)

3.3.2.1 *PPARG* gene and protein function

PPAR-γ is one of the most extensively investigated members of the PPAR family and is encoded by single gene located on the chromosome on 3p25 in human [127]. PPAR-γ is characterized in several species including mice [128–130], hamsters [131], rat [132,133] and humans [129,133–136] and is predominantly expressed in adipose tissue, and also found in heart, spleen and large intestine, placenta, and macrophages whereas it is barely detectable in muscle [135–138].

PPAR-γ is well known to regulate adipocyte differentiation as well as glucose homeostasis and insulin sensitivity in response to several structurally distinct compounds including thiazolidinediones (TZDs) and fibrates [139–141]. A class of PPARγ ligands such as TZDs including rosiglitazone and pioglitazone, has been applied in clinical practice for improving glycemic control via insulin sensitization in T2D patients [142,143]. Activation of PPARγ in adipocytes was shown to be sufficient for whole-body insulin sensitization equivalent to systemic TZD treatment [144]. PPAR-γ has been also described to be involved in regulation of inflammatory responses as a negative regulator of macrophage activation [137] and of immune system in T cells [145,146], B cells [147] and dendritic cells [148,149]. Other studies also demonstrated that PPARγ is involved in other diseases such as different human tumors [150–153], cardiovascular disease [154] and Alzheimer's Disease [155].

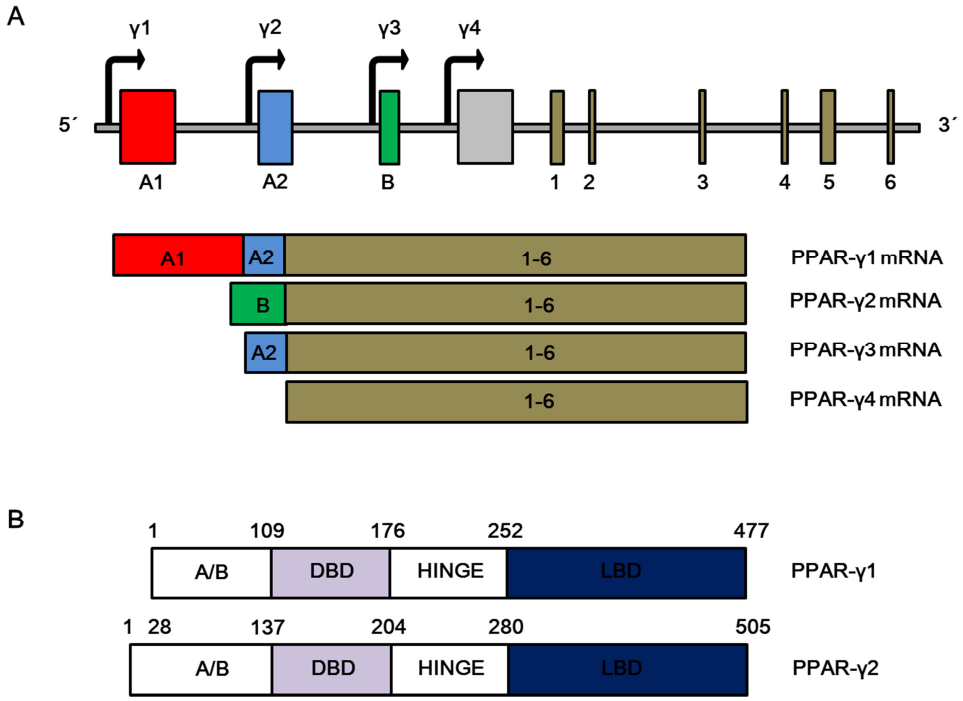


Figure 4. Schematic structure of human *PPARG* gene on chromosome 3p25 (Redrawn and modified after Al-shali et al. 2004 [156] and Sabatino et al. 2012 [157])**.** A. Schematic structure of *PPARG* gene and four *PPARG* mRNA isoforms. The black arrows indicate the transcription start sites for each specific mRNA isoform. Exons are indicated as boxes on the genomic map. Exons A1 and A2 are untranslated, and exon B is translated [156]. B. *PPARG-1*, *-3*, and *-4* mRNA isoforms are translated into the unique 477-amino-acid PPARγ1 protein whereas the *PPARG2* mRNA translates into a 505-amino-acid protein with 28 extra amino acids at the N-terminal end. A/B: variable region; DBD: DNA-binding domain; LBD: Ligand-binding domain; PPAR: Peroxisome proliferator-activated receptor; PPRE: PPAR response element; RXR: Retinoid X receptor (reviewed in Sabatino et al. 2012 [157]).

Differential promoter usage and alternative splicing of *PPARG* gene result in four isoforms, *PPARG-1*, *-2*, *-3* and *-4* [134,135,156,158]. The *PPARG* mRNA isoforms and promoters are illustrated in Fig. 4A. *PPARG-1*, *-3*, and *-4* mRNA isoforms are translated into the identical 477-amino-acid protein. In contrast, *PPARG2* mRNA gives rise to a 505-amino-acid protein with 28 additional N-terminal amino acids, indicating that the differences in the promoters and non-coding exons result in differential tissue expression [156]. *PPARG1* mRNA is relatively widely expressed in several tissues including heart, intestines, colon, kidney, pancreas, spleen and skeletal muscle [135,136] whereas *PPARG2* mRNA is most highly expressed in adipose tissue and regarded as a key isoform related to adipose metabolism [135,136]. *PPARG3* mRNA expression is restricted to adipose tissue and large intestine [134], and primer extension studies reported that *PPARG4* mRNA is also present in adipose tissue, which might be important for fat depot distribution and metabolism *in vivo* [156,159]. Among isoforms of *PPARG* mRNA, *PPARG-1* and *-2* have been most intensively studied. Several studies demonstrated the differential regulation of *PPARG* isoforms, especially *PPARG1* and *PPARG2* in metabolic networks via basal cell differentiation [160]. The tissue-specific distribution of isoforms and the variable ratio of *PPARG1* to *PPARG2* differ in individuals, suggesting that isoform expression might be modulated in disease states like non-insulin-dependent diabetes mellitus [136]. It was also reported that during early stages of adipocyte differentiation *PPARG2* mRNA was increased whereas *PPARG1* mRNA remained unchanged. Moreover, a C/EBP site was identified only in the human *PPARG2* promoter, indicating that *PPARG2* may initiate adipocyte differentiation [161]. In addition, two new exons, exon C and exon D at the *PPARG* locus were established by Zhou and his colleagues in monkeys, showing sequences identical to exons C and D in the human chromosome 3. These additional exons generate four novel *PPARG* subtypes, *PPARG-4*, *-5*, *-6*, and *-7*. *PPARG4* and *PPARG5* were found only in macrophages while *PPARG6* and *PPARG7* were expressed both in macrophages and adipose tissues [158]. Later, the presence of *PPARG-4*, *-5*, and *-7* transcripts in human THP-1 macrophages was confirmed by Reverse transcription polymerase chain reaction (RT-PCR) and sequencing while transcript corresponding to *PPARG6* was not detected in human [162].

3.3.2.2 Genetic variants at *PPARG* locus

PPARG is the first identified gene for reproducible association with T2D [163]. A number of *PPARG* variants have been identified to be associated with T2D [164–169], but the most widely

studied variant of the *PPARG* gene is the Pro12Ala (rs1801282) [170,171]. The Pro12Ala variant shows substitution of a proline for alanine at position 12 of the protein which is observed in about 12% of the European population [172]. The Pro12Ala is located in exon B of the *PPARG* gene expressing the splice variant *PPARG2* mRNA and encoding the protein target for the thiazolidinedione class of drugs used to treat T2D [170]. The Pro12Ala variant has been also shown to be associated with reduced transcriptional activity, increased insulin sensitivity and protection against T2D [163,173,174]. In meta-analysis of 60 association studies involving 32,849 T2D cases and 47,456 controls, the T2D odds ratio (OR) for the Pro12Ala variant was 0.85 using a fixed-effects model and 0.86 using a random-effects model (comparing alanine to proline) [175]. These results confirm findings from previous meta-analyses [170,176–179] that the *PPARG* Ala12 variant was associated with a reduced T2D risk, especially in lean individuals. Also, in other study using a Pro12Ala knock-in mouse model, Ala/Ala mice on chow diet were leaner and showed improved insulin sensitivity, plasma lipid profiles and longer lifespans [180]. However, Ek et al. showed an association of the Ala12 variant with increased weight gain in obese individuals [181]. Taken together, these results indicate that gene-environment interactions play a key role as high-fat feeding eliminates the beneficial effects of the Pro12Ala variant on adiposity, plasma lipids and insulin sensitivity [180]. Thus, the Pro12Ala variant of the *PPARG2* gene appears to be sensitive to environmental influence, and the influence of PPARG on diabetes might, at least in part, be highly dependent on gene-environment interactions such as exercise and dietary context [182–185]. Of note, it is a paradoxical fact that the minor Ala12 allele, associated with enhanced insulin sensitivity in humans blunts the transcriptional activity of the insulin-sensitizing PPAR-γ2 transcription factor. This fact suggests that the elusive *PPARG* T2D signal instead of the Pro12Ala comes from regulatory variants that affect *PPARG2* expression [98]. Indeed, in the work of Claussnitzer and colleagues we demonstrated that the adverse effect of PRRX1 homeobox factor on lipid metabolism and systemic insulin sensitivity, dependent on the rs4684847 C risk allele that initiates PRRX1 binding. This provides a valuable contribution to the translation of genetic association signals to disease-related molecular mechanisms [98].

3.4 *FTO*

3.4.1 *FTO* gene and protein function

The Fat mass and obesity associated (also known as FTO) gene was first cloned after identification of a fused toe (Ft) mutant mouse, whose phenotype resulted from a 1.6-Mb deletion of six genes, including Fto on mouse chromosome 8 [186,187]. In human, the *FTO* gene located on chromosome 16q12.2 encodes a protein consisting of 9 exons and spanning more than 400kb [188]. *FTO* mRNA is widely expressed in a wide range of tissues, especially in the brain, but also in skeletal muscle, liver and adipose tissue [110,188–191]. In a study using a wild-type mouse model, *FTO* mRNA was most abundant in the brain, particularly in hypothalamic nuclei governing energy balance, which was regulated by feeding and fasting state. It suggests the role of FTO in controlling food intake [110], which is consistent with the results of Stratigopoulos et al. [188]. *FTO* mRNA expression showed an age- and BMI-dependent regulation in skeletal muscle whereas it was regulated by age and sex in subcutaneous adipose tissue. Moreover, the age-dependent *FTO* expression was shown to be associated with peripheral defects of glucose and fat metabolism [191]. A study in mouse brain suggested that *FTO* gene expression levels showed a strong negative correlation with expression levels of neuropeptides in the hypothalamus, which in turn is known to impact feeding behavior [192]. In contrast, other study on Fto-deficiency mice showed no significant change of neuropeptide expression including agouti-related protein (*Agrp*), neuropeptide Y *(Npy)* and proopiomelanocortin (*Pomc*) in the fed state, but slight reduction of both *Npy* and *Pomc* expression in the fasted state, consistent with the lack of hypophagia in the Fto deficient mice. Moreover, Fto–knock-out mice displayed a reduced body length and leanness with mild improvement in insulin sensitivity and increased circulating adrenaline concentration in blood as a consequence of increased energy, despite of reduced spontaneous locomotor activity and relative hyperphagia [193]. Although scientists had been struggling for years to understand the function of FTO, its role has been not fully defined. Previous study using sequence analysis suggested that *FTO* gene encodes for a protein 2-oxoglutaratedependent nucleic acid demethylase involved in fatty acid metabolism, DNA repair, and posttranslational modifications for example, proline hydroxylation and histone lysine [194]. *In silico* analysis of the human *FTO* sequence revealed that the *FTO* gene seems to have sequence homology with Fe(II)- and 2-oxoglutarate (2OG) oxygenases that catalyze oxidative reactions on multiple substrates using non-heme iron as a co-factor and 2OG as a co-substrate [110,195]. Recent studies

demonstrated that recombinant FTO has efficient oxidative demethylation activity targeting the abundant N6-methyladenosine (m6A) residues in RNA *in vitro* [196]. In *in vitro* differentiated primary human preadipocytes and in human Simpson-Golabi-Behmel Syndrome (SGBS) preadipocytes, expression of *FTO* gene and its nearby gene *FTM* was down-regulated during adipogenic differentiation [189]. Other study in FTO deficient mice demonstrated that FTO deficiency resulted in a prominent reduction of adipocyte size [193], which is consistent with the results of Tews et al. [197]. Moreover, FTO-deficient SGBS adipocytes exhibited the increased expression of uncoupling protein 1 (UCP-1), thereby inducing a brown adipocyte phenotype [197]. These data suggest that FTO might have a role not only in brain, but also in other tissues.

3.4.2 Genetic variants at *FTO* locus

Over the past years, several independent GWAS demonstrated that variants in the first intron of the *FTO* gene are highly associated with T2D and BMI, suggesting that the *FTO* locus exerts its primary effect on obesity and subsequently impact on T2D [107,164]. Non-coding variants within a 47-kilobase (kb) region of high LD in introns 1 and 2 of the *FTO* locus have been implicated as the strongest genetic association with risk to obesity in human [198–202]. Common variants in the first intron of the *FTO* locus (e.g rs9939609) showed an increased risk predisposing to both childhood and adult obesity through an effect on BMI. SNP rs9939609 represents a cluster of 44 SNPs in the first intron of the *FTO* gene that are highly correlated with each other ($r^2 > 0.8$, 1000 Genomes CEU data [172]). Moreover, subjects homozygous for the risk allele tend to be weighing approximately 3 kg more and having a 1.67-fold increased risk of obesity compared to those homozygous for the protective allele. These results were replicated in 13 cohorts with 38,759 participants [107]. These findings were replicated in different study populations and supported the association between the *FTO* gene variants and obesity in such populations [203–205]. However, Wing et al. demonstrated that there is no significant evidence of the association in African Americans, suggesting that the effect of the *FTO* variants on adiposity phenotypes may show some genetic heterogeneity dependent on ethnicity [206].

The association for the *FTO* variant (rs9939609), both with regard to BMI and T2D was replicated in several studies [111,199,202,207]. Interestingly, the significant association of rs9939609 with T2D (odds ratio 1.13 [95% CI 1.06 –1.20], $P = 9 \times 10^5$) was abolished after adjusting for BMI (1.06 [0.97–1.16], $P = 0.2$) [207]. Contrary, a meta-analysis from the

Scandinavian HUNT, MDC, and MPP studies suggested that the association between rs9939609 and T2D was strong after adjustment for age and sex (OR 1.13 [95%CI 1.08–1.19]; *P* = 4.5 3 1028) and remained significant after BMI correction (OR 1.09 [95% CI 1.04–1.15]; *P* = 1.2 3 1024) [208]. The *FTO* variant (rs9939609) was reported not to be involved in the regulation of energy expenditure, but may have a role in the control of food intake and food choice, suggesting an association with a hyperphagic phenotype or a preference for energy-dense foods [111]. In several studies was shown that dietary factors [209] and physical activity [207,209] may accentuate the susceptibility to obesity by the *FTO* variants. During the *FTO* gene confers risk for T2D in Caucasian, with rs9939609 A allele increasing BMI by approximately 0.4 kg/m^2 [107], other studies in Asian [210] and African population [211] showed no significant association between *FTO* gene variants and BMI or obesity. In addition, several studies failed to show association between *FTO* expression and obesity associated *FTO* genotype, rs9939609 in adipose tissue and skeletal muscle [191,212] and rs8050136 in adipose tissue [213]. Despite of the wide investigation of the association of the *FTO* variants with obesity, it is not fully understood how the *FTO* gene variants exert their effect on obesity. Interestingly, Smemo et al. recently reported that these obese-linked *FTO* introns do not interact with the promoter for *FTO* in adult mouse brain. Instead, these introns are functionally connected with the promoter for homeobox *IRX3* at megabase distance. Consistent with these findings, the analysis of the ENCODE data also revealed interaction between IRX3 and the obesity association interval in the first intron of the *FTO* gene, but no interaction between the *FTO* promoter and the association interval, suggesting that the obesity-associated *FTO* intron mediates functional interactions with IRX3 in the human, mouse and zebrafish genomes. Moreover, for IRX3 function in regulation of body mass and composition a reduction in body weight of 25 to 30% in Irx3-deficient mice was observed, suggesting that IRX3 is a functional long-range target of obesity-associated variants within the *FTO* gene and represents a novel determinant of body mass and composition [102]. However, possible functional implications of long-range gene regulation from the *FTO* locus in other tissues such as adipose remain elusive. The *FTO* risk variants are shown to be associated with other diseases or traits including polycystic ovary syndrome (PCOS) [214,215], Alzheimer's disease [216], acute coronary syndrome [217], myocardial infarction [218], melanoma [219] and end-stage renal disease (ESRD) [220]. However, several studies suggested that these effects may to be secondary to weight increase since the associations are abolished or attenuated after adjusting for BMI or other factors [107,221–223].

3.5 *TCF7L2*

3.5.1 *TCF7L2* gene and protein function

The transcription factor-7-like 2 (also known as TCF4), encoded by *TCF7L2* is a high mobility group box-containing family of transcription factors influencing the transcription of several genes, thereby exerting a large variety of functions within the cell [224]. As a component of the bipartite transcription factor β -catenin/TCF, TCF7L2 plays a crucial role in conveying Wnt signaling during embryonic development and in regulating gene expression during adulthood [225]. Aberrations in the Wnt signaling pathway may lead to the development of diseases in humans such as congenital malformations, cancer, osteoporosis and T2D [226–229]. The Wnt signaling pathway negatively regulates adipogenesis. Mice overexpressing adipose tissue-specific Wnt-10b displayed up to ~50% lower adipose mass and were resistant to HFD-induced obesity. Wnt-10b null mice exhibited increased adipogenic potential during repressive Wnt ligand Wnt-5b was shown to promote adipogenesis (reviewed in Ip et al. 2012 [230]). Moreover, TCF7L2 was shown to be involved in stimulating proliferation of pancreatic β-cells and production of the incretin hormone glucagon-like peptide-1 in intestinal endocrine L cells [225]. TCF7L2 also regulates proglucagon gene expression by β-Catenin and glycogen synthase kinase-3β in enteroendocrine cells [231].

The *TCF7L2* gene spans 215.9 kb comprising 17 exons and includes two major domains: a catenin-binding domain (exon 1) and a central DNA-binding HMG domain (exons 10 and 11) [232]. In total, 17 exons were identified, of which 5 were alternative. The alternative use of 3 consecutive exons localized in the 3' part changes the reading frames used in the last exon, leading to the synthesis of a number of human *TCF7L2* isoforms with short, medium, or long-size COOH-terminal ends. It suggests functional significance of TCF7L2 due to its ability to interact functionally with C-terminal binding protein (CtBP), a corepressor protein required to mediate transcriptional repression of the TCF family activity [233]. The *TCF7L2* gene is highly expressed in most human tissues including mature pancreatic β-cells, with the exception of the skeletal muscle [234,235]. In obese T2D patients, *TCF7L2* expression is significantly decreased in the subcutaneous and omental fat compared with obese normoglycemic individuals [235]. In other study it was shown that *TCF7L2* expression in human islets was increased 5-fold in the islets of pancreas in T2D patients, and overexpression of *TCF7L2* in human islets reduced glucose-stimulated insulin secretion [224].

3.5.2 Genetic variants at *TCF7L2* locus

In the past decade, researchers have been given huge efforts to find T2D associated genes through numerous candidate-gene studies and fine-map linkage signals. Genetic fine mapping is to identify potentially causal variants such as *cis*-regulatory variants modulating gene expression, which contribute to increase susceptibility of diseases or to affect phenotypes. The classical linkage analysis was used in mapping T2D-causal genes, but mostly not successful except two genes, *CAPN10* and *TCF7L2* (reviewed in Ahlqvist et al. 2011 [74])[236]. The *TCF7L2* locus was mapped to chromosome 10q in both Icelandic and Mexican-American populations [237,238]. In 2006, Grant and his colleagues fine-mapped using a microsatellite marker DG10S478 throughout a 10.5 Mb interval on 10q in Icelandic population and identified various common T2D susceptibility variants within intron 3 in the *TCF7L2* gene, which were replicated in Danish and US cohorts [76]. The association between T2D and a number of variants in the *TCF7L2* gene has been demonstrated in different populations [235,239–250], implicating SNP rs7903146 of *TCF7L2* as a most significant determinant of T2D [234].

In Danish and American cohorts, the heterozygous and homozygous individuals for the risk-associated alleles showed relative risks of T2D of 1.45 and 2.41 compared to the non-risk allele carriers, respectively [76], which was replicated in numerous subsequent studies (reviewed in Pang et al. 2013 [251]). The risk T allele of rs7903146 was associated with impaired insulin secretion, incretin effects and enhanced rate of hepatic glucose production [224]. However, the precise mechanisms by which variants in the *TCF7L2* gene increase the risk of developing T2D, or which variants play a role in alternative splicing, gene expression, or protein structure remain to be fully understood [251]. Several studies suggest the enteroendocrine role of *TCF7L2* in the pathogenesis of T2D (reviewed in Grant et al. 2006 [76]) and the involvement of *TCF7L2* in the colorectal carcinogenesis [252]. TCF7L2 null mice were shown to die within 24 h after birth due to the depletion of an intestinal epithelial stem cell compartment [253]. Moreover, the progucagon gene (glu) encodes glucagon, expressed in pancreatic islets and the insulinotropic hormone production of glucagon-like peptide-1 (GLP-1), expressed in the intestines. These two hormones exhibit critical and opposite effects on blood glucose homeostasis, and TCF7L2 is involved in the regulating proglucagon gene transcription and the GLP-1 in the intestinal endocrine L cells [231], suggesting that the *TCF7L2* variants may increase the risk of T2D by affecting the production of the incretin hormone GLP-1 [76]. In a study of TCF7L2 the most obvious potential target tissue is the pancreatic islets. A study of Gaulton et al. demonstrated that a *TCF7L2* intronic variant strongly associated with T2D (rs7903146) is

located within epigenetic marks of regulatory region in human pancreatic islets. Moreover, human islet samples heterozygous for the rs7903146 showed allele-specific effects and increased enhancer activity for risk allele (T allele) compared with nonrisk allele (C allele), indicating that genetic variation at this locus acts in *cis* with local chromatin and regulatory changes [254,255]. It was confirmed by the fact that an increased risk of T2D may be associated with overexpression of TCF7L2 in specific tissues [256]. Savic et al. demonstrated that T2D-association interval (92-kb) harboring *cis*-regulatory elements regulates the spatial–temporal expression patterns of *TCF7L2*, including expression in tissues involved in the control of glucose homeostasis. Moreover, by the selective deletion of the T2D-associated interval, the enhancers situated within the association interval were shown to be critical for robust *TCF7L2* expression in tissues regulating glucose metabolism. In support of a role for *cis*-regulatory variation in T2D susceptibility, they also showed that a null Tcf7l2 allele led, in a dose-dependent manner, to lower glycemic profiles. Furthermore, Tcf7l2 null mice (Tcf7l2$^{-/-}$) displayed enhanced glucose tolerance coupled to significantly lowered insulin levels, suggesting that these mice are protected against T2D. These observations confirm the role of variation in *cis*-regulatory elements in T2D susceptibility and strengthen the evidence that *cis*-regulatory variants may be a paradigm for genetic predispositions to common disease [256].

4. Experimental approach for identification of allele-specific binding proteins

4.1 Requirements for identification of allele-specific binding proteins

GWAS improved largely the understanding of the genetic components of complex traits by identification of numerous SNPs associated with phenotypic traits and diseases [55]. From a large set of genome-wide variants, GWAS investigated to identify a few SNPs that are statistically significantly associated with human complex diseases or traits. One of the key challenges of GWAS data interpretation is to identify causal SNPs and provide profound evidence on the mechanism how they affect the traits. Currently, researches are focusing on identification of candidate causal variants from the most significant SNPs of GWAS-identified loci. However, researches based on classical GWAS approach are limited to annotate SNPs to nearby genes [257]. To date, in most cases there is lack of a functional link between variants and development of diseases or traits because the majority of identified SNPs are located in non-coding regions without obvious expected phenotype or function. On the other hand, due to the tight genetic linkage of SNPs in a haplotype-block, only few of

these variants identified by one GWAS signal are likely to exert a functional effect [258]. Recently, there has been an increased effort to identify non-coding genomic elements that regulate gene transcription [55,93,100,259–262]. The most frequent genomic elements affected are transcriptional enhancers and silencers. These elements typically regulate transcription through long-range interactions, mediated by the formation of chromatin loops. The identification of expression quantitative trait loci (eQTL) can be used for predicting the target genes of *cis*-regulatory variants, however, this approach usually provides only indirect evidence of an association. Thus, other experimental approaches would be required for representing certain biological mechanistic relevance. There are more direct methods such as 3C and its derivatives. Chromatin conformation capture (3C) has already been used for successful identification of target genes of several regulatory variants identified by GWAS (reviewed in Edwards et al. 2013 [263]). Subsequently, a number of different 3C- derivatives have been developed to overcome the limitations of 3C and answer different biological questions, including 3C/3C-qPCR, 3C-seq/4C-seq, 4C (3C-on-a chip), Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET), 5C (3C carbon copy) and Hi-C and Targeted Chromatin Capture (T2C). However, such methods are often time-consuming, cost-intensive and limited to the amount of available cell material [264].

The knowledge of allele-specific protein binding will also give important clues since the majority of regulatory functions such as chromatin looping and transactivation are mediated through transcription factors (TFs) and other proteins (reviewed in Edwards et al. 2013 [263]). The importance of a coordinated interaction between TFs, coregulators, and the basal transcriptional machinery for regulation of gene expression in metabolism has been well documented [265]. To date, there are few published examples in which TFs affect gene expression in allele-specific manner [98,266–269]. However, understanding of the mechanisms of allele-specific expression is still incomplete. Transcript based studies alone are not enough to resolve the mechanisms that regulate these genes and cause allele-specific expression. Therefore, there is a great need to find effective approach analyzing the interactions between specific regulatory sequences, transcriptional regulators and chromatin structures [270].

4.2 Analysis of DNA-protein interaction

Protein-DNA interaction is critical to the life of cells. The interactions between proteins and nucleic acids are prevalent in many biological processes such as recognition of specific nucleotide sequences, regulation of transcription, regulation of gene expression, and

chromosome assembly and disassembly (reviewed in Chu et al. 2014 [271])[272–274]. Besides the RNA polymerases, there are histones, chromatin remodeling proteins, general transcription factors, their cofactors, and a host of sequence-specific TFs that directly initiate transcription to specific promoters (reviewed in Helwa et al. 2010 [275]). In terms of GO biological process, transcription regulator activity and nucleic acid binding are significantly over-represented in the oncogenes in cell cycle, cell-growth and/or maintenance and developmental processes [276]. In addition, a variety of proteins have been identified with specific or general affinity to DNA (reviewed in Helwa et al. 2010 [275]). However, there are still numerous proteins involved in gene regulation, DNA repair and oncogenesis that are not yet fully understood. Thus, taken the importance of functions of DNA-binding proteins in cells, development of an unbiased, proteome-wide analytical approach in order to identify DNA-binding proteins was necessary. A number of techniques, both computational and experimental have been developed to identify DNA–binding proteins and model interactions, such as electrophoretic mobility shift assay (EMSA), Nitrocellulose filter binding assay, footprinting, Methylation interference assay, Chromatin immunoprecipitation (ChIP), DNA adenine methyltransferase identification (DamID), Surface plasmon resonance (SPR), systematic evolution of ligands by exponential enrichment (SELEX), yeast one-hybrid system, microarray-based assays [272–275] and published phylogenetic module complexity analysis (PMCA) [98] .

A classical method used to detect DNA-protein complexes is EMSA. EMSA is based on the principle that the electrophoretic mobility of a protein-DNA complex is dependent on their size, charge and to some extent, shape. EMSA is useful as qualitative and quantitative assay for the characterization of protein-nucleic acid interactions. This basic technique is simple to perform and sensitive. Since this assay uses radioisotope-labeled nucleic acids, it is highly sensitive, enabling assays to be performed with small protein and nucleic acid concentrations in small sample volumes. There are also other variants available using fluorescence, chemiluminescence and immunohistochemical detection. A wide range of nucleic acid sizes (lengths from short oligonucleotides to several thousand nt/bp) and structures (single-stranded, duplex, triplex and quadruplex nucleic acids as well as small circular DNAs) can be used in EMSA assays. However, samples could be not at chemical equilibrium during the electrophoresis step. Rapid dissociation during electrophoresis can prevent detection of complexes while even slow dissociation can result in underestimation of binding density. Moreover, many complexes are often present in other physiological solution which could result in more significant stability of complexes *in vitro* than *in vivo*. Furthermore, these

complexes provide little direct information about the location of the nucleic acid sequences bound by proteins [277]. Other techniques are also available for the detection and characterization of protein-nucleic acid complexes such as nitrocellulose filter-binding [278,279] and footprinting [280–282]. Filter binding is sensitive, simple to perform, and the procedure is rapid enough to allow both, kinetic studies and equilibrium measurements [283]. Moreover, the required equipment of filter binding is inexpensive, and like the EMSA this assay is a non-equilibrium technique. This assay is also not limited by salt concentration of the sample and is useful for a very large range of nucleic acids (e.g., the phage λ genome (48,502 bp) [284,285]. However, this assay is not suitable for the detection of more than one protein-nucleic acid complex [283]. Footprinting assay is based on the principle that a specific nucleic acid sequence will be labeled with radioactivity or fluorescence at the 3' or 5' end followed by incubation with protein extracts. The nucleic acid sequences will be cleaved using chemical or enzyme reagents that will cut at protein-free locations while nucleic acid templates bound to proteins will be protected. At the end, electrophoretic sequencing gels will be performed to analyze footprinting of nucleic acid fragments [282]. The advantage of this assay is to provide the information about nucleic acid sequences within or near protein binding sites. The appearance of binding sites that are hypersensitive to modification can additionally provide evidence of conformational change in the target nucleic acid [286]. There are however, some limitations in using DNase I that does not cut DNA randomly, and its activity is affected by local DNA structure and sequence resulting in an unequally fragments. In turn, it prevents the precise prediction of the protein binding site on DNA molecule [287,288]. As the bound protein does not protect the DNA, the gel can be difficult to interpret, which could alter the photoreactions in the vicinity [289].

While EMSA, filter binding and footprinting assays are usually limited to evaluate *in vitro* interactions by incubating protein extracts with labeled DNA probes, ChIP is employed to explore the binding and interaction of post-translationally modified histones or transcription factors with specific DNA sequences *in vivo* using specific antibodies to proteins of interest [274]. ChIP is very efficient and specific from the use of approximate antibody, and there is no need of further PCR amplification for the study of the precipitated DNA [290]. However, the success of ChIP assay depends critically on both, a *priori* knowledge of target proteins and the access to the respective high quality antibodies. In addition, ChIP assay is mainly limited to known biological targets and low throughput [274]. The traditional ChIP assay is combined with PCR to identify the sequence identities of the precipitated DNA fragments. However,

PCR method can only be used to identify known target genes for a given protein (reviewed in Mundade et al. 2014 [291]). In addition, the traditional ChIP assay requires large numbers of cells (~10 million), which can be especially challenging in small model organisms. Therefore, recent advances in ChIP method have overcome such limitations, and complementary assays have been developed including Chromatin immunoprecipitation coupled with microarrays (ChIP-chip) or short-tag sequencing (ChIP-seq). The ENCODE Consortium has performed hundreds of ChIP-seq experiments [100] and has used this experience to develop a set of working standards and guidelines [292] (reviewed in Furey et al. 2012 [293]). These assays are not appropriate to identify unknown DNA interacting factors or to study the dynamics of gene regulation, a complex process requiring the interaction of numerous factors [274]. Reece-Hoyes and his colleagues introduced gateway-compatible yeast one-hybrid (Y1H) assay providing a convenient gene-centered (DNA to protein) approach to identify TFs that can bind a DNA sequence of interest. They showed Y1H resources including clones for 988 of 1,434 (69%) predicted human TFs that can be used for detection of both, known and new interactions between human DNA regions and TFs. YIH assay simplifies the mapping of human gene-centered regulatory networks. Thus, the human enhanced Y1H (eYIH) pipeline will be a powerful complement to TF–centered methods such as ChIP and ChIP-seq, by enabling large-scale characterization of the DNA-binding activity of transcription factors that may be expressed or active only under restricted conditions, or in a few cells. These resources can also be useful for mating or direct DNA transformations of one or a few human DNA baits in small-scale studies [294]. However, there is a major limitation to this system that frequency of false positives generated by yeast endogenous TFs appeared to be higher than that of true positives [295]. False positive interactions may be produced when multiple members of the same TF family with highly similar consensus binding sites bind to the same enhancers, and only a subset of these actually bind to the enhancer *in vivo.* The rate of false negatives in eY1H assay is relative high. TFs that mostly interact with DNA as heterodimers or after post-translational modification by another human protein will not be found in this system. Furthermore, eY1H assay is unable to detect cooperative interactions with multiple TFs [296]. In addition, when TFs are in low abundance, they are likely to be not detected and the information about post-translational modification or cofactors of TFs is missing [297]. Moreover, TFs require post-translational modifications or cofactors that are not available in yeast for binding that specific promoter. Also, the chromatin context in the yeast nucleus can interrupt the detection of some interactions [298].

In the past decade, there have been remarkable advances in proteomic technologies. The evolution of mass spectrometry (MS)–based proteomic technologies has greatly facilitated *in-depth* characterization of the protein components of biological systems and enabled deep insights into the composition, regulation and function of molecular complexes, and pathways to address diverse biological questions. Moreover, MS-based proteomics allowed the analysis and identification of proteins in high throughput (reviewed in Schulze et al. 2010 [299] and Mallick et al. 2010 [300])[301,302]. Previously, the advances of MS resulted in the identification of hundreds of unknown proteins as nuclear proteins that are potentially involved in the regulation of gene expression, DNA replication and repair [303]. In principle, all mass spectrometers measure biological samples according to the mass-to-charge ratio (m/z) of freely moving gas-phase ions in electric and/or magnetic fields (reviewed in Soldi et al. 2013 [304]). The great advances in protein MS has been driven by the development of soft protein ionization methods such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) [299]. Proteins and peptides are polar, nonvolatile species that require an ionization method to transfer them into the gas-phase, without extensive degradation and significant loss of sample integrity. Currently, MALDI ionization is commonly used in combination with other mass analyzer, ALDI/time-of-flight (TOF). ESI is usually coupled "on-line" with liquid chromatography (LC) instrument to achieve continuous or high throughput analysis. For instance, reverse phase high-pressure liquid chromatography (RP-HPLC) has been widely adopted in proteomics to resolve very complex peptide mixtures prior to MS analysis (LC-MS) due to its high resolution, efficiency, reproducibility and mobile phase compatibility with ESI. A further development of this technology is nano-ESI that the flow rates are lowered to a nanoliter-per-minute regime to improve the sensitivity of the method, and is compatible with capillary RP-HPLC columns allowing high sensitivity (reviewed in Soldi et al. 2013 [304]). For MSn (n refers to sequential number of MS) approaches, peptides are first selected for fragmentation inside the mass spectrometer and then are fragmented by one of several methods such as collision-induced dissociation (CID) or electron capture detection (ECD). Typically, the most intensive ions are selected for fragmentation. Tandem MS approach (MS/MS) is now most widely used among MSn, which enables the measurement and the identification of peptides at a rate of thousands of sequences per day with better than femtomole sensitivity (10^{-15} mol, or subnanogram) in complex biological samples (reviewed in Mallick et al. 2010 [300]). Therefore, MS/MS is a key technique for protein or peptide sequencing and post-translational modifications (PTM) analysis (reviewed in Soldi et al. 2013 [304]). MS technique itself is limited to the protein identification.

Hence, in order to explore protein-DNA interactions, the protein of interest has to be isolated and purified from the mixtures prior to the identification of proteins by MS technique. Each protein purification step usually results in some degree of protein loss. Therefore, an ideal protein purification strategy has to have the fewest steps with the highest purity of protein. The selection of purification method is dependent on various properties of the target proteins such as size, charge and solubility [305].

4.3 Affinity chromatography using magnetic beads

The isolation of proteins and peptides is often performed using a variety of chromatography, electrophoretic, ultrafiltration, precipitation and other procedures such as affinity chromatography (AC) (reviewed in Safarik et al. 2004 [306]). Protein affinity purification coupled to MS technique is currently the most used technology for the systematic isolation and identification of the DNA-binding proteins (reviewed in Gingras et al. 2007 [307], Kocher et al. 2007 [308] and Musso et al. 2007 [309]) and more recently for the study of transient complexes [310,311]. The strength of column AC has been shown in thousands of successful applications, particularly in the laboratory scale. Such standard column LC procedures are difficult to handle with the samples containing particulate material, so they are not suitable for work in early stages of the isolation and purification process where suspended solid and fouling components are present in the sample. In this case, applications of magnetically stabilized fluidized beds or magnetically modified two-phase systems are advantageous (reviewed in Safarik et al. 2004 [306]). The identification of DNA-binding protein using magnetic beads system has previously described in several studies [98,312–316], and benefits from its magnetic properties have been reported for use in diverse bioapplications [317–319]. The basic principle of batch magnetic separation is very simple and fast with only few handling steps. Samples can be used directly as crude cell lysates, whole blood, plasma, ascites fluid, milk, whey, urine, cultivation media, wastes from food and fermentation industry, and many others (reviewed in Safarik et all. 2004 [306]). The binding of a DNA template to protein of interest occurs in a few minutes, magnetic separation takes seconds, and washes and elution take a few minutes. The magnetic beads have high stability, low particle-particle interaction and high dispersibility. For example, the biotin-streptavidin complex is extremely strong ($K_{association\ rate\ constant}$ (K_{ass}) = 10^{15} M^{-1}) and resistant to high concentrations of salt and urea. Moreover, with the same starting material, it yields for 30 min. procedure higher purity than the normal method using columns and requiring several days [316]. All steps of the purification procedure can take place

in one single test tube or another vessel. Moreover, there is no need for expensive liquid chromatography systems, centrifuges, filters or other equipment. The separation process can be performed directly in crude samples containing suspended solid material. Magnetic beads can be relatively easily and selectively removed from the sample, which is successful for recovery of small magnetic particles (diameter ca. 0.1 – 1 μm) in the presence of biological debris and other fouling material of similar size. Moreover, this system is in particular useful for large scale operations and for high-throughput approaches which can principally be automated (e.g. magnetic-particle based immunoassay). Magnetic separation is usually less stringent to the target proteins or peptides. While large protein complexes tend to be broken up by traditional column chromatography, they may remain intact in magnetic system (reviewed in Safarik et al. 2004 [306]). To further advance, surface-enhanced Raman spectroscopy (SERS)-encoded magnetic bead was successfully applied to flow cytometry separation analysis, which is useful for drug screening, medical diagnostics, or combinatorial chemical synthesis with the advantage of being a simpler, more convenient, and cost-effective method [320].

4.4 Label-free proteomic analysis using LC-MS/MS

Accurate and reliable protein quantification is necessary for understanding basic biological responses as well as discovering valuable biomarkers for disease treatment and diagnosis. Advances in instrumentation (e.g. modern mass spectrometers), computing power and bioinformatics have enabled identification of a large number of proteins in biological samples. However, the accurate quantitation of proteins which are differentially expressed has remained a challenge [321,322]. While previous MS techniques were so far dominated by qualitative identification of proteins in biological sample mixture, recently quantitative measurement of MS is widely available and contributes to generate comprehensive and quantitative information on protein modifications which requires several experimental approaches, significant amounts of pure starting material and special expertise and time (reviewed in Mallick et a. 2010 [300]). In systems biology, such qualitative and quantitative properties of MS have enabled the advances in development of mathematical models of the behavior of pathways and networks in several model organisms [323–325]. For example, MS allowed the identification of thousands of phosphorylation sites in a quantitative manner and significantly contributed to the present knowledge of signaling pathways [326]. Quantitative protein-DNA proteomics, coupling affinity chromatography with Liquid Chromatography-

tandem Mass Spectrometry (LC-MS/MS) was reported for the identification of enhancer-binding proteins [312,314] and enabled the identification of allele-specific DNA binding proteins [98,269,312,327]. Generally, LC-MS/MS is not quantitative due to different physical and chemical properties of different tryptic peptides. Differences in charge state, peptide length, amino acid composition or posttranslational modifications result in great differences in ion intensities for the peptides, even from the same protein (reviewed in Schulze et al. 2010 [299]). Thus, most studies based on the LC-MS/MS approach have often required stable isotope labeling techniques, e.g. with 15N, 18O, stable isotope labeling by amino acids in cell culture (SILAC) and isotope-coded affinity tags (ICAT) to provide relative quantification [328]. Protein extracts were first labeled either metabolically with SILAC [312] or chemically with ICAT [314]. Stable isotope labeling combined with MS provides the greatest accuracy for the protein quantification. However, despite its high sensitivity and accuracy, labeling with stable isotopes requires time-consuming processing, high-cost or inefficient labeling [329], may be limited by missing data points due to under-sampling. In addition, some labeling process involves complex processes which could cause artifacts [328,330,331], and possible signal caused by co-eluting components of similar mass could interfere with the precise protein quantification in complex samples [332]. It also can provide computational difficulties to reliably define the isotopic "pairs" for relative quantification since there could be differences in the LC elution time of the labeled forms, incomplete mass spectrometric resolution of the isotopic pairs [333,334], or the presence of other unresolved components [328]. Such difficulties of quantitative proteomics with isotope labeling encouraged researchers to develop an alternative approach, i.e. label-free quantitative proteomics that is attractive due to its simplicity and low costs [322]. Ideally, samples for label-free comparisons are run consecutively on the same LC-MS/MS setup to avoid variations in ion intensities due to differences in the system setup (column properties, temperatures) and thereby to allow precise reproduction of retention times. The complexity of the sample is not increased by the mixing of different proteomes. Therefore, label-free approach usually exhibits high analytical depth and dynamic range when large, global protein changes between treatments are expected (reviewed in Schulze et al. 2010 [299]). Label-free quantification is based on spectral counting or on precursor signal intensity. Spectral counting takes account into the abundance of a protein using the number of distinct peptides observed or the number of times a peptide sequenced from a protein. In this approach, peptides from more abundant proteins could be more sequenced and identified than peptides from less abundant proteins. In addition, quantification accuracy decreases significantly when only a few peptides are observed for a given protein. Thus, spectral

counting is not exceptionally sensitive to small changes in abundance and cannot provide information on the change in abundance of a peptide relative to a protein, such as frequently arises by truncation or modification of a protein. In contrast, peak intensity is a more direct measurement of abundance than the spectral count, and thus offers some advantages, i.e. linearity and accuracy (reviewed in Mallick et al. 2010 [300]). Wang and his colleagues evaluated the reproducibility and linearity of this approach using various amounts of proteins from highly complex proteomes in p53-deficient HCT-116 human cells. More than 50% and nearly 90% of the peptide ion ratios deviated less than 10% and 20%, respectively, in duplicate runs. Algorithms for outlier-resistant mean estimation and for adjusting statistical significance threshold in multiplicity of testing were applied to reduce the rate of false positives [322]. The label-free approach has no general dynamic range limitation while fold changes of protein abundance are greater than ~20:1, measured in isotope-labeled samples with very large errors in ratio determination [329]. In addition to these advantages, it allows high coverage of quantified proteins and applications to any biological material that may be crucial factor for the use of human materials (reviewed in Schulze et al. 2010 [299])[328]. Due to the reproducibility and accuracy of the label-free proteomic approach, it is especially suitable for biomarker discovery in large sample sets [321].

5. Material and Methods

5.1 Material

5.1.1 Cell lines

The following cell lines were obtained from the American Type Culture Collection (ATCC): 3T3-L1 (Mouse white pre-adipocyte cell line, adherent), SGBS (Human pre-adipocyte cell line, adherent), 293T (Human embryonic kidney cell line, adherent), Huh7 (Human hepatocyte cell line, adherent), INS-1 (Rat pancreatic beta cell line, adherent) and C2C12 (Mouse myoblast cell line, adherent). HIB 1B (Mouse brown pre-adipocyte cell line, adherent) cell line was kindly provided by B. Spiegelman (Harvard university, USA). Primary human pre-adipocytes were obtained from all patients who donated biological samples. Medical ethical committee approval for this study was obtained from 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.

5.1.2 Probes and primers

The oligonucleotides used for electrophoretic mobility shift assay (EMSA), affinity chromatography and real time quantitative PCR are listed. All oligonucleotides used in this study were purchased from Eurofins MWG Oligo Synthese-Report (Ebersberg, Germany), except for YY1 (Santa cruz, CA, USA).

Oligonucleotides	**Strand**	**Length (bp)**	**Modification**
Sequence (5´→ 3´)			
PPARG			
rs4684847 (long)	for	40	Cy5-/Biotin-
TTTAAATCATCTCTAATTCT[C/T]ACAACTCCGAAAAGATAAG			
rs4684847 (short)	for	31	Cy5-/Biotin-
TTTAAATCATCTCTAATTCT[C/T]ACAACTCCGA			
rs7647481	for	40	Cy5-/Biotin-
CAACTCCCCCACTTTATTCC[A/G]TGATGTTCAGACCCAGCCA			
rs17036342	for	40	Cy5-/Biotin-
GCTCTCCCAAAGAATTGTAA[A/G]TTCCCAGAGTGTAGGACCA			
rs2881479	for	40	Cy5-/Biotin-

GCAAGACTCTGTCTCAAAAA[A/T]AAATAAATAAATAAATAAA			
rs13085211	for	40	Cy5-/Biotin-
TTGCGGTGAGCTATGATCGC[A/G]CTGCTGCACTCCAGCCTGG			
rs1801282	for	40	Cy5-/Biotin-
TGGGAGATTCTCCTATTGAC[C/G]CAGAAAGCGATTCCTTCAC			
Sp1	for	22	Cy5-/Biotin-
ATTCGATCGGGGCGGGGCGAGC			
1-10	for	27	Cy5-/Biotin-
GAAGAATTCATGCAAATGAATTCGAAGAAG			
Scramblea	for	40	unmodified
AGCAAACCCTGACTAGTTATAGAGTCAAGACCGCCCACTT			
PPRX1b	for	22	unmodified
GTCGTAACTAATTAACTAGGAC			
YY1c	for	27	unmodified
CGCTCCCCGGCCATCTTGGCGGCTGGT			
MyoD	for	22	unmodified
CCCCCAACAGCTGTTGCCTGA			
CdxA	for	31	unmodified
GCATTTTATTACCACGCCTGCACTGTTGGTA			
FTO			
rs1421085 (long)	for	62	Cy5-/Biotin-
AATATTGATTTTATAGTAGCAGTTCAGGTCCTAAGGCATGA[T/C]ATTGATTAAGTG TCTGATGA			
rs1421085 (short)	for	36	Cy5-/Biotin-
GGTCCTAAGGCATGA[T/C]ATTGATTAAGTGTCTGATGA			
TCF7L2			
rs7903146	for	45	Cy5-/Biotin-
AGAGCTAAGCACTTTTTAGATA[T/C]TATATAATTTAATTGCCGTATG			

Table 1a. Probes used for EMSA, affinity chromatography in this study.

[a]predicted as non-specific binding site which contains random oligonucleotide sequence (designed by Dr. Bernward Klocke, Genomatix, Germany), [b]PRRX1 consensus sequence (adapted from MatBase tool, Genomatix, Germany), [c]YY1 consensus sequence (Santa cruz biotechnology, USA).

Target genes	**Primer (forward)**	**Primer (reverse)**
Sequence (5´→ 3´)		
PPARG1	CGTGGCCGCAGATTTGA	AGTGGGAGTGGTCTTCCATTAC
PPARG2	GAAAGCGATTCCTTCACTGAT	TCAAAGGAGTGGGAGTGGTC
YY1	CGAGTTCTCGGTCACCATGT	CTGCCAGTTGTTTGGGATCT
RYBP	CTGCACCTTCAGAAACAGTGC	GTGCCACCAGCTGAGAATTG
Leptin	ACACGCAGTCAGTCTCCTCCAA	TGGAAGGCATACTGGTGAGGAT
HPRT	TGAAAAGGACCCCACGAAG	AAGCAGATGGCCACAGAACTAG

Table 1b. Primers used for qPCR amplification in this study.

5.1.3 Bacterial strain

All transformation experiments were performed with the chemical competent *E. coli* strain DH5α (C-2988) purchased by NEB (New England Biolabs, Hitchin, U.K.). All bacterial cells were stored in 50µl aliquots at 80°C.

5.1.4 Plasmids

Name	Insert vector	Source
pcDNA 3.1	Empty	M. Kern, USA
FLO6	Human PRXX1 cDNA into pcDNA 3.1	M. Kern, USA
pcMV6-XL4/5/6	Empty	M. Kern, USA
pcMV6-XL5 PRRX1	Human PRXX1 -flag cDNA into pcMV6-XL4/5/6	M. Kern, USA
pcDNA 3.1 (-)	Empty	M.Klar, Germany
pcDNA 3.1 (-)-hYY1	Human YY1 cDNA into pcDNA 3.1 (-)	M.Klar, Germany
TK control	752 bp thymidine kinase (TK) promoter into TK control	M.Claussnitzer, Germany
rs4684847 C/T	40 bp allelic DNA for rs4684847 at mid-position TK control	M.Claussnitzer, Germany
rs7647481 G/A	40 bp allelic DNA for rs7647481 at mid-position TK control	M.Claussnitzer, Germany
rs2881479 A/T	40 bp allelic DNA for rs2881479 at mid-position TK control	M.Claussnitzer, Germany
rs17035342 A/G	40 bp allelic DNA for rs17035342 at mid-position TK control	M.Claussnitzer, Germany

Table 2. Plasmids used for transformations in this study.

5.1.5 Antibiotics

Name	Source
Penicllin-Streptomycin (P/S)	PAA Laboratories, Pasching, Austria
Gentamicin	Roth, Karlsruhe, Germany
Ampicillin	Sigma, Steinheim, Germany

Table 3. Antibiotics used in this study.

5.1.6 Antibodies

Name	Host	Specificity	Code	Source
PRRX1	Rabbit	H, M		M. Kern, USA
TF1	Rat			B. Kempkes, Germany
YY1 (C-20)	Rabbit	H, M, R, C, B, P, A, E	sc-281	Santa Cruz Biotechnology, CA, USA
Sp1 (PEP 2)	Rabbit	H, M, R, C, B, P, A	sc-59	Santa Cruz Biotechnology, CA, USA
normal IgG	Mouse		sc-2025	Santa Cruz Biotechnology, CA, USA
normal IgG	Rabbit		sc-2027	Santa Cruz Biotechnology, CA, USA

Table 4. Antibodies used in this study.

Abbreviations: H =Human, M =Mouse, R =Rat, C=Canine, B=Bovine, P=Porcine, A=Avian, E=Equine.

5.1.7 siRNAs

Name	Source
ON-TARGETplus Non-targeting Control Pool	Dharmacon, CO, USA
ON-TARGETplus Human RYBP (23429) siRNA - SMARTpool	Dharmacon, CO, USA
ON-TARGETplus Human YY1 (7528) siRNA - SMARTpool	Dharmacon, CO, USA

Table 5. siRNAs used in this study.

5.1.8 Buffers, solutions and agar plates

Name	Composition
Affinity chromatography (magnetic beads)	
Binding and Wash buffer 2x (2x B&W)	
	10 mM Tris-HCl, pH 7.5
	1 mM EDTA
	2 M NaCl
	Dest. H_2O
Binding and Wash buffer with biotin 1x (1x B&W with biotin)	
	1x B&W
	2 ng/µl biotin
Binding buffer 5x w/o salt (5x BB w/o salt)	
	20% (v/v) glycerol
	5 mM $MgCl_2$
	2.5 mM EDTA
	2.5 mM DTT
	50 mM Tris-HCl, pH 7.5
	Dest. H_2O
Elution buffer	
	0.8x BB w/o salt
	different conc. of NaCl (50 – 1000 mM)
	Dest. H_2O
Wash buffer	
	1x BB w/o salt
	different conc. of NaCl (10 – 50 mM)
	Dest. H_2O
Affinity chromatography (sepharose beads)	
Blocking buffer	
	0.2 M glycerol
	Dest. H_2O, final pH ~8.0
Coupling buffer	
	0.1 M $NaHCO_3$
	0.5 M NaCl
	Dest. H2O, final pH ~8.4
Heparin affinity (HA) buffer	
	20 mM HEPES, pH 7.9
	different conc. of KCl (0 – 1000 mM)
	5 mM $MgCl_2$
	0.5 mM DTT
	0.1% (v/v) protease inhibitor cocktail (3.7x Complete, mini)
	0.1% (v/v) phosphatase inhibitor cocktail (10x Phosstop)

	100 µg/ml insulin
	8% (v/v) glycerol

EMSA

Gel binding buffer 5x (5x GBB)	
	20% (v/v) glycerol
	5 mM $MgCl_2$
	2.5 mM EDTA
	2.5 mM DTT
	250 mM NaCl
	50 mM Tris-HCl, pH 7.5
	Dest. H_2O
Gel binding buffer 4x (4x GBB)	
	40 mM HEPES, pH 7.9
	4 mM EDTA, pH 8.0
	1mg/ml BSA
	800 mM KCl
	4mM DTT
	16% (v/v) Ficoll
	Dest. H_2O
Loading Buffer 10x	
	250 mM Tris-HCl, pH 7.5
	0.2% (w/v) orange G
	40% (v/v) glycerol
	Dest. H_2O
Tris/Borate/EDTA (TBE) buffer 5x	
	445 mM Tris Base
	445 mM Boric acid
	10 mM EDTA, pH 8.0
	Dest. H_2O, final pH ~8.3

PBS

instamed PBS Dulbecco w/o Mg^{2+} , Ca^{2+}

Protein extracts

High salt buffer (HSB)	
	20 mM HEPES, pH 7.9
	1.5 mM $MgCl_2$
	1.2 M KCl
	20 mM NaF
	0.2 mM EDTA
	added fresh to the buffer:
	0.5 mM DTT
	25% (v/v) glycerol
	8% (v/v) protease inhibitor cocktail (3.7x Complete, mini)
	1% (v/v) phosphatase inhibitor cocktail (10x Phosstop)
	Dest. H_2O
Homogenization buffer	
	10 mM HEPES, pH 7.9
	1.5 mM $MgCl_2$
	10 mM KCl
	20 mM NaF
	added fresh to the buffer:
	0.5 mM DTT
	8% (v/v) protease inhibitor cocktail (3.7x Complete, mini)
	1% (v/v) phosphatase inhibitor cocktail (10x Phosstop)
	Dest. H_2O

Low salt buffer (LSB)	
	20 mM HEPES, pH 7.9 1.5 mM MgCl2 20 mM KCl 20 mM NaF 0.2 mM EDTA added fresh to the buffer: 0.5 mM DTT 25% (v/v) glycerol 8% (v/v) protease inhibitor cocktail (3.7x Complete, mini) 1% (v/v) phosphatase inhibitor cocktail (10x Phosstop) Dest. H_2O
Schreiber buffer A (hypotonic buffer) for 293T, Huh7, C2C12 and INS-1 cell lines	
	10 mM HEPES, pH 8 10 mM KCl 0.1 mM EDTA 0.1 mM EGTA added fresh to the buffer: 0.5 mM PMSF 1 mM DTT 1.5% (v/v) protease inhibitor cocktail (7x Complete, mini) 1.5% (v/v) phosphatase inhibitor cocktail (10x Phosstop) Dest. H_2O
Schreiber buffer C (hypertonic buffer) for 293T, Huh7, C2C12 and INS-1 cell ines	
	20 mM HEPES, pH 8 0.4 M NaCl 1 mM EDTA 1 mM EGTA 20% (v/v) glycerol added fresh to the buffer: 1 mM PMSF 1 mM DTT 1.5% (v/v) protease inhibitor cocktail (7x Complete, mini) 1.5% (v/v) phosphatase inhibitor cocktail (10x Phosstop) Dest. H_2O
Schreiber buffer A (hypotonic buffer) for 3T3-L1 and SGBS cell lines	
	10 mM HEPES, pH 8 10 mM KCl 0.1 mM EDTA 0.1 mM EGTA added fresh to the buffer: 1 mM PMSF 1 mM DTT 1% (v/v) protease inhibitor cocktail (7x Complete, mini) 1% (v/v) phosphatase inhibitor cocktail (10x Phosstop) Dest. H_2O
Schreiber buffer C (hypertonic buffer) for 3T3-L1 and SGBS cell lines	
	20 mM HEPES, pH 8 0.4 M NaCl

	0.1 mM EDTA 0.1 mM EGTA 20% (v/v) glycerol added fresh to the buffer: 1 mM PMSF 1 mM DTT 1% (v/v) protease inhibitor cocktail (7x Complete, mini) 1% (v/v) phosphatase inhibitor cocktail (10x Phosstop) Dest. H_2O
Transformation	
Agra plates	
	1.5% (w/v) Agar-agar in LB medium 100 µg/ml ampicillin
LB media	LB media 1.0% (w/v) peptone 1.0% (w/v) yeast extract 85.6 mM NaCl
Tris/EDTA (TE) buffer 10x	
	100 mM Tris-HCl, pH 8.0 10 mM EDTA Dest. H_2O
WB	
Blocking buffer: 5 % Milk or 2% ECL-blocking buffer	
	5% (w/v) Milk powder in TBS-T 5% (w/v) ECL-blocking reagent in TBS-T
Laemmli buffer 5x	
	310 mM Tris pH 6.8 173.4 mM SDS 25 % (v/v) Glycerol 2.5 mM EDTA 2 mM DTT
Running buffer 10x	
	247.6 mM Tris 2 M Glycine 0.1 % (v/v) SDS Dest. H_2O, final pH ~8.3
Semidry blotting buffer	
	25mM Tris 192 mM Glycin 20% (v/v) Methanol Dest. H_2O, final pH ~8.4
Tris/Buffer/Saline (TBS) 10x	
	200 mM Tris 1.37 M NaCl Dest. H_2O, final pH ~7.5
Tris/Buffer/Saline/Tween (TBS-T) 10x	
	1x TBS 0.1% Tween20 Dest. H_2O

Table 6. Buffers, solutions and agar plates used in this study.

5.1.9 Chemicals, reagents and cell culture media

Name	Source
2-Mercaptoethanol	Merck, Darmstadt, Germany
Acrylamide solution 40 % (37:5:1)	Carl Roth, Karlsruhe, Germany
Agar-Agar	Sigma-Aldrich, Steinheim, Germany
APS	Merck, Darmstadt, Germany
Biotin	Roth, Karlsruhe, Germany
Boric Acid	Carl Roth, Karlsruhe, Germany
Bromphenol Blue sodium salt	Sigma-Aldrich, Steinheim, Germany
Calcium pantothenate	Roth, Karlsruhe, Germany
Chaps	Omnilap, Bremen, Germany
COmplete, Mini, EDTA-Free	Roche Diagnostics, Risch, Switzerland
Dexamethason	Sigma-Aldrich, Steinheim, Germany
DMEM (high glucose, L-glutamine)	GibcoTM, Invitrogen, Karlsruhe, Germany
DMEM:F12	GibcoTM, Invitrogen, Karlsruhe, Germany
DMSO	Roth, Karlsruhe, Germany
DTT	Applichem, Darmstadt, Germany
ECL Prime Blocking Reagent	GE Healthcare, NJ, USA
EDTA	Carl Roth, Karlsruhe, Germany
EGTA	Merck, Darmstadt, Germany
Ethanol	J.T. Baker, Deventer, Holland
FCS Gold	PAA Laboratories, Pasching, Austria
Ficoll®PM 400	Sigma, Steinheim, Germany
Gentamicin (10 mg/ml)	Roth, Karlsruhe, Germany
Glucose	Sigma-Aldrich, Steinheim, Germany
Glycine	Applichem, Gatersleben, Germany
Glycerol	Merck, Darmstadt, Germany
HEPES	Carl Roth, Karlsruhe, Germany
IBMX	Serva, Heidelberg, Germany
Insulin solution human (10 mg/ml)	Sigma-Aldrich, Steinheim, Germany
KCL	Merck, Darmstadt, Germany
KH_2PO_4	Merck, Darmstadt, Germany
L-Glutamine	Sigma-Aldrich, Steinheim, Germany
Lb-Agar	Applichem, Gatersleben, Germany
LB-Medium powder	Applichem, Gatersleben, Germany
Lipofectamine 2000	GibcoTM, Invitrogen, Karlsruhe, Germany
Magnesium chloride	Sigma-Aldrich, Steinheim, Germany
Magnetic beads (M-280)	Invitrogen, Karlsruhe, Germany
Methanol	Sigma-Aldrich, Steinheim, Germany
Milk powder	Carl Roth, Karlsruhe, Germany
NaCl	Roth, Karlsruhe, Germany
Na_2HPO_4	Carl Roth, Karlsruhe, Germany
$NaHCO_3$	Merck, Darmstadt, Germany
NaOH	J.T. Baker, Deventer, Holland
Nonidet® P40	Sigma-Aldrich, Steinheim, Germany
NaF	Sigma-Aldrich, Steinheim, Germany
Nonidet® P 40 Substitute solution	Sigma-Aldrich, Steinheim, Germany
Oil-red-O	Sigma-Aldrich, Steinheim, Germany
ON-TARGETplus Non-targeting Control Pool	Dharmacon, CO, USA
ON-TARGETplus Human RYBP (23429) siRNA - SMARTpool	Dharmacon, CO, USA
ON-TARGETplus Human YY1 (7528) siRNA - SMARTpool	Dharmacon, CO, USA
OPTI-MEM®I	GibcoTM Invitrogen, Karlsruhe, Germany
Orange G	Roth, Karlsruhe, Germany
PBS solution	Biochrom AG, Berlin, Germany
Phosphatase inhibitor cocktail	Roche Diagnostics, Risch, Switzerland
PMSF	Sigma-Aldrich, Steinheim, Germany

PBS Dulbecco	Biochrom AG, Berlin, Germany
Poly[d(l-C)]	Roche Diagnostics, Risch, Switzerland
Precision Plus Protein Dual Color Standards	Bio-rad, CA, USA
Protein Standard (BSA for Bradford assay)	Sigma-Aldrich, Steinheim, Germany
Penicillin-streptomycin (P/S)	PAA Laboratories, Pasching, Austria
RNase Zap	Ambion, Woodward-Austin, USA
Roti-Quant solution	Carl Roth, Karlsruhe, Germany
RPMI 1640	GibcoTM, Invitrogen, Karlsruhe, Germany
Sodium dodecyl sulphate (SDS)	Roth, Karlsruhe, Germany
TEMED	Merck, Darmstadt, Germany
Tris base	Applichem, Gatersleben, Germany
Triton X-100	Sigma-Aldrich, Steinheim, Germany
TRIZOL®Reagent	Invitrogen, Carlsbad, CA, USA
Typsin-EDTA(10x)	PAA Laboratories, Pasching, Austria
Trypan Blue Solution (0.4 %)	Sigma-Aldrich, Steinheim, Germany
Tween® 20	Sigma-Aldrich, Steinheim, Germany
Nuclase free distilled Water	Carl Roth, Karlsruhe, Germany

Table 7. Chemicals and reagents used in this study.

5.1.10 Kits

Name	Description	Source
Absolute SYBR Green ROX Mix	Real time quantitative PCR	ABgene, Hamburg, Germany
Dual-Luciferase® Reporter Assay	Lificerase gene assay	Promega, Madison, USA
High-Capacity cDNA Reverse Transcription Kit	Reverse transcription PCR	Applied Biosystems, Weiterstadt, Germany
NE-PER Nuclear and Cytoplasmic Extraction Reagents	Nuclear extracts preparation	Pierce, IL, USA
NucleoSpin RNA II (RNA isolation)	RNA isolation	Macherey&Nagel, Düren, Germany
Pure Yield^{TM} Plasmid MaixPrep	DNA isolation	Promega, Madison, USA

Table 8. Kits used in this study.

5.1.11 Consumables

Name	Source
Bottle top filter (0.2 µm)	Corning, NY, USA
Cell culture flasks (25 cm^2)	Greiner bio-one, Frickenhausen, Germany
Cell scrapers	Greiner bio-one, Frickenhausen, Germany
	Techno Plastic Products (TPP), Trasadingen, Switzerland
Centrifugation tubes (15, 50 ml)	Greiner bio-one, Frickenhausen, Germany
Cryogenic vials (2 ml)	Greiner bio-one, Frickenhausen, Germany
Eppendorf tubes	Eppendorf, Hamburg, Germany
Filter tips	Kisker, Steinfurt, Germany
Gloves	Meditrade, Kiefersfelden, Germany
Laboratory glass bottles	Duran, Mainz, Germany
Multiply-µStrip 0.2 ml chains	Sarstedt, Nümbrecht, Germany
PCR plate 96 well	Eppendorf, Hamburg, Germany
PCR plate sealer	Schubert&Weiss, Munich, Germany

PCR tubes (0.5 ml)	Biozym, Oldendorf, Germany
pH indikatorpapier	Merck, Darmstadt, Germany
Pipetus	Brand, Wertheim, Germany
Pipette tips	Biozym, Oldendorf, Germany
	Corning, NY, USA
	Gilson, Bad Camberg, Germany
	Kisker, Steinfurt, Germany
PS-tubes (5 ml)	Greiner bio-one, Frickenhausen, Germany
Solid cotton buds	Zefa Laborservice, Harthausen, Germany
Surgical disposable scalpels	B. Braun Melsungen AG, Melsungen, Germany
Syringes (2 ml, 5 ml, 10 ml)	Bection Dickison GmbH, Germany
Tissue culture dish (150 x 25 mm)	Falcon, BD Bioscience, NJ, USA
	Sarstedt, Nümbrecht, Germany
Tissue culture flasks (175 cm^2)	Falcon, BD Bioscience, NJ, USA
Tissue culture flasks (75 cm^2)	Falcon, BD Bioscience, NJ, USA
	Techno Plastic Products (TPP), Trasadingen, Switzerland
Tissue culture flasks TPP (25 cm^2)	Falcon, BD Bioscience, NJ, USA
	Techno Plastic Products (TPP), Trasadingen, Switzerland
Tissue culture plates (6, 12, 24, 48 well)	Corning, NY, USA
	Techno Plastic Products (TPP), Trasadingen, Switzerland
Tissue culture stripettes (1, 2 ml)	Corning, NY, USA
Tissue culture stripettes (5,10,25,50 ml)	Greiner bio-one, Frickenhausen, Germany
Whatman Gel Blotting Paper	GE Healthcare, Chalfont, UK

Table 9. Consumables used in this study.

5.1.12 Laboratory instruments

Name	Source
Bacteria Shaker	B.Braun Biotech, Melsungen, Germany
Berthold detection system	TITERTEK BERTHOLD, Eilat, Israel
Blot-Apparatus (Semidry blotting)	Biometra, Göttingen, Germany
Cell culture bench (HeraSafe)	Heraus, Hanau, Germany
Cryo Freezing Container	Nalgene, NY, USA
Magna-Sep™ Magnetic Particle Separator	Invitrogen, Karlsruhe, Germany
MasterCycler Gradient	Eppendorf, Hamburg, Germany
MasterCycler realplex gradients	Eppendorf, Hamburg, Germany
Multi-dispender pipette	Brand, Wertheim, Germany
Incubator	Binder, NY, USA
EMSA gel chamber	Biometra, Göttingen, Germany
Improved Neubauer chamber	Brand, Wertheim, Germany
Centrifuges 5415/5810 R/ 5417 R	Eppendorf, Hamburg, Germany
Potter homogenizer	Sartorius, Göttingen, Germany
Low Voltage Power Supplies	Biometra, Göttingen, Germany
Laminar flow hood	Köttermann, Uetze-Hänigsen, Germany
Microplate reader	TECAN, Männedorf, Switzerland
NanoQuant Plate™	TECAN, Männedorf, Switzerland

Photo Leica DC 300F	Leica, Bensheim, Germany
Overhead shaker	Heidolph Instruments, Schwabach, Germany
pH meter	Mettler-Toledo, Inc., OH, USA
Pipetting aid	Gilson, Bad Camberg, Germany
	Eppendorf, Hamburg, Germany
Reax Control	Heidolph Instruments, Schwabach, Germany
Rocking Platform	VWR International, PA, USA
Vortex mixer	Scientific industries, NY, USA
Rollers synchronous mixer RM5-30V	CAT Ing., Staufen, Germany
Rotor Sorvall SLA-1500	Kendro, Asheville NC, USA
Thermo haake K20 (circulator pump)	Thermo haake, NH, USA
Thermo leader	Uniequip, Planegg, Germany
Thermomixer comfort	Eppendorf, Hamburg, Germany
Titramax 100	Heidolph, Schwabach, Germany
Typhoon TRIO+	GE Healthcare, NY, USA
UV Lamp	Vilber Lourmat, Marne-la-Vallée, France
Water bath	Julabo, Seelbach, Germany
Weighting scale	Denver Instrument, Göttingen, Germany

Table 10. Laboratory instruments used in this study.

5.1.13 Software

Name	Description	Source
Adobe Illustrator CS6	Image processing	Adobe, CA, USA
Adobe Photoshop CS	Image processing	Adobe, CA, USA
Citavi 3.4	Reference management	Swiss Academic Software GmbH, Zürich, Switzerland
Excel	Data analysis	Microsoft Corporation, Redmond, USA
GraphPad Software	Statistical analysis, graphics	GraphPad Software, CA, USA
IBM SPSS Statistics 20	Statistical analysis, graphics	IBM Corporation, NY, USA
i-control™	Microplate Reader	Tecan, Männedorf, Switzerland
Image J	Image processing	Wane Raspand, Bethesda, USA
MS word	Word processing	Microsoft Corporation, Redmond, USA
Odyssey V3.0	Infrared Image reader	LI-COR Biosciences, Bad Homburg, Germany
Powerpoint	Image processing	Microsoft Corporation, Redmond, USA
Scan control software	Gel plate reader of Typhoon Trio™	GE Healthcare, CT, USA

Table 11. Software used in this study.

5.2 Experimental methods

5.2.1 Cell culture

HIB 1B cell culture and differentiation

Hibernoma 1B (HIB 1B) cell line is derived from brown fat tumor of a transgenic mouse which functions as an energy-dissipating tissue [335]. HIB 1B cells were cultured in proliferation medium. For differentiation, HIB 1B cells were cultured to confluence up to 100 % and then exposed to the differentiation medium. Cell medium was changed every 2 days. After 9 days (day 9) cells were harvested for further analysis.

Media	Composition
Basal medium	
	1.2 % (w/o) DMEM:F12 powder
	0.014 mM $NaHCO_3$
	0.016 mM biotin
	0.004 mM calcium pantothenate
	0.015 mM glucose
	13.5 mM HEPES pH 7.4
	Dest. H_2O, final pH ~7.3
	sterile filtered with bottle top filter (0.2 μm)
Proliferation medium	
	Basal medium
	10 % FCS G
	10 μg/ml gentamicin
Differentiation medium	
	Basal medium
	7 % FCS G
	10 μg/ml gentamicin
	17 nM insulin

Table 12. HIB 1B cell culture media.

3T3-L1 cell culture and differentiation

3T3-L1 white pre-adipocyte, derived from Swiss albino mouse embryo tissue, is the best established cell line model for studying adipogenesis *in vitro* [336]. 3T3-L1 cells were cultured in proliferation medium. For differentiation, 3T3-L1 cells were seeded in a 6-well plate and grown. After reaching ~80% confluence, the medium was then changed to induction of adipocyte differentiation with induction medium (day 0). On day 3, medium was replaced by differentiation medium (day 3). After day 5, medium was changed with differentiation medium every 2 days until adipocyte differentiation. After two weeks (day 14) cells were stained with Oil-red-O or harvested for further analysis.

Media	Composition
Proliferation medium	
	DMEM
	10% FCS G
	penicillin (100 units/ml) and streptomycin (100 μg/ml)
Induction medium	
	Proliferation medium
	861 nM insulin
	250 mM dexamethasone
	0.5 mM IBMX
Differentiation mdeium	
	proliferation medium
	861 nM insulin

Table 13. 3T3-L1 cell culture media.

SGBS cell culture and differentiation

Human Simpson-Golabi-Behmel Syndrome (SGBS) pre-adipocytes is obtained from an adipose tissue specimen of a diseased patient with Simpson-Golabi-Behmel syndrome. The SGBS cell line is characterized by a high capacity for adipogenic differentiation. When the SGBS cells are once differentiated, they function as primary isolated human fat cells [337]. SGBS pre-adipocytes were grown in proliferation medium until desired density. For induction of adipocyte differentiation SGBS cells were cultured in serum free induction medium (day 0). 3 days after induction cell medium was replaced by differentiation medium (day 3). After day 5, medium was replaced by differentiation medium every three days until adipocyte differentiation. After two weeks (day 14) cells were stained with Oil-red-O or harvested for further analysis.

Media	Composition
Basal medium a	
	D-MEM:F 12 (1:1)
	33 μM pantothenic acid
	17 μM Biotin
Basal medium b	
	MCDB-131
Proliferation medium	
	Basal medium a
	10% FCS F
	penicillin (100 units/ml) and streptomycin (100 μg/ml)
Induction medium	
	Feeding medium
	2 μM rosiglitazone
	25 mM dexamethasone
	0.5 mM IBMX
Feeding medium	
	2/3 Basal medium a and 1/3 basal medium b
	10 μg/ml human transferrin
	66 nM insulin
	100 nM cortisol
	1 nM triiodothyronine

Table 14. SGBS cell culture media.

Human primary adipocytes isolation, cell culture and differentiation

Primary human adipocyte progenitor cells were obtained by lipoaspiration or surgical excision of subcutaneous adipose tissue, and were isolated and cultured as previously described

[98,338,339]. Briefly, after expansion and freezing, the cells were cultured in 6-well plates in 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 desired 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) and harvested for further analysis.

C2C12 cell culture and differntiation

Mouse myoblast C2C12 cell line was cultured in DMEM medium supplemented with penicillin (100 units/ml) and streptomycin (100 µg/ml) and 10 % FBS. For differentiation of myoblasts, C2C12 myoblasts were cultured in DMEM medium containing 10% horse serum to induce differentiation for 7 days (day 7).

293T, Huh7 and INS-1 cell culture

Human embryonic kidney 293T cell line and human hepatoma Huh7 cell line were cultured in DMEM medium supplemented with penicillin (100 units/ml) and streptomycin (100 µg/ml) and 10 % FBS. Rat insulinoma cell line INS-1 was cultured in RPMI medium supplemented with penicillin (100 units/ml) and streptomycin (100 µg/ml), 10 % FBS, 100 mM sodium pyruvate, and 50 µM 2-mercaptoethanol.

5.2.2 Cell culture: general information

All the above cell lines were maintained at 37°C in a thermostatically-controlled incubator containing 5% CO_2 in humidified environment. The each cell line was cultured in the above described medium (5.2.1) before splitting. For adherent cell lines when the cells reached the desired density, the cell medium was removed and cell layers were washed carefully with 5 ml PBS two times and incubated with 1 ml trypsin for 3-6 min (depends on cell type). Trypsin reaction was stopped by adding 9 ml fresh medium. After passaging cell proliferation rates were measured by direct cell counting using a Neubauer Hermocytometer with trypan blue under light microscopy. The stock cell suspension was then diluted with corresponding

medium to each cell line were seeded at a desired ratio in a new flask or cell plates followed by incubation o/n at 37 °C. For the suspension cell line, there is no need to trypsinize because the cells are already suspended in growth medium. Instead, the cells were maintained by directly diluting with the fresh medium every 2 to 3 days at a desired ratio in a new flask or cell plates followed by incubation o/n at 37 °C. For long term storage, cells were pelleted by centrifugation at 1500 rpm for 5 min. Then, the supernatant was resupended in the corresponding proliferation medium additionally supplemented with 10% DMSO and transferred in 2 ml polypropylene cryogenic vial (10^6 cells per tubes). The vials were placed into a 4°C pre-chilled isopropanol box and kept at -80°C freezer for 3 days and then transferred to liquid nitrogen.

5.2.3. Oil red O staining

Oil red O is used to stain cytoplasmic lipids, triglycerides, and some lipoproteins, particularly in culture or in tissues. It is commonly used to evaluate adipogenesis *in vitro* [247]. The differentiated adipocytes were washed once with PBS and fixed with 3.7 % formaldehyde for 1 h. Then, formaldehyde was removed and replaced by working solution of Oil Red O and incubated for 1 h. The plated was then washed once with PBS and observed under light microscopy. Subsequently, isopropyl alcohol was added to the stain culture dish. The Oil Red O solution was discarded and the stained cells were washed with PBS followed by visualizing under light microscope. Additionally, the Oil Red O in triglyceride droplets was extracted with 100% isopropanol and the absorbance at OD 510 nm was determined with a spectrophotometer. The whole process was performed at RT.

5.2.4. Transformation

Transformation is often used for non-viral DNA transfer in non-animal eukaryotic cells such as bacteria. In this study transformation was used to produce a large scale of recombinant DNA. 15 µl of the competent *E.coli* strain DH5α were thawed slowly on ice and mixed with 1 µg plasmid DNA followed by incubation on ice for 5 min. The bacterial cells were then shock heated for 90 sec. at 42 °C and immediately replaced on ice for further 5 min. 500 µl of the LB medium was added to the reaction mixture and incubated for 30 min at 37 °C in a bacterial shaker (200 rpm). Finally, 500 ml of the LB-medium containing 100 ug/ml ampicillin (for selection of transformed cells) was added to the bacteria cells followed by incubation o/n (at least for 16 h) at 37 °C in the bacterial shaker (200 rpm). On next day,

plasma DNA isolation was performed using the Promega PureYield^{TM} Plasmid Maxiprep System (Promega) according the manufacter´s instruction manual. Plasmid DNA was eluated with nuclease free H_2O and stored at -20 °C. Concentration of Plasmid DNA was determined using NanoQuant Plate™ reader (Tecan).

5.2.5 Transfection of eukaryotic cell lines

Transient transfection allowed expression of specific proteins in eukaryotic cells without the integration of foreign gene into the genome. All plasmids used for transfection were kindly provided by M.Kern (USA) and M. Klar (Germany) purified using the PureYield™ Plasmid Maxiprep System (Promega) according to the manufacturer´s instruction. The PRRX1 and PRRX1-flag expression constructs were derived from pcDNA 3.1 and pCMV, respectively. One day before the transfection, 293T cells were seeded in 10 cm^2 plate to achieve 80-90 % confluence on day of transfection. 2 hours before transfection the medium was replaced with opti-MEM® (Invitrogen) (containing no serum). In the mean time for each transfection sample, DNA-Lipofectamine 2000 complexes were prepared as follows: desired amount of DNA tranfected was diluted in opti-MEM® (Invitrogen), in parallel Lipofectamine 2000 (Invitrogen) was mixed with opti-MEM® and incubated at RT for 5 min. The diluted Lipofectamine was combined with the diluted DNA and the mixture (total medium=660µl) incubated for 20 min at RT. After 20 min, the DNA-Lipofectamine complex was added to the cells incubated at 37°C o/n. In 3-4 hours, the transfection medium was removed and the cells were supplemented with 2 ml fresh medium. 24 h after transfection cells were harvested for further analysis.

5.2.6 Knockdown using siRNA

siRNA transfection was used to silence expression of specific genes in eukaryotic cells. Transfections were performed with Hiperfect Reagent (Invitrogen). SGBS cells were cultured in 6-well plates as described above (see 5.2.1) and transfected with 25 nM siRNA targeting YY1, RYBP or non-targeting (NT) control siRNA (ON-TARGETplus human siRNA SMARTpool, Dharmacon, USA) using HiPerFect (Qiagen, Germany) according to the manufacturer's instructions (Knock-down efficiency was 30-60%). 72h after transfection confluent cells were harvested using the RNeasy-Minikit (Qiagen) to extract total RNA. The high capacity cDNA Reverse Transcription kit (Applied Biosystems, Germany) was used for

transcription of 1µg total RNA into cDNA. qPCR analysis of the human PPARG1 and PPARG2 isoform transcripts[98], GAPDH as housekeeping gene, YY1 and RYBP to control for knockdown efficiency, was performed using a qPCR SYBR-Green ROX Mix (ABgene, Germany) and 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 and by agarose gel electrophoresis to assess the size of PCR products (primers are shown in Table 1a). Mean target mRNA level was calculated by the ΔΔCT method relative to the level of the GAPDH gene expression level based on technical duplicates. siRNA transfection experiments were performed five to eight times and *P*-values of qRT-PCR were calculated using one-sample *t*-test.

5.2.7 Luciferase reporter gene assay

Luciferase reporter gene assay is the common method to measure transcriptional activity mediated by inserted gene including promoter or enhancer region in eukaryotic cells. To assess transcriptional activity mediated by SNP-adjacent regions, luciferase reporter gene assay was performed as described previously [98] with some modifications. C2C12 cells in 48 wells at approximately 80% confluence were differentiated for 7 days as described above. Differentiated C2C12 cells and 293T cells, INS-1 cells, Huh7 cells and undifferentiated C2C12 cells at 80-90 % confluence were transfected in 48-well plates by Lipofectamine 2000 transfection reagent (Invitrogen). 293T cells, Huh7 cells and C2C12 (undifferentiated and differentiated) cells were transfected with 0.3 µg of the respective firefly luciferase reporter vector, 0.04 µg of the ubiquitin promoter vector and 1 µl differentiated Lipofectamine reagent. INS-1 cells were transfected with 1.2 µg of the respective firefly luciferase reporter vector, 0.16 µg of the ubiquitin promoter vector and 2 µl differentiated Lipofectamine reagent. 3-4 h after transfection the medium was replaced by fresh medium followed by incubation at 37°C. 24 hours after transfection, the cells were washed one time with PBS and lysed in 1x passive lysis buffer (Promega) on rocking for 20 min at RT. Firefly and renilla luciferase activity were measured using luminometer (Berthold), respectively. The ratios of firefly luciferase expression to renilla luciferase expression were calculated and normalized to the TK promoter control vector. All experiments were performed at least 5-7 times.

5.2.8 RNA preparation

Cells were seeded in 6-well plate until to achieve the desired confluence. The medium was removed and the cells were washed twice with 1 ml pre-chilled PBS. Cell cultured were lysed with 1 ml TRIzol reagent (Invitrogen) per well by scrapping on ice. Lysates were transferred to a sterile 1.5 ml RNAse free reaction tube. RNA isolation was performed using the NucleoSpin Kit (Macherey-Nagel) according the manufacter´s instruction manual. To prevent the degradation by RNase, all workplace were cleaned with RNase Zap from Ambion. For determination of RNA concentrations absorptions at 260 nm and 280 nm of samples were determined using NanoQuant Plate™ reader (Tecan). Optimal RNA preparations are characterized by an OD 260 /280 of 2.0. Isolated RNA was stored at -20 °C for further analysis.

5.2.9 PCR analysis

5.2.9.1 Reverse Transcription PCR

The synthesis of cDNA was performed using cDNA Reverse Transcription kit (Applied Biosystems) according the manufacturer´s instruction manual. For cDNA synthesis, 1 µg RNA was reverse transcribed into cDNA. For each reaction, 2x RT buffer, 2x RT random primer, 80mM of dNTP mix and 50U of MultiScribe™ reverse transcriptase were added to RNA in a total volume of 10 μl. Reaction was carried out in a thermocycler (Eppendorf). Conditions for reverse transcription PCR was performed see below (Table 12). The generated cDNA was then used as template for quantitative real time PCR reactions.

Step	Temperature (°C)	Time	Description
1	25	10 min	Priming
2	37	120 min	Transcription
3	85	5 min	Enzyme inactivation
4	4	ever	Storage
5	end		

Table 15. Reverse transcription – PCR condition.

5.2.9.2 Real time quantitative RCR (qPCR)

To determine the expression levels of genes, qPCR was performed. qPCR analysis of the genes *PPARG1*, *PPARG2*, *Leptin* and *HPRT* (Table 1b) was performed using a qPCR SYBR-Green ROX Mix (ABgene, Germany) and using the MasterCycler Realplex system (Eppendorf, Germany). For each reaction, 1x Absolute QPCR SYBR Green ROX MIX buffer, 70mM forward primer, 70 mM reverse primer were added to cDNA in a total volume of 20 μl. The samples were analyzed as triplicates. qPCR was performed under conditions (Table 16). The mRNA levels of target genes were normalized to those of hypoxanthin phosphoribosyltransferase (HPRT) using the *ΔΔCT* method. A melting curve profiles in which temperature is decreased from 95°C to 68°C with simultaneous measurement of fluorescence allowed verification of the amplification specificity of transcripts at the end of each PCR. All process was performed according qPCR SYBR-Green ROX Mix Kit ´s protocol.

Step	Temperature (°C)	Time	Description	Cycle
1	95	15 min	Activation of enzyme	
2	95	15 sec	Denaturation	
3	60	30 sec	Annealing	40 cycels
4	72	30 sec	Extension	

Table 16. Real time quantitative PCR condition.

5.2.10 Preparation of whole extracts

After overexpression of PRRX1 expression vector in 293T cells, the transfected cells were harvested as whole extracts. Medium was removed and the cells were washed twice with pre-chilled 1ml PBS. Cells were collected by scraping in the treatment with lysis buffer (20mM Tris-HCL (pH 8.0), 0.4 M NaCl, 0.1 mM EDTA, 0.1 mM EGTA, 1 mM PMSF, 1 mM DTT, 20 % v/v glycerol). The solution was incubated for 20 min followed by three times freezing and thawing in liquid nitrogen. Finally, the solution was centrifuged at 14000 rpm at 4 °C for 10 min. The supernatant was transferred in a new tube and the protein concentration was determined by the Bradford assay and stored at -80 °C for further analysis.

5.2.11 Preparation of nuclear extracts

Nuclear protein extracts from 3T3-L1 cells, SGBS cells, primary pre-adipocytes, 293T cells, C2C12 cells and INS-1 cells were prepared as described previously [340]. Cells were washed two times with PBS and scraped into pre-chilled Schreiber buffer A followed by incubation for 25 min on ice. 0.6 % Nonidet® P40 were added and two times vigorously mixed to the

suspension for destruction of cell membrane. Cells were collected by centrifugation at 14000 rpm for 3 min. After centrifugation, the pellet of nuclei was washed twice with buffer A to completely remove Nonidet® P40 and centrifuged at 14000 rpm for 3 min. speed. In case of 3T3-L1 cells, SGBS cells, the remaining fat residue were totally removed using cotton swabs. Then pre-chilled hypertonic Schreiber buffer C was added to the pellet (3 times of pellet volume) and incubated at 4 °C for 20 min by shaking vigorously on rotating wheel. Finally, the supernatant was recovered by centrifugation at 14000 rpm for 5 min and used for further analysis. HIB 1B nuclear extracts were prepared by high salt extraction according to standard procedures [341]. Cultured HIB 1B cells were washed two times with PBS and gently scraped off in pre-chilled homogenization buffer. The cells were collected by centrifugation at 3300 g for 15 min and the supernatant was discarded. The pellet was then resuspended in low salt buffer. Subsequently, high salt buffer containing 1.2 M KCl was added to the suspension to swell the cells. Swollen cells were broken at 4°C for 30 min by vigorous shaking on rotating wheel and the supernatant was recovered by centrifugation at 20817g for 30 min to pellet the nuclei and used for further analysis. Nuclear protein extracts from adult mouse whole brain were prepared using NE-PER Reagent (Pierce) according the manufacturer´s instruction manual.

5.2.12 Protein concentration by Bradford Assay

Bradford assay was used as a common method for determining protein concentration. Bovine serum albumin (BSA) as a standard was used at concentrations of 0.0625, 0.125, 0.25, 0.5, 1 and 2 mg/ml. BSA was diluted in 10 % of respective final buffer for each extract. 5 µl of standards, blanks and samples (diluted 1:10) were measured in plates in a 1:10 dilution with dest. H_2O. Roti-Quant solution (Carl Roth) was mixed with dest. H2O (in a ratio of 1:2.8, respectively). 45 µl of mixture was added per well and incubated for 5 min at RT under dark condition. All samples were duplicated. Measurement was performed at OD 595 using a photometer (Tecan).

5.2.13 Silver staining

The SDS-PAGE gel was first fixed by incubating in fixer solution (50:5:45 v/v/v methanol: acetic acid: H_2O) for 20-30 min. After fixing, the gel was washed several times with H_2O for 10 min. and incubated in H_2O further for 1 h on a shaking form. The gel was then sensitized

with 0.02% sodium thiosulphate for 1-2 min. Following sensitization, the gel was washed two times with H_2O for 20 s (10 s per each) and then stained with pre-cooled staining solution (0.1% $AgNO_3$) for 30 min at 4°C. Then, the gel was washed two times with H2O for 30 s. The gel was developed with developing solution (0.04% formaldehyde in 2% sodium carbonate). The development was stopped with stop solution as soon as it turned yellow and the bands were clearly visible. The gel was then washed with 1% ethanol and stored n the same solution.

5.2.14 Western Blotting

Western blot is a widely used analytical method for determining specific protein levels. The proteins are separated on SDS-PAGE gel and transferred to a membrane PVDF or nitrocellulose. Detection of interested proteins is performed using antibodies anti target proteins.

5.2.14.1 Sodium dodecyl sulfate- polyacrylamide gel electrophoresis (SDS-PAGE)

For applying sample into the gel 20 µg of protein extracts was mixed with 1x Laemmli buffer solution and dest. H_2O was added to perform 20 µl of total volume. The mixture was incubated at 95°C for 5 min on heating block. The samples were subjected into SDS-PAGE on stacking gel (5% acrylamide gel) and resolving gel (10% acrylamide gel). Electrophoresis was performed in a vertical electrophoresis chamber using 1x running buffer. 4 µl of Dual Marker (Bio-rad) was run in parallel as a protein weights reference. Gel was run to the stacking gel at 100 V until the loading dye reached the end of the stacking gel, and then at 120 until the loading dye is approximately 1 cm from the bottom of the resolving gel.

5.2.14.2 Electroblotting

The separated proteins were transferred out of the gel and electroblotted onto the nitrocellulose membrane using semi-dry blotting system (Bio-rad). After electrophoresis gels are freed from the frame and rinsed with ddH_2O. The gel was placed over the membrane overlayed with Whatman paper soaked in Semidry blotting buffer. Air bubbles were carefully removed with roller because bubbles could disrupt the transfer of proteins onto the membrane. Proteins were dry transferred to a nitrocellulose membrane at low voltage (25 V) for 45 min

pro gel. After blotting the membrane was washed for 3 times for each 5 min with PBS buffer and either directly used or stored rapped at 4°C.

5.2.14.3 Immunodetection

For detecting specific proteins non-specific proteins were first blocked with ECL blocking buffer (Amersham) or 2% Milk in PBS-T for 1 h at RT (blocking buffer and time depends on specific proteins). The membrane was incubated with affinity purified rabbit polyclonal antibodies to PRRX1 at 4 °C o/n or H3 at RT for 1h on a horizontal shaker. To control and correct for loading error, an internal control was always used. In this study H3 (about H3) was used as internal control for nuclear proteins in the Western blot analysis. Antibody to PRRX1 (Dr. Kern, Norway) was used at 1:5000 dilution, antibody to α-tubulin (Santa cruz, 2027) was used at 1:1000 dilution. After 1 h of blocking in a PBS-T buffer containing 2% (w/v) ECL powder (GE Healthcare), the membrane was washed three times with PBS-T buffer. Afterwards, the membrane was incubated in fresh PBS-T buffer with IRDye 680, 800 labeled anti-rabbit IgG secondary antibodies with a dilution of 1: 20000 (LI-COR Biossciences) for 1 h at RT, and immune complexes quantified (scanned) using the Odyssey Infrared Imaging system (LICOR Biosciences).

5.2.15 Preparation of oligonucleotides

Complementary pairs of single stranded oligonucleotides were annealed by heating at 80 - 90°C for 5 min in TE buffer. The appreciated temperature depends on the length of DNA sequence that was generally 10°C above melting temperature. After annealing the oligonucleotides were slowly cooled down o/n. For effective purification of double stranded oligonucleotides, the reaction was shifted on a 12% polyacrylamide gel (0.5x TBE , 12 % 37.5:1 acrylamide/biacrylamide (40 %), 2.7 % v/v glycerol, 0.075% APS, 0.05 % TEMED; 5.3% polyacrylamide gel is also available) to remove remaining single stranded oligonucleotides. The gel was lay on the thin layer chromatography plate covered with fluoresce and was lightened with UV light to visualized the DNA. The gel slice with double stranded oligonucleotides were cut out and extracted in TE buffer by incubation o/n in shaking at 37 °C. Finally, the reaction was transferred to the filter and the gel slice was removed by centrifugation at 13200 rpm for 1 min. The DNA concentration was determined using TECAN microplate reader on NanoQuant Plate™ at a wave length of 260nm. Samples

were measured undiluted (1.5 µl) with TE as a blank. The concentration of oligonucleotides was measured duplicated. The purification of oligonucleotides was examined by OD 260 to OD 280 ratio. The ratio of ~1.8 is ideal for DNA in high purity; a ratio < 1.8 indicate presence of contaminants such as RNA. Purified oligonucleotide samples were stored at -20 °C for long storage.

5.2.16 EMSA

The Electrophoretic Mobility Shift Assay (EMSA) is a type of affinity electrophoresis and gives information about protein-DNA interactions. In this study, EMSA techqunics were used as a large part to observe allele-specific binding of proteins. Non-radioactive EMSA was performed with Cy5-labelled oligonucleotide as described previously [98] with some modifications. Initial EMSA was performed with 5 µg nuclear extracts, 0.35 µg poly (dI-dC), 1.5 µl of 5x GBB and H_2O to the total volume of 9 µl for each reaction. To optimize DNA-protein interactions, conditions such as amount of proteins (2-10 µg), concentration of poly (dI-dC) (0.1-1 µg) as a nonspecific competitor, and salt concentration in the binding buffer were varied. Optimizing EMSA binding conditions is a crucial step as the reaction conditions for affinity chromatography are in general, similar to those used for EMSA. The reaction was performed for 10 min on ice. Then, 1 µl (conc. 1 ng/µl) of Cy5-labelled oligonucleotides was added to each sample and incubated for 20 min on ice. After incubation, 1 µl 10x loading buffer was added. In the meantime, gel was pre-runned at 200 V for 1-1.5 hours in 0.5x TBE buffer. Samples were applied onto the non-denaturing 5.3 % polyacrylamide gel (0.5x TBE , 12 % 37.5:1 acrylamide/biacrylamide (40 %), 2.7 % glycerol, 0.075% APS, 0.05 % TEMED) and the gel was run at 200 V. for ca. 3 hours at 4 °C until 10x loading buffer is ~ 4/5 from the top of the gel. Finally, the gel was removed from the frame and scanned with a Typhoon TRIO+ fluorescence scanner at a wavelength of 650 nm with high sensitivity and a resolution of 100 microns. Quantification of band signal intensity was performed using software Image J. In competition EMSA, 33-fold molar excess of unlabelled specific probe was included as competitor to the reaction prior to addition of Cy5-labeled DNA probes to compete DNA-binding between Cy5-labelled and unlabelled oligonucleotides. The reactions were incubated for 20 min at 4°C. In supershift experiments, nuclear extracts were pre-incubated with 0.8 µg of antibody anti-YY1 (sc-281, Santa Cruz Biotechnology) or IgG (sc-2027, Santa Cruz Biotechnology) were added to the reaction mixture and incubated for 30 min at 4 °C, respectively. All EMSA experiments were replicated at least three times.

5.2.17 Affinity chromatography using magnetic beads

To isolate and enrich allele-specific binding proteins, we performed magnetic beads-based affinity chromatography. Streptavidin coupled magnetic beads (Dynabeads M-280, 10 mg/ml, Invitrogen) were washed using B&W buffer and Magnetic particle separator (Magna-Sep™, Invitrogen) according to the procedures provided by the manufacturer and supernatant was discarded by pipetting prior to coupling to biotinylated oligonucleotides. Magnetic beads were coupled with biotinylated oligonucleotides by incubating at 4 °C o/n. Different conditions was tested for optimal binding of beads to oligonucleotides, and found that both the concentration of oligonucleotides and beads, and incubation time were critical for the coupling efficiency of oligonucleotides to beads. To prevent undesired reaction with streptavidin, magnetic beads were incubated with the 1x B&W buffer containing 2 ng/µl free biotin for 1 h at RT. The magnetic beads were then washed two times with wash buffer followed by equilibration with 1x BB and incubated with nuclear extracts for 20 min in 1x BB containing 50 mM NaCl and 0.01% CHAPS using rotator. For establishing methods protein amount ranging from 500 µg – 7 mg nuclear protein from HIB 1B cells was used for binding reaction of one allele. Indeed, increasing amount of nuclear extracts resulted obviously in more enriched protein of interest eluted which were confirmed by EMSA and LC-MS/MS data, suggesting that protein amount is an important factor for efficient detection. To reduce non-specific DNA binding, poly (dI-dC) was subsequently added to the mixture and incubated further for 10 min. Competition with poly (dI-dC) supports the specific binding of proteins to the oligonucleotides [342]. Then, the supernatant containing unbound proteins was recovered. Subsequently, the beads were washed three times with binding buffer containing 10-50 mM NaCl and the bound proteins were eluted by 1x binding buffer with increasing concentration of NaCl (50, 100, 200, 300, 400, 500, 600 and 1000 mM) in a volume of 100 µl (eluate E50 – E1000). All steps were performed at 4°C. Finally, a 5-10 µl of protein from supernatants, washes and eluates were used for EMSA. The remaining eluates were subjected to LC-MS/MS analysis. All affinity chromatography experiments were replicated at least three times using same conditions.

5.2.18 Affinity chromatography using sepharose beads

Sepharose is a crosslinked, beaded-form of agarose-based gel filtration matrix. A common application for the material is in chromatographic separations of biomolecules within the broad fractionation range. Sepharose is often used in combination with some form of

activation chemistry, enzymes, antibodies and other proteins and peptides through covalent attachment to the resin. The most widely used method for activation is cyanogen bromide (CNBr) activation. Proteins and other molecules containing primary amino groups are then coupled directly to the pre-activated sepharose gel. First, 20 g of sephoarse beads were suspended in 200 ml H_2O and inverted several times. After 10 min, swollen beads were slowly stirred at RT and washed several times in the hood until the beads were free of any supernatant. The beads were then activated by adding of 5 g CNBr directly to the stirred bead slurry. pH and temperature were controlled during the activation. 1M NaOH was added to the mixture to keep pH at around 11-11.5. Temperature control (between 20 and 25 °C) was achieved by addition of a small piece of ice to the mixture. The reaction was finished completely within 15 min. The activated beads are quickly filtered and washed then several times with 5 bead volumes of H_20 and coupling buffer, and at the end transferred to a falcon tube containing the ligand in coupling buffer. 50 ml of the mixture in the falcon tube was incubated with continued stirring o/n at 4°C. The next day, the buffer was discarded and the beads were washed with 3 bead volumes of blocking buffer and incubated o/n at 4°C to block unbound residues. On the third day, the beads were washed with 5 bead volumes of coupling buffer, acetate buffer and TE buffer, respectively. The beads were then stored at 4°C in TE containing 10 mM sodium azide until use. Affinity chromatography using sepharose beads was performed as described previously [342] with some modifications. First, oligonucleotide sequence with $(AC)^5$-overhang (5'NH2-ACACACAC-3') was generated. 50 nmol of DNA pro 1 g sepharose was added to CNBr-preactivated Sepharose. 1mg of nuclear extracts was diluted in a final volume of 20 ml HA buffer without KCl. The (AC)5–Sepharose support was added to the mixture and equilibrated in the buffer. Then, the nuclear extract mixture was incubated for 10 min at 4 ◦C with heparin (20000 U Heparin pro ml CNBr-preactivated Sepharose) and 10 µg poly (dI-dC) as competitors. After washing with HA buffer (100 mM KCl), proteins were eluted with increasing concentration of KCl (200-1000 mM KCl) in a volume of 1 ml. All steps were performed at 4°C. Finally, input proteins, flow through, washes and fractions (Fraction 1-14) were collected and used for EMSA followed by LC-MS/MS analysis. The affinity chromatography experiment was performed only once.

5.2.19 Filter-aided sample preparation (FASP) and non-targeted liquid chromatography–tandem mass spectrometry (LC-MS/MS) (in close cooperation with Dr. Hauck HMGU)

Salt eluted fractions were processed as described before [269,327] in an adaptation of the FASP approach [343] using Microcon devices YM-30 (Millipore). The LC-MS/MS analyses were performed as described previously on a LTQ-Orbitrap XL (Thermo Scientific) [344] with the following adjustments: A nano trap column was used (300 µm inner diameter × 5 mm, packed with Acclaim PepMap100 C18, 5 µm, 100 Å; LC Packings) before separation by reversed phase chromatography (PepMap,25 cm,75 µm ID, 2µm/100 Å pore size, LC Packings) operated on a RSLC (Ultimate 3000, Dionex) using a nonlinear 170 min LC gradient from 5 to 31% of buffer B (98% acetonitrile) at 300 nl/min flow rate followed by a short gradient from 31 to 95% buffer B in 5 min and an equilibration for 15 min to starting conditions. 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 were at least doubly charged. During fragment analysis a high-resolution (60,000 full-width half maximum). The MS spectrum was acquired in the Orbitrap with a mass range from 300 to 1500 Da.

5.2.20 Protein identification and label-free relative quantification

The RAW files (Thermo Scientific) were further analyzed using the *Progenesis* LC-MS software (version 4.0, Nonlinear Dynamics), as described previously[344,345], with the following changes: Spectra were searched using the search engine Mascot (Matrix Science) against the Spectra were searched against the Ensembl mouse database (Release 69; 50512 sequences). A Mascot-integrated decoy database search using the Percolator algorithm calculated an average peptide false discovery rate of < 1% when searches were performed with a Percolator score cut-off of 13 and a significance threshold of $P < 0.05$. Peptide assignments were re-imported into *Progenesis* LC-MS. Normalized abundances of all unique peptides were summed up and allocated to the respective protein.

5.2.21 Enrichment of allele-specific binding proteins at predicted *cis*-regulatory variants

Pairwise comparison of the number of allele-specific binding proteins (fold-change > 2 or < 0.5, $P < 0.01$, Table 25 identified at the predicted *cis*-regulatory variants rs4684847, rs7647481 (165 and 142 proteins, respectively, 307 proteins for both *cis*-regulatory variants)

and at the predicted non *cis*-regulatory variants rs17036342 and rs2881479 (44 and 82 proteins, respectively, 126 proteins for both non *cis*-regulatory variants) with the respective set of all proteins identified at the respective variant (824, 869, 951 and 933 proteins for rs4684847, rs7647481, rs17036342, rs2881479, respectively) was performed using a two-sided two-group binomial test. *P*-values of *cis*-regulatory versus non *cis*-regulatory predicted variants are highlighted in bold. The test considers for each identified protein a variable X with X=1 if a significant, allele-specific differential binding (fold-change > 2 or < 0.5, $P < 0.01$, see Table 2) is found, and X=0 if no allele-specific binding for the identified proteins is found. Assuming Bernoulli distribution of X, with the unknown probability parameter *P* with X~B(*P*), X is equal to 1 with probability *P*. Comparing two groups, X~B(*P1*) for group 1 and X~B(*P2*) for group 2 (all proteins are assumed independent) is assumed with the hypotheses H0: *P1=P2* versus H1: *P1*!= *P2*. A *P*-value below 5% denotes that with confidence 95% one can say that two groups are not comparable.

5.2.22 GePS-tool GO-term and signaling pathway analysis

The Genomatix GePS-tool (Genomatix, Munich, Germany) was used to assess the enrichment of molecular function GO-terms (Supplementary table S1) and signaling pathways (Supplementary table S3) in the protein / gene data sets identified by LC-MS/MS, using all identified proteins.

5.2.23 GePS-tool transcription factor / transcriptional coregulator co-citation analysis

To calculate the enrichment of transcriptional co-regulators, the *Genomatix GePS tool* was used to build a co-citation based network for *YY1, NFATC4, and PRRX1*. All identified proteins/genes which are annotated as cofactors at the rs4684847 and rs7647481 (Supplementary table S2) were used as input gene list for the tool *Characterization of gene sets* and enrichment was calculated using the *Gene Ranker* tool. For visualization of co-citation interactions, the GePS tool *Characterization of gene sets* was used. All identified proteins / genes annotated as cofactors (Supplementary table S2) and the respective candidate transcription factors *PRRX1, YY1* and *NFATC4* were used as input gene list. Networks were created using the settings Co-citation level: *Function word level*; Co-citation filter: *1*; Network generation: *Simple network*; and Additional interactions per gene: *3*. Genes with interactions are shown in the Fig. 29A and Fig. 30.

5.2.24 Regression analysis of human adipose tissue samples

The insulin-resistance measure HOMA-IR and *YY1, RYBP* and *PRRX1* mRNA levels were measured in a cohort comprising 30 obese (BMI > 30 kg/m^2) otherwise healthy and 26 nonobese (BMI < 30 kg/m^2) healthy women [346], all pre-menopausal and free of continuous medication and investigated in the morning after an overnight fast. Venous blood sample was obtained for measurements of glucose, insulin and for preparation of DNA. HOMA-IR was calculated by the formula fP-Glucose (mmol/L) x (fS-Insulin (microU/ml)/ 22.5)[347]. Following blood sampling abdominal subcutaneous adipose tissue biopsy was obtained by needle aspiration and adipose microarray analysis was performed exactly as described [346] using the Affymetrix GeneChip miRNA Array protocol with 1μg of total adipose RNA from each subject (Gene and miRNA expression 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 with R, version 3.0.2 (R: A Language and Environment for Statistical Computing: from the R Development Core Team of the R Foundation for Statistical Computing, Vienna, Austria, 2014, http://www.R-project.org/) to assess correlation of YY1, RYBP and PRRX1 adipose tissue mRNA levels with HOMA-IR (adjusted for age, age/BMI and without adjustment) for all subjects as well as for risk-allele and nonrisk allele carriers. Adipose tissue samples were genotyped for rs1801282, rs4684847 and rs7647481 with a concordance rate of > 99.5% using the MassARRAY system with iPLEX™ chemistry (Sequenom, USA), as previously described [348]. The study was approved by the local ethics committee of the Karolinska University Hospital, Stockholm, Sweden; written informed consent was obtained from all patients who donated biological samples.

5.2.25 Statistical Analysis

All results were expressed as mean ± SD. Student´s *t* tests and Wilcoxon tests were used to compare two groups (two alleles/ complex and non-complex SNP regions). All statistical analyses were done using the Graph Pad Prism software v5.02 (GraphPad Software, CA, USA) or the Statistical Software SPSS v20.0 (IBM Corporation, NY, USA). The applied statistical methods for each experiment are given in the respective figure legend. Statistical differences of results were shown in P <0.05 (*), P <0.01 (**), P <0.001 (***).

6. Aim of study

GWAS have identified over 70 common risk loci for T2D [18](reviewed in Sun et al. 2014 [86]) [164,349]. Interestingly, most of the identified variants associated with diseases are located in non-coding regions of the genome, which might affect transcriptional regulation [100,259–262]. However, GWA signals typically identify one variant which tags numerous other variants in high LD. Such variants have rarely been traced to the causal variants and even more rarely to the mechanisms by which they may increase disease risk [98,350]. Thus, in most cases the causal *cis*-regulatory variants which might affect gene expression remain unknown. Also, little is known about regulatory effects of those variants on nearby genes (*cis*-eQTLs) or on genes at longer genomic distances (*trans*-eQTLs) [351]. Additionally, only few studies deciphered the allele specific binding of transcription factors (TFs) which ultimately affects gene expression [266,267,312]. Therefore, identifying TFs binding at genomic regions is an essential step for the understanding regulation of allele-specific gene regulation [352] and thereby the precise molecular mechanisms underlying associations between variants and disease risk. Combining of the functional cell type- and differentiation-specific epigenetic marks of regulatory region data [353] with the novel computational transcription factor binding sites (TFBS) modularity analysis, PMCA [98] enabled the prediction of *cis*-regulatory activity. With the growing interest in the field, there has been also an increasing need for methods to experimentally evaluate *cis*-regulatory variants and their functional consequence [354–356]. In this study, several variants at the different T2D risk-loci were selected, which were predicted to be *cis*-regulatory and non *cis*-regulatory by PMCA. Previous EMSA and reporter gene assay results [98] showed that predicted *cis*-regulatory variants could obviously alter the protein-binding capacity and transcriptional activity according to allele. Thus, the main aim of this study is to further investigate understanding the molecular mechanisms underlying the effect of *cis*-regulatory variants on gene expression at T2D associated loci including *PPARG*, *FTO* and *TCF7L2*.

Specifically:

Aim 1. To establish a highly sensitive label-free quantitative DNA protein interaction approach for systematic identification of allele-specific protein binding at variants predicted to be *cis*-regulatory.

Aim 2. To experimentally examine the presence of multiple causal variants within a given locus, providing more precise insights into the pathophysiological mechanism through newly identified transcriptional regulators and their coregulators binding to the variants.

To prove the power of the approach to find regulatory proteins at different loci:

2.1 At the *PPARG* locus (T2D) to further analyze if several variants also may modulate *PPARG* gene expression.

2.2 At the well-established *FTO* obesity risk and the *TCF7L2* locus T2D risk loci, to find proteins binding at predicted *cis*-regulatory variants.

7. Results

In the **first results chapter,** the ***Study design*** of the presented work including loci selection is introduced. In the **second chapter**, the ***development and optimization*** of a method for the identification of the allele-specific binding proteins at predicted *cis*-regulatory variants is described. In the **third chapter**, the application of this approach at the *PPARG* locus is presented including ***in-depth*** **analysis** of new identified transcription factors and their coregulators. The results verify computational *cis*-regulatory predictions and analysis of epigenetic marks on the regulatory region and support the power of the here introduced proteomics approach to find both, transcription factors and coregulators involved in allele-specific gene regulation. Parts of the results presented in this study are included in the following publications:

1. Claussnitzer M, Dankel SN, Klocke B, Grallert H, Glunk V, Berulava T, **Lee H**, Oskolkov N, Fadista J, Ehlers K, et al. 2014. Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms. *Cell* **156**: 343–358

2. **Lee H**, von Toerne C, Claussnitzer M, Hoffmann C, Glunk V, Wahl S, Breier M, Molnos S, Grallert H, Dahlmann I, Arner P, Hauner H, Hauck SM, Laumen H. 2014. Unbiased allele-specific quantitative proteomics unravels molecular mechanisms modulated by *cis*-regulatory variation at the *PPARG* locus (in revision).

7.1 Study design

In the past decade, GWAS have identified numerous risk loci associated with human diseases (reviewed in Sun et al. 2014 [86])[18,164,349]. Most of the identified variants in those loci are located in non-coding DNA regions [55,100,172], which hampered the further progress to assign the functional roles of those SNPs. Recently, advances in high throughput technologies enabled analysis of genome-wide expression quantitative trait locus (eQTL) [262,357–360], DHS-, RNA- and ChIP-seq [100,361] in target tissues or cell types related to diseases or traits, suggesting that *cis*-regulatory SNPs might affect transcriptional regulation [100,259–262]. Data based on eQTL [262,357,358,360,362], DHS-, RNA- and ChIP-seq [100,353,361] have been frequently used to define functional candidates for further investigation of disease causing genes. Moreover, bioinfomatics approaches support the prediction of *cis*-regulatory functionality, like the recently published phylogenetic module complexity analysis (PMCA) methodology assessing the occurrence of conserved patterns of TFBS in *cis*-regulatory modules (CRMs) within the

genomic region flanking a non-coding variant [98]. However, an essential problem facing previous studies is a lack of validation data for the identification of *cis*-regulatory SNPs. In only few studies, it was shown that several TFs bind differentially to genetic variants, which alter gene expression [98,266,267]. These results were recently confirmed and expanded by the ENCyclopedia Of DNA Elements (ENCODE) consortium [100]. However, these results can only annotate SNPs to genes. Since a gene can be involved in a variety of pathways, further annotation of genes to pathways, which represent certain biological mechanisms of the complex disease [257], still remains to be fully understood. Thus, understanding the precise molecular mechanisms underlying associations between variants and disease risk at the molecular level is a big challenge in human genetics.

In most biological processes, proteins always interact with DNA, RNA or other proteins. Interactions of proteins with nucleic acids (e.g. DNA, RNA) mediate metabolic and signaling pathways, cellular processes and organismal systems. Due to their central roles in biological function, protein interactions also control mechanisms resulting in healthy and diseased states in organisms [363]. Therefore, many researchers have attempted to identify proteins binding to *cis*-regulatory variants using quantitative protein-DNA proteomics [98,312,314]. Differential fluorescent dye labeling has been developed to allow relative protein quantification. Isotope labeling techniques (e.g., ICAT, ^{18}O-incorporation) have been integrated with LC-MS/MS for relative protein quantification. Protein samples are first labeled separately with stable isotopes either by metabolic incorporation or chemical reaction (reviewed in Wang et al. 2006 [322]). However, stable isotope labeling [312] or chemical labeling [314] faces different limitations. First, experiments for the identification of *cis*-regulatory SNPs often require human tissues of interest. However, such labeling approaches give some limited access to disease-relevant human tissues. Moreover, those approaches can be time-consuming, have limitations due to high-cost or inefficient labeling [329], may cause artifacts [330], and may be limited by missing data points due to under-sampling. For these reasons, in this study an effort was made to establish a label-free quantitative proteomics approach by testing various conditions at different predicted *cis*-regulatory variants and different loci (*PPARG*, *FTO*, *TCF7L2*) associated with T2D or obesity (**results chapter 7.2**). Subsequently, this study focused on the elucidation of the mechanisms mediated by *cis*-regulatory SNPs in the development of T2D through *in-depth* analysis of *PPARG* locus (**results chapter 7.3**).

This study includes development of a strategy for identification of allele-specific binding proteins that is particularly well suited for the identification of interacting protein partners.

This strategy involves three steps: *i) Sample preparation*, *ii) Discovery phase* and *iii) Validation*. *I) Sample preparation.* First, SNPs of interest were selected based on PMCA and evidence from published data. Oligonucleotides containing surrounding sequence around each SNP were labeled with Cy5 and used for EMSA experiments. To find optimal conditions for binding of proteins to DNA sequence surrounding SNPs, EMSA experiments were performed under various conditions including different concentration of poly (dI-dC), oligonucleotides, or amount of nuclear extracts [342]. After optimization of EMSA conditions, the same oligonucleotides labeled with biotin were bound to streptavidin immobilized magnetic beads, which in turn were incubated with nuclear extracts. The oligonucleotide-bound proteins were eluted from the beads with increasing concentration of NaCl. *II) Discovery phase.* Samples prepared were subsequently analyzed by LC-MS/MS. 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. The mass spectrometry data were analyzed using the software *progenesis*. *III) Validation.* Based on MS data and biological relevance, several candidate proteins were selected and proved by competition/supershift EMSA for their allele-specific binding to SNPs. In addition, functional assays such as siRNA knockdown and overexpression were performed to observe the effect of candidate proteins on target gene expression. Moreover, MS data were further analyzed by GePS for putative pathways involved and for protein-protein-interaction networks.

Selection of loci was based on the PMCA analysis [98] and epigenetic marks on the regulatory region, i.e. ChIPseq data [353]. In a previous study [98], PMCA analysis was applied to the GWAS loci associated with T2D, such as *MTNR1B*, *TCF7L2*, *PPARG*, *CENTD2*, *FTO, GCK, CAMK1D*, *KLF14* including 200 non-coding variants in LD with the respective tagSNPs ($r^2 \geq$ 0.7, 1000G) [164,364]. PMCA analysis predicted 64 complex and 136 non-complex variants at eight loci as complex, i.e. predicted a *cis*-regulatory function and a non *cis*-regulatory function, respectively. The allele-specific binding of proteins at the predicted *cis*-regulatory variants was confirmed by EMSA and luciferase gene assay *in vitro* [98].

PPARγ is an important regulator of adipogenesis, lipid- and glucose metabolism, and involved in whole-body insulin sensitivity (reviewed in Poulsen, Lars la Cour et al. 2012 [139])[140,141,144,365,366]. The association of *PPARG* gene with T2D and related diseases has been well demonstrated in various studies [166,167]. *PPARG* is the first reported gene for its reproducible association with T2D [163]. Many follow-up GWA studies have confirmed that polymorphisms in *PPARG* gene are associated with T2D [164–169]. It also has been reported that

PPARG genetic variants are associated with cardiovascular disease [367,368], obesity, metabolic syndrome and related traits obesity [126,369–371]. However, the questions still remain whether and how the genetic variants at the *PPARG* locus influence risk of T2D in the general population. ***FTO*** was a gene of unknown function in an unknown pathway [107], which was first cloned after identification of a fused toe (Ft) mutant mouse from a 1.6-Mb deletion on mouse chromosome 8 [186,187]. The *FTO* gene is expressed in a variety of tissues including brain, skeletal muscle, liver and adipose tissues [110,188–191]. A number of genetic variants in the *FTO* gene have shown the strongest associations with obesity [107,372,373], inflammation and cardiovascular disease risk [221] in Europeans. The strong associations of variants at the *FTO* locus with T2D were found to be through an association with BMI [107,164], which was confirmed by other studies [111,207,208]. However, the molecular mechanisms underlying the *FTO* variants in modulating obesity and obesity-related traits in different populations still remain unclear, which should provide valuable clues to the exact biological roles of FTO. **TCF7L2** (also known as TCF4) is involved in the Wnt signaling pathway [374], which is essential for regulating cell morphology, proliferation, motility, oncogenesis and tumor suppression [374,375]. The *TCF7L2* gene is well known as the most important T2D susceptibility gene [76]. Since its discovery, the association has been replicated in different ethnic groups [235,240–242,244–247]. Despite intensive research, it remains an open question how mechanistically genetic variants at the *TCF7L2* locus affect the risk of T2D. For these reasons, several variants were selected from the different loci including *PPARG* (rs4684847, rs7647481, rs17036342, rs2881479), *FTO* (rs1421085) and *TCF7L2* (rs7903146) for further studies.

Here, the allele-specific binding of proteins to DNA sequence was first confirmed and optimized under various conditions by EMSA experiments. Afterwards, affinity chromatography using bead based-approaches enabled isolation and enrichment of the allele-specific binding proteins. After the sample preparation procedure, the proteins present in each eluate were proteolytically digested, detected and quantified by label-free shotgun peptide sequencing LC-MS/MS analysis. Subsequently, the proteins were identified via comparison with public available data (i.e., Ensemble). After selection of the candidate proteins, their allele-specific binding to the target DNA sequence was validated by competition and supershift EMSA. Next, the proteomic data were further analyzed using pathway and GO term-based GePS tool (Genomatix, Munich, Germany) to investigate the upstream signaling pathways involving the identified proteins and predict protein-protein-interactions. Finally,

siRNA knockdown of identified TFs was performed to investigate the biological meaning at the *PPARG* locus in adipogenesis (see chapter 7.3 and Claussnitzer et al. 2014 [98]).

7.2 Development of a method based on magnetic beads immobilized affinity chromatography coupled to label-free proteomic analysis

This section consists of four parts: *i)* confirmation of allele-specific binding of proteins at the predicted *cis*-regulatory variants by EMSA, *ii)* isolation/enrichment of the allele-specific binding proteins by affinity chromatography, *iii)* detection and identification of proteins by LC-MS/MS and *iv)* the verification of identification using functional assays. In the *i)* EMSA and *ii)* affinity chromatography parts, various conditions were tested to optimize conditions which facilitated more systematic approach for the entire method development. For the development of the method, four variants predicted as *cis*-regulatory SNPs [98], rs4684847, rs7647481 (*PPARG*), rs1421085 (*FTO*) and rs7903146 (*TCF7L2*) were selected at the three different loci, *PPARG*, *FTO*, *TCF7L2*. PMCA analysis [98] and epigenetic marks of regulatory region data [353] supported that both SNPs, rs4684847 and rs7647481 at the *PPARG* locus might contribute to *PPARG* regulation, which showed consistent cell stage-dependent density distributions [353]. The rs1421085 at the *FTO* locus is located within a highly conserved intronic regulatory element, and the previous fine-mapping study predicted this variant to have allele-specific binding affinities for different transcription factors [376], which was confirmed by PMCA [98]. Thus, this variant is the most interesting candidate for follow-up functional evaluation [376]. The rs7903146 is the most well-established variant at the *TCF7L2* locus for its association with T2D [76,249,377,378]. The rs7903146 showed consistent T2D association in samples across diverse ethnic groups [379], and the rs7903146 T risk allele exhibited islet-selective epigenetic marks of regulatory region in human islets [254]. Claussnitzer et al. demonstrated that the rs7903146 created both, allele-specific binding of proteins and luciferase reporter activity in a beta-cell line. Interestingly, the rs7903146 T risk allele constructs showed significantly greater enhancer activity than the C nonrisk allele [98], which is consistent with the previous findings [254,255]. The allele-specific regulatory properties for this variant have been largely limited to pancreatic beta cells [98,254,255]. However, Savic et al. confirmed these findings and further exhibited the allelic-specific properties of the rs7903146 in other cell lines including myoblasts and neuronal cells [380]. Together, these data suggest that the rs7903146 at the *TCF7L2* locus may be of interest in this study.

7.2.1 *PPARG* locus

7.2.1.1 Optimization of conditions for allele-specific binding of proteins

The PMCA approach predicted six out of 24 non-coding variants at the T2D associated *PPARG* locus as complex, i.e. predicted a *cis*-regulatory function (tag SNP rs1801282). Among the six variants, the rs4684847 was solely observed with cell stage-dependent histone H3-lysine 27 acetylation (H3K27ac) density distributions. In addition, only the rs4684847 showed direct overlap to a distinct homeobox TFBS matrix as a specific feature of T2D susceptibility variants inferred from PMCA [98]. Based on these data, the rs4684847 was assumed be a *cis*-regulatory variant at the *PPARG* locus. Indeed, the homeobox overlap analysis inferred binding of the homeobox TF PRRX1 to the rs4684847, which was also identified here by affinity chromatography (see chapter 7.2.1.2) and was confirmed as a repressor of *PPARG2* expression in further experiments [98].

In order to assess DNA-protein interaction properties, EMSA was performed using Cy5-labeled 40 bp oligonucleotide sequence surrounding the SNP rs4684847 (C/T) at a mid-position. Initially, nuclear extracts isolated from two cell lines, 3T3-L1 (mouse white pre-adipocytes) and HIB 1B (mouse brown pre-adipocytes) were used in EMSA experiments (Fig. 5A). In line with the previous data [98], allelic change from C to T at the site of the rs4684847 led to slight reduction in signal intensity in both cell lines, 3T3-L1 adipocytes (day 15) and HIB 1B adipocytes (day 9). The slight allelic difference in 3T3-L1 adipocytes might result from background signal around the allele-specific band (lane 1-2), while HIB 1B adipocytes indicated obvious allelic difference in the formation of DNA-protein complex (lane 3-4) (Fig. 5A). In order to control the quality of nuclear extract prepared, EMSA was performed using oligonucleotides containing either specificity protein 1 (Sp1) or Octamer-binding transcription factor 1 (Oct-1) consensus binding sites, respectively. EMSA results revealed that Sp1 and Oct-1 binding was detected in 3T3-L1 adipocytes (lane 1 and 2, respectively) as well Sp1 in HIB 1B adipocytes (lane 3) (Fig. 5B). Since obvious allelic difference is necessary for isolation of the allele-specific binding proteins, EMSAs were performed under various conditions such as different amount of nuclear extracts, and different concentration of poly (dI-dC) and Cy5-labeled probes for optimal protein-DNA interaction. The signal intensity of the allele-specific band was obviously reduced with decreased amount of nuclear extracts in both cell lines, indicating that the amount of protein is one of the critical parameters for EMSA band intensity (Fig. 5C). Poly (dI-dC) is the most widely used non-specific competitor

as synthesized double stranded DNA fragments to prevent non-specific binding proteins [381]. As expected, the incubation with poly (dI-dC) (4.4 ng/µg) showed slight, but visible reduction in non-allele-specific signal and led to the more obvious allelic difference in protein binding (lane 7-12) compared to without poly (dI-dC) in 3T3-L1 adipocytes (lane 1-6) (Fig. 5C). Moxley et al. reported that high concentration of poly (dI-dC) (in range 100–400 ng/µl) could exhibit little diminishing effect on non-specific binding, but also on specific binding [342]. In further EMSA experiments, using two-fold serial dilution of poly (dI-dC) from 0 to 88 ng/µl enabled to optimize obvious specific binding and reduce non-allele-specific signal. In the range of 35 to 70 ng/µl of poly (dI-dC), there is no significant difference in allele-specific complex formation (data not shown). Thus, in all subsequent EMSAs and affinity chromatography, the concentration of poly (dI-dC) (35 ng/µl) was used as this concentration showed an increased specific binding activity of proteins with non-deleterious effect on specific shift band, and for economic purpose. Next, EMSA was performed using 10 µg HIB 1B nuclear extracts with different concentration of Cy5-labeled oligonucleotides to observe the dependence of allele-specific binding of proteins on the DNA concentration. To determine its optimal concentration, an oligonucleotide titration (in the range 1.2 pM-77 fM) was performed using fixed amount of nuclear extracts (10 µg) (Fig. 5D). As shown in Fig. 5D, the concentration of oligonucleotides (in range 19.3-77 fM) seemed to be directly correlated with the general signal intensity. For more selective binding of protein of interest, low concentration of oligonucleotides was suggested as the concentration should be sufficient for obvious binding of protein of interest [342]. These experiments were repeated at least three times and revealed similar results (data not shown). Based on the EMSA results, the condition; 5 µg nuclear extracts, 38.5 fM of oligonucleotides and 35 ng/µl of poly (dI-dC), was chosen for further EMSA experiments. In order to assess specific nuclear protein binding to the rs4684847 during adipogenesis, EMSA was performed using nuclear extracts from 3T3-L1 (lane 1-4 and 10-13) and HIB 1B (lane 5-6 and 14-15) at different adipogenesis stages (Fig. 5F). Interestingly, the signal intensity of the allele-specific binding proteins was decreased during adipogenesis in 3T3-L1 cells (Fig. 5F), and a similar pattern was obtained also in HIB 1B cells in repeated experiments (data not shown). In addition to nuclear extracts from adipocytes, nuclear extracts from 293T (Human embryonic kidney cells) (lane 7 and 16), Huh7 (Human hepatocytes) (lane 8 and 17) and INS-1 (Rat pancreatic beta cells) (lane 9 and 18) were included for comparative purposes. Similar migration patterns were obtained in all cell lines tested (Fig. 5F). However, a very weak allele-specific complex formation was observed when incubated with nuclear extracts from INS-1 cells compared to other cell lines

analyzed. Taken together, these results confirmed and extended the previous observations that the rs4684847 variant (C > T allele) does modulate an allele-specific binding to particular proteins under certain conditions.

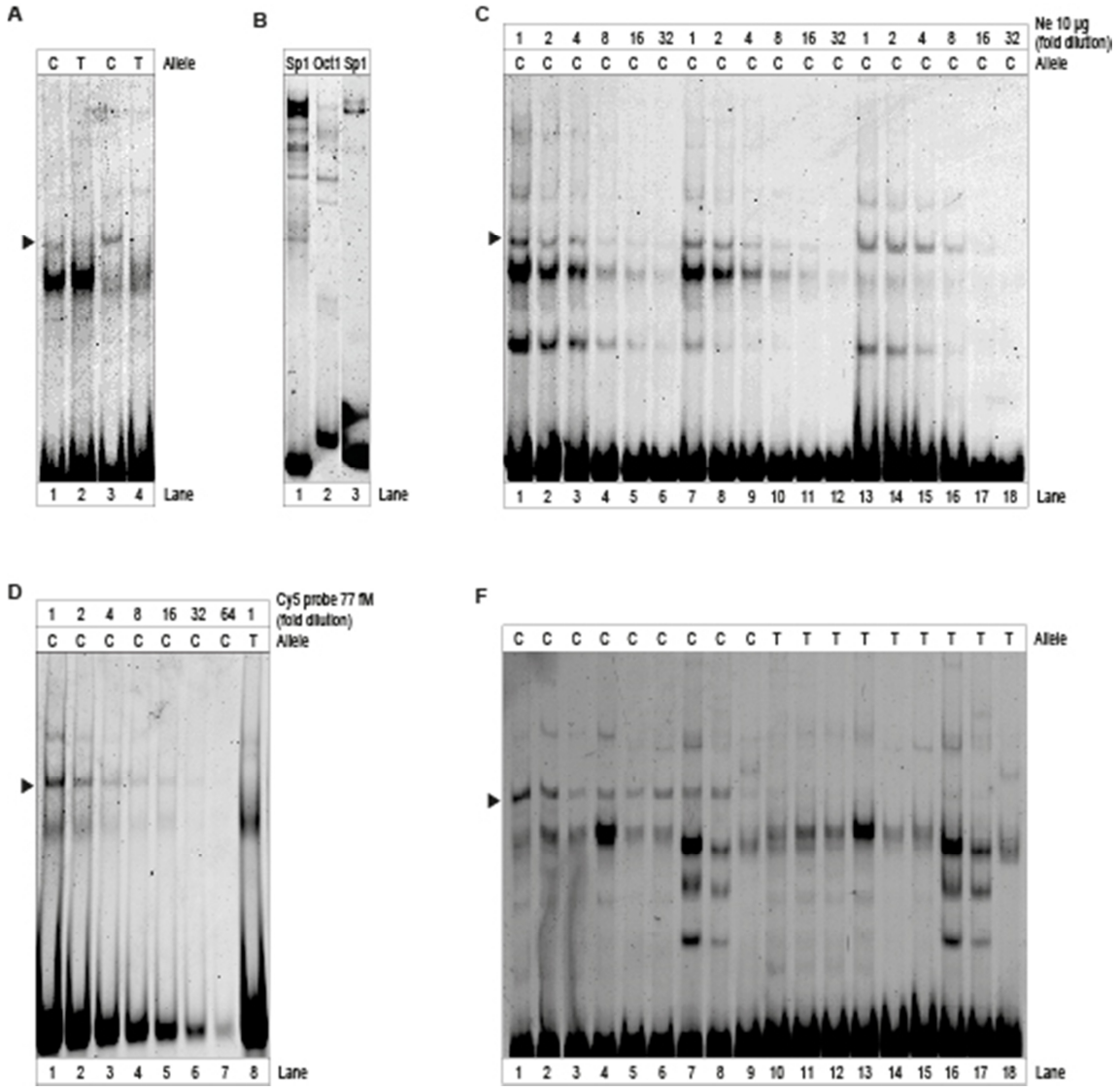


Figure 5. Analysis of allele-specific protein-DNA interaction at the predicted *cis*-regulatory variant, rs4684847 of the *PPARG* locus.

To examine the allele-specific binding of proteins at the rs4684847 of the *PPARG* locus, EMSA experiments were performed using allelic Cy5-labeled probes for the rs4684847 and nuclear extracts from different cell lines. (A) The predicted *cis*-regulatory SNP, rs4684847 showed allele-specific differential binding affinity of proteins in 3T3-L1 adipocytes (lane 1-2) and HIB 1B adipocytes (lane 3-4). (B) Sp1 and Oct-1 motif oligonucleotides were used in EMSA as control for quality and quantity of nuclear extract from 3T3-L1 adipocytes (lane 1-2) and from HIB 1B adipocytes (lane 3). (C) Cy5-labeled probe for the rs4684847 was incubated with decreasing

amounts of nuclear extracts. Reduced amount of nuclear extract proteins (in the range 0.3-10 µg) resulted in decreased total signal intensity, thereby improving resolution of the allele-specific bands in both 3T3-L1 adipocytes (lane 1-12) and HIB 1B adipocytes (lane 13-18). In addition, 44 ng/µg of poly (dI-dC) was added to the reaction to reduce non-allele-specific signal (lane 7-12). (D) Signal intensity of the allele-specific band showed dependence on concentration of Cy5-labeled oligonucleotides (in the range 1.2-77 fM). (F) EMSA was performed using nuclear extracts from different cell lines: 3T3-L1 during the entire differentiation on day 0 (lane 1 and 10), day 1 (lane 2 and 11), day 3 (lane 3 and 12) and day 15 (lane 4 and 13), HIB 1B pre-adipocytes (day 0) (lane 5 and 14) and HIB 1B adipocytes (day 9) (lane 6 and 15), 293T cells (lane 7 and 16), Huh7 cells (lane 8 and 17) and INS-1 cells (lane 9 and 18). A black arrow indicates the allele-specific band.

Next, EMSA was performed in order to determine to which extent shortened oligonucleotides might improve specific DNA-protein signals and reduce non-allele-specific signal. Previously, Cy5-labeled 40 bp oligonucleotides (C/T allele at position 21, (+) strand) (Fig. 6A) were used for EMSA experiments. Here, 31 bp oligonucleotides (C/T allele at position 21, (+) strand) (Fig. 6B) by reduction 9 bp oligonucleotides were used for comparison with 40 bp oligonucleotides. To compare putative binding of transcription factors to the DNA sequence between 40 bp to 31 bp, an *in silico* analysis was performed using a bioinformatics tool, MatInspector (Genomatix, Munich, Germany). The analysis using the MatInspector software revealed that seven TF families could bind solely to 40 bp oligonucleotides as described in Table 17. Among these TF families, four predicted TF families such as *Brachyury gene, mesoderm developmental factor*, *Interferon regulatory factors*, *Heat shock factors* and *SOX/SRY-sex/testis determinig and related HMG box factors* directly overlap to the allele position (position 21, (+) strand), and the variant directly affected the core of these TFBS (data not shown). In turn, the destruction of TFBS could abrogate its ability to transactivate the target gene. Moreover, the other TFBSs lying outsides of allele position could also influence the transcriptional regulation with other transcription factors in regulatory networks [382]. Indeed, the comprehensive cross-species TFBS pattern analysis demonstrated that the T2D-distinct clustering of the homeobox TFBS matrix such as homeobox TF PRRX1 binds in close proximity to the C/T allele of the rs4684847 at the *PPARG* locus. Moreover, in the recent publication by Claussnitzer and colleagues [98]. PRRX1 was further demonstrated to be a repressor of *PPARG2* expression in adipose cells and its adverse effect on lipid metabolism and systemic insulin sensitivity, dependent on its binding at the rs4684847 risk allele [98].

Matrix Family	Detailed Family Information	Binding position[a]	Strand[a]	Matrix similarity[b]	Sequence (5'-3')
V$BRAC	Brachyury gene, mesoderm developmental factor	14-36	-	0.888	tcttttcggagtTGTGagaatta
V$ZF35	Zinc finger protein ZNF35	25-37	+	0.877	actccgAAAAgat

V$KLFS	Krueppel like transcription factors	20-36	-	0.849	tcttttcGGAGttgtga
V$IRFF	Interferon regulatory factors	20-40	+	0.823	tcacaactccGAAAagataag
V$HEAT	Heat shock factors	8-32	-	0.821	ttcggagttgtgAGAAttagagatg
V$SORY	SOX/SRY-sex/testis determinig and related HMG box factors	9-33	+	0.820	atctctAATTctcacaactccgaaa
V$IKRS	Ikaros zinc finger family	22-34	-	0.818	ttttCGGAgttgt

Table 17. TF families identified by *in silico* analysis (MatInspector, Genomatix, Munich, Germany) at the 40 bp sequence with the rs4684847 variant at midposition.

[a]Nucleotide position relative to +1 and Watson (+) or Crick (-) strands, [b]Sequence similarity relative to the matrix.

Subsequently, EMSA experiments were carried out using 31 bp and 41 bp oligonucleotides, respectively. In addition, nuclear extracts from 3T3-L1 and HIB 1B adipocytes were used in EMSA, respectively. The signal intensity of the allele-specific band increased at the C allele with reduced non-allele-specific signal by using 31 bp oligonucleotides in both 3T3-L1 and HIB 1B adipocytes (Fig. 6C and D, respectively), which was also seen at the T allele. However, the use of 31 bp oligonucleotides resulted in large smears of signal migrated at the T allele in close proximity to the allele-specific band with poor resolution, which could be disadvantageous for isolation of the allele-specific binding proteins. In addition, 9 bp shortening of oligonucleotides (40 →31 bp) could lead to the alteration in protein-protein-interactions and disrupt the TFBS-modules predicted by PMCA prediction [98]. The *in silico* analysis (Table 17) and EMSA experiment (Fig. 6) indicated that the use of 31 bp oligonucleotides could provide rather disadvantage compared to 40 bp oligonucleotides. For these reasons, the use of 31 bp oligonucleotides was excluded for further analysis.

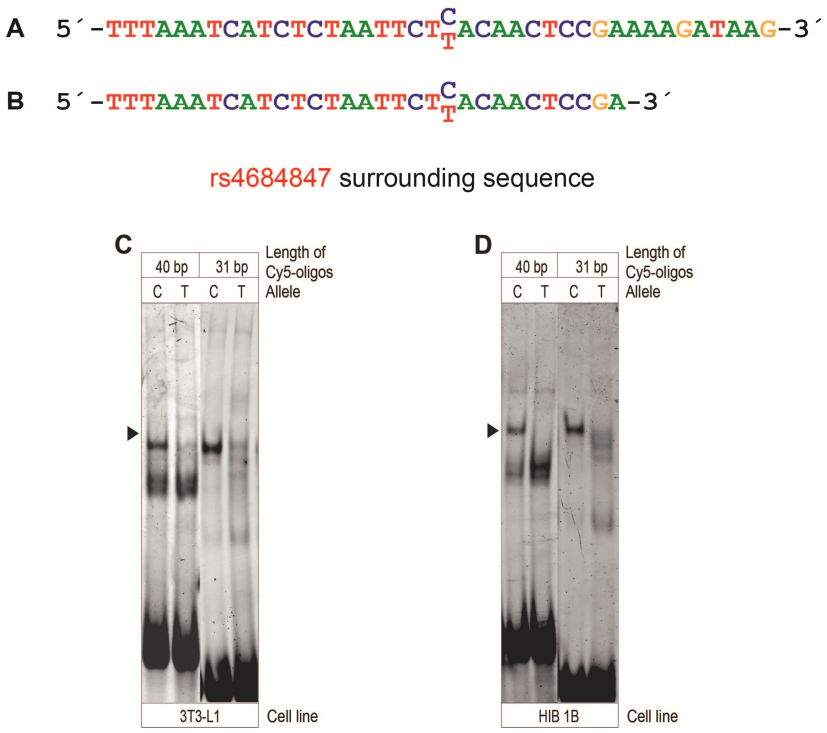


Figure 6. Analysis of allele-specific protein-DNA interaction at the predicted *cis*-regulatory variant, rs4684847 using variable length of Cy5-probes.

EMSA experiments with the rs4684847 allelic Cy5-labeled probes of variable length were performed to examine dependence of the allele-specific binding properties of proteins on the length of oligonucleotides. (A-B) The predicted *cis*-regulatory SNP, rs4684847 surrounding sequence for 40 bp (A) and 31 bp (B). (C-D) EMSAs with different length of the rs4684847 allelic Cy5-labeled probes in 3T3-L1 adipocytes (C) and HIB 1B adipocytes (D). A black arrow indicates the allele-specific band.

7.2.1.2 Affinity chromatography coupled to mass spectrometry: optimization

Identification and characterization of the allele-specific binding proteins is essential to understand the molecular mechanism modulated by *cis*-regulatory variants at T2D associated loci and their effects on gene expression. The isolation and identification of proteins of interest have often been provided by a combination of gel electrophoresis with mass spectrometry (MS). However, affinity chromatography can also be used as part of a traditional gel electrophoresis–MS workflow for effective protein separation, which would be also useful for the identification of protein–protein interactions [383]. Several studied followed this workflow and achieved desired results [98,327,384]. Here, isolation of the allele-specific binding proteins was performed using oligonucleotides trapping method, which has been used as a powerful tool for protein purification in several studies [342,385]. For establishment of the

method, two methods using *i)* magnetic beads and *ii)* sepharose beads were tested under various conditions. In both cases, EMSA was routinely performed immediately after affinity chromatography to observe the presence and enrichment of the target proteins. Following protein isolation and enrichment, liquid chromatography tandem-mass spectrometry (LC-MS/MS) was used as a very sensitive, accurate and efficient method for detection and characterization of proteins [321]. Stable isotope labeling in cell culture is not suitable for detection of small protein amounts [344]. However, TFs are usually low abundant in cells (ranging between 10^3 and 10^5 molecules per cell) and the purification of TFs requires appropriate cells or tissues [386]. Thus, in this study a label-free quantitative proteomics technique is used. The so called label-free LC-MS/MS-based comparative proteomics allows accurate quantification of tissue samples without additional error during *in vitro* labeling reactions [344] and is faster, cleaner, and provides simpler results (reviewed in Zhu et al. 2010 [387]). All works of mass spectrometry part in this study were performed in close co-operation with the group of Dr. Hauck from Research Unit Protein Science, Helmholtz Zentrum München, Germany.

7.2.1.2.1 Small scale affinity purification

First, a streptavidin-biotin system was applied for isolation and enrichment of the allele-specific binding proteins. Purification of DNA-binding proteins using magnetic beads has been widely used and described in several studies [98,312,314–316,320,327,388]. In this system, streptavidin linked magnetic beads (Dynabeads® M-280 Streptavidin (Invitrogen), a 2.8 µm spherical structure with a monolayer of streptavidin) were first coupled to biotinylated oligonucleotides (Fig. 7A). For the purification of proteins, two cell lines, HIB 1B adipocytes (day 9) and 3T3-L1 pre-adipocytes and adipocytes (day 0 and day 15, respectively) were selected due to the obvious allelic difference observed in EMSA experiments (see chapter 7.2.1.1) and the biological relevance for both, the *PPARG* gene and T2D [98,165–171]. In the first affinity step, each allelic biotinylated oligonucleotide (26 pmol) containing the identical sequence as Cy5-labeled probes was coupled to the Dynabeads® M-280 beads (250 µg corresponding to the binding capacity of 50 pmol biotinylated oligonucleotides) as described in Methods (see chapter 5.2.17) (Fig. 7A). Since the use of 26 pmol oligonucleotides do not lead to saturation of the entire binding sites for 50 pmol on surface of magnetic beads, an excess of free biotin was added to the reaction solution to prevent the non-allele-specific binding of proteins. The following incubation with HIB 1B nuclear extracts (250 µg HIB 1B nuclear extracts per each allele) should result in binding of proteins of interest to the

biotinylated oligonucleotides coupled to the beads. In affinity chromatography, detergents were reported to improve the protein-protein interaction complexity [342]. Non-ionic and mild surfactants such as Triton X-100 or CHAPS were found to have no significant effect on trypsin digestion with surfactant concentrations up to 1%. However, they interfered with the subsequent peptide analysis by MALDI-MS [389]. They were also shown to interfere with MS analysis in this study (data not shown). For these reasons, the amount of CHAPS was used less than 1% (v/v) in the reaction. Moreover, to prevent undesired non-allele-specific DNA-protein interactions in binding to DNA-magnetic beads, the binding sites were competed with non-specific DNA-binding proteins using poly (dI-dC). Moxley et al. indicated that poly (dI:dC) showed a slight detrimental effect from the range 100–400 g/ml, but little effect in the range of 6 to 50 g/ml in EMSA experiments [342]. Thus, relative low concentration of poly (dI-dC) (70 ng/µl) was added to each reaction solution followed by incubation for 20 min at 4 °C. After 20 min. incubation at 4 °C, the beads with the specific protein adsorbed were washed three times with 1x binding buffer containing 50 mM NaCl. Some non-specific proteins could still bind the magnetic beads. The addition of excess of poly (dI-dC) to the reaction and the subsequent stringent wash steps will, however, largely remove the contamination with non-specific proteins in solutions [316]. The beads were washed three times with wash buffer. Afterwards, the specifically bound proteins were eluted by increasing salt concentration which dissociates the DNA-bound proteins from the magnetic beads. The proteins were eluted by resuspending in elution buffer containing 400 mM to 1000 mM NaCl. After 2 min. incubation, the beads were removed by magnetic separation. Finally, EMSA was performed to monitor the efficiency of isolation and enrichment of the specific proteins using all collected samples (Fig. 7B). During the elution process, the binding conditions of proteins used in EMSA could be altered caused by the possible change of buffer composition such as pH and ionic strength, which could vary the stability of proteins [390]. Thus, when complexes are not stable in solution, it may require very short runs so that the observed binding pattern is closely approximate to the distributions of species present in the initial samples [277]. Additionally, running time of electrophoresis is dependent on the length of nucleic acid and protein complex [277], which could be also reduced since protein fractions in samples might be highly enriched for the sequence-specific DNA-binding protein after affinity chromatography [316]. Moreover, slow dissociation could also result in underestimation of binding density [277]. For these reasons, the electrophoresis was performed in shortened time (3 h → 1.5 h) to avoid the loss of the specific protein activities. To observe the presence of the allele-specific binding proteins in eluates, EMSA was performed using Cy5-labeled C-allele probe which showed

stronger allelic binding in previous EMSAs (Fig. 6B). Overall, neither obvious nor differential signal for the allelic band was observed (Fig. 7B), which was consistent with results from a silver staining experiment confirming no difference between both alleles (data not shown). Taken together, the results suggested that there was still need to improve the enrichment of the target proteins in the eluates. Thus, increased amount of HIB 1B nuclear extracts was used for further affinity chromatography (125 µg → 500 µg) with the equal amount of biotinylated oligonucleotides (26 pmol) in order to determine solely the effect of protein amount for the enrichment of proteins. Indeed, after the affinity chromatography using four times more amount of HIB 1B nuclear extracts, the allele-specific binding of proteins was obviously observed in eluate E400 for the C allele while extremely faint signal for the target proteins was seen in eluate E400 for the T allele, as assessed by EMSA (Fig. 7C). This result indicated that the amount of protein extracts is a critical factor for the enrichment of proteins. Finally, the eluates E400 for both alleles were applied for MS analysis.

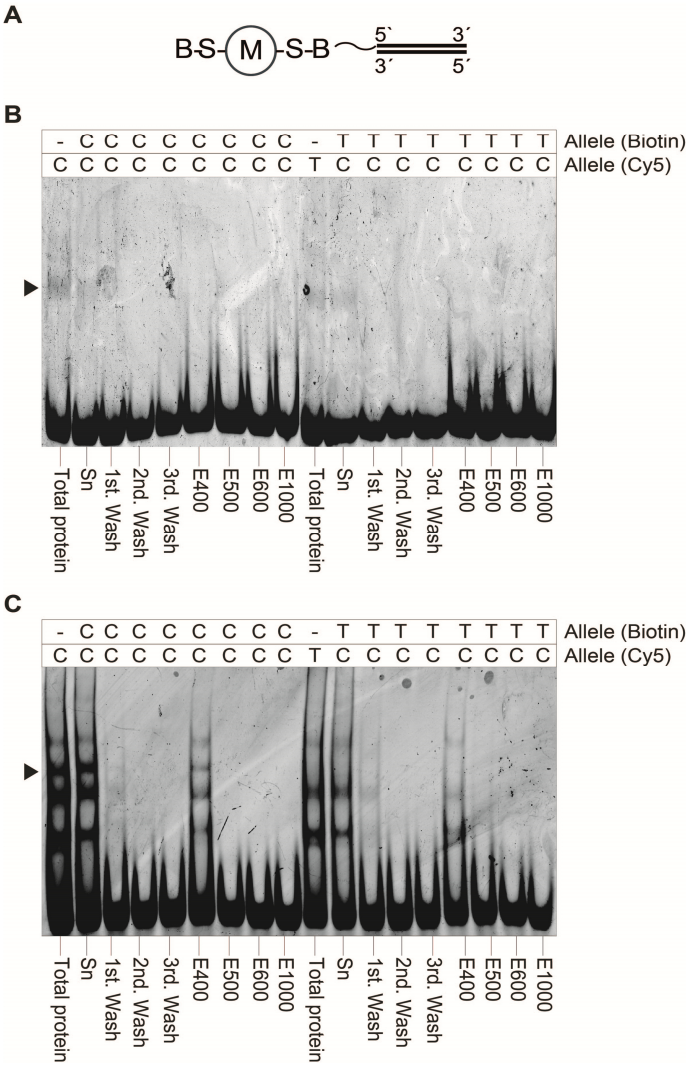


Figure 7. EMSA analysis of eluates obtained during affinity purification process for the rs4684847 using nuclear extracts from HIB 1B adipocytes.

(A) Immobilization of biotinylated oligonucleotides on streptavidin encoded magnetic beads. B: Biotin, M: Magnetic bead, S-Streptavidin. (B)-(C) Nuclear extracts prepared from HIB 1B adipocytes were subjected to affinity purification as depicted in Methods (see 5.2.17). The initial nuclear extracts and specified chromatographic fractions (wash and eluates) were analyzed to confirm the enrichment of the allele-specific binding proteins using C allelic Cy5-labeled probe of the rs4684847. A black arrow indicates the allele-specific band. *Total protein:* nuclear extracts from HIB IB adipocytes, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (50 mM). *E400-1000*: Elution of proteins with increasing concentration of NaCl (400 -1000 mM NaCl, respectively). Affinity chromatography using magnetic beads was performed using 125 µg (B) and 500 µg (C) nuclear extracts from HIB 1B adipocytes for each allele, respectively. All experiments were performed in triplicates.

As shown in the previous EMSA (Fig. 5F), the signal of the allele-specific band was more intensive in pre-adipocytes than that in adipocytes, suggesting that the allele-specific binding proteins might be regulated during adipocyte differentiation. To explore this question, the affinity chromatography was performed using nuclear extracts from both, 3T3-L1 pre- and adipocytes under the same condition (125 µg of nuclear extracts and 26 pmol of biotin-labeled oligonucleotides for each allele) as above (Fig. 6B). As shown in Fig. 8A, the allele-specific band was faintly detected by EMSA using C allelic probe of the rs4684847 in the 3T3-L1 pre-adipocytes. However, this band was more clearly seen in 3T3-L1 adipocytes (Fig. 8B). Interestingly, these results were in contrast to the previous EMSA data (Fig. 5F), which will be discussed in a later part of discussion (see chapter 8.2). For further identification of the allele dependent binding proteins in 3T3-L1 adipocytes and comparison with those in HIB 1B adipocytes, all E400 eluates from 3T3-L1 adipocytes were analyzed by LC-MS/MS.

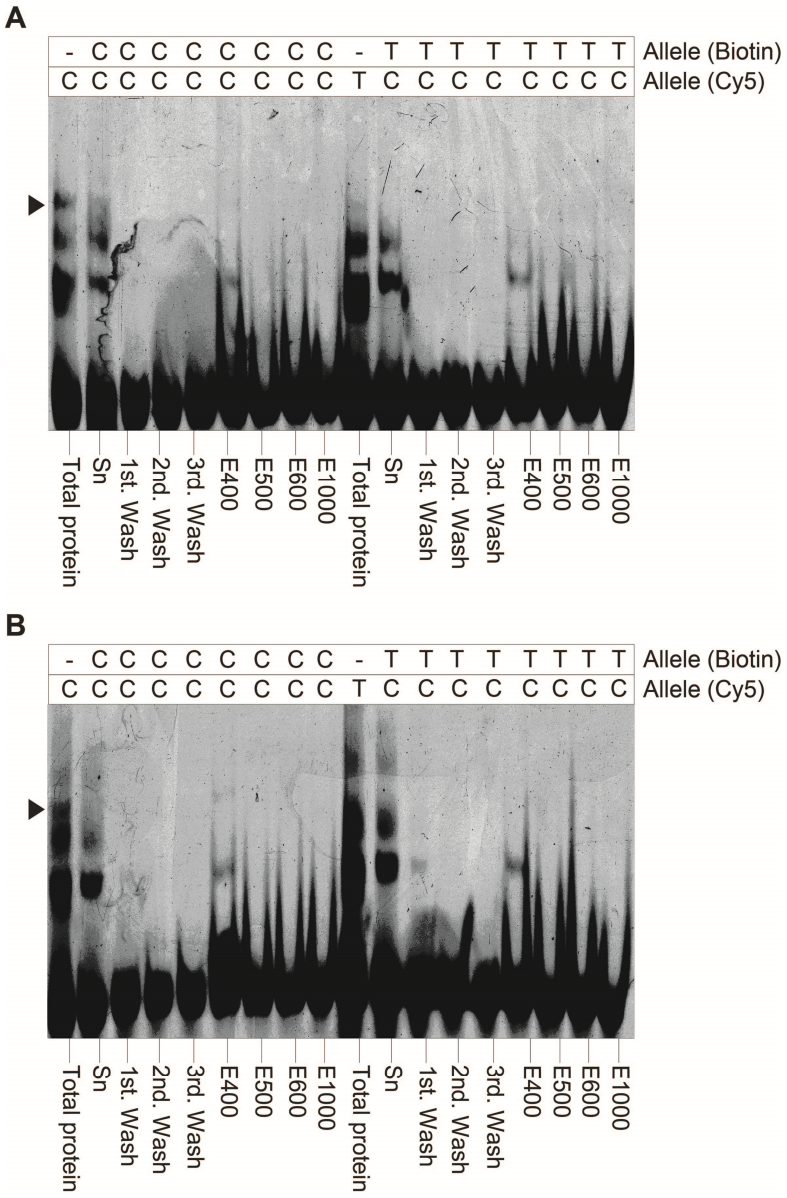


Figure 8. EMSA analysis of eluates obtained during affinity purification process for the rs4684847 using nuclear extracts from 3T3-L1 pre- and adipocytes.

Nuclear extracts prepared from 3T3-L1 pre- (A) and adipocytes (B) were subjected to affinity purification as depicted in Methods (see 5.2.17). The initial nuclear extracts and specified chromatographic fractions (wash and eluates) were analyzed to confirm the enrichment of the allele-specific binding proteins using C allelic Cy5-labeled probe of the rs4684847. A black arrow indicates the allele-specific band. *Total protein:* 3T3-L1 nuclear extracts, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (50 mM). *E400-1000*: Elution of proteins with increasing concentration of NaCl (400 -1000 mM NaCl, respectively). Affinity chromatography using magnetic beads was performed using 125 µg from 3T3-L1 pre- (A) and adipocytes (B) for each allele, respectively. All experiments were performed in duplicates.

All E400 eluates collected from affinity chromatography (Fig. 7 and Fig. 8) were analyzed quantitatively by LC-MS/MS using *Progenesis* software The LC-MSMS analysis was performed on an Ultimate3000 nano HPLC system (Dionex, Sunnyvale, CA) online coupled to a LTQ OrbitrapXL mass spectrometer (Thermo Fisher Scientific) as described previously [344]. From the high resolution MS pre-scan, 10 most abundant peptide ions were selected for fragmentation. During fragment analysis a high-resolution (60,000 full-width half maximum) MS spectrum was acquired in the Orbitrap with a mass range from 300 to 1500 Da [344]. The RAW files (Thermo Scientific) were analyzed using the *Progenesis* LC-MS software (version 4.0, Nonlinear Dynamics) [344,345]. Spectra were searched using the search engine Mascot (Matrix Science) against the Ensembl mouse database (Release 69; 50512 sequences) (see Methods). In total, 208 proteins were detected in E400 eluates (n = 2) obtained from affinity chromatography using 125 µg 3T3-L1 pre- and adipocytes nuclear extracts, respectively. In contrast, 667 proteins were detected in E400 eluates (n =3) obtained from affinity chromatography using 500 µg HIB 1B nuclear extracts (Table 18). Even if the results came from two different cell lines (Table 18), it suggests that the amount of protein input in affinity chromatography might be, to some extent, proportional to the number of proteins detected by LC-MS/MS.

Cell line	Allele-specific signal by EMSA	NE amount applied for AC	No. of proteins identified[a]
3T3-L1 (d0)	faint, but visible	125 µg	208
3T3-L1 (d15)	faint, but visible	125 µg	208
HIB 1B (d9)	clear visible	500 µg	667

Table 18. Total number of allele-specific binding proteins identified by mass spectrometry at the predicted *cis*-regulatory SNP, rs4684847 in 3T3-L1 pre- and adipocytes, and HIB 1B adipocytes.

[a]Number of all proteins identified by LC-MS/MS, NE: nuclear extracts, AC: affinity chromatography. LC-MS/MS data were quantitative analyzed by *Progenesis*.

However, most of proteins identified in both 3T3-L1 pre- and adipocytes were non-specific cytoplasmic proteins including keratin. Only few proteins were found as transcription factors including TF1 (referred to as TF1 in this thesis as data are not yet published) and PRRX1 (recently published based on these experiments) [98], however with non-significant fold change (Table 19), which might be resulted from poor enrichment of proteins during purification process. For this reason, it was difficult to compare and select the interesting proteins responsible for adipogenesis between 3T3-L1 pre-adipocytes and adipocytes.

Next, for the selection of candidate proteins involved in *cis*-regulatory activity, 667 proteins identified by LC-MS/MS in HIB 1B adipocytes (500 µg nuclear extracts) were first sorted according to annotation as transcription factors (MatBase tool, Genomatix). Then, the proteins were ranked according to the fold change, *P*-value and high accuracy (number of identified peptides for quantification > 2, Mascot percolator score > 13, FDR < 1%, see Methods) (data not shown). Two proteins were selected as putative allele-specific binding proteins for follow -up studies, i.e. the transcription factor TF1 and the homeodomain transcription factor, paired related homeobox 1 (PRRX1). In HIB 1B adipocytes, TF1 was identified with significant fold change (4.5 at C/T allele), *P*-value (4.5×10^{-4}) with 11 unique peptides at the rs4684847, as shown in Table 19. TF1 was also detected in 3T3-L1 adipocytes and pre-adipocytes, however with no significant fold change (0.6-0.9 at C/T allele) and only 3 unique peptides, respectively (125 µg nuclear extracts). These results indicated that the intensities of peptides identified by LC-MS/MS are to some extent proportional to the protein input amount, strengthening the previous observation (see above). Next, homeobox TFs are reported to be involved in embryonic and tissue developmental processes [391–393]. As mentioned above (see chapter 7.2.1.1), among the six predicted *cis*-regulatory variants at the *PPARG* locus, only the rs4684847 showed direct overlap to a homeobox TFBS matrix, which is a T2D-specific feature inferred from PMCA. In contrast, the other five predicted *cis*-regulatory variants showed no overlap of homeobox TFBS [98]. Indeed, only PRRX1 as a homeobox transcription factor was detected in MS data for both 3T3-L1 and HIB 1B cells, although it was detected under the criteria of non-significant fold change (1.0-1.3 at C/T allele), as shown in Table 19. To increase the enrichment and improve the fold change in DNA-binding affinity to PRRX1, and also find other possible protein candidates, further affinity chromatography was performed with increased amount of nuclear extracts.

Cell line	Gene symbol	Allelic fold change (C/T)[a]	*P-value*[b]	Peptide count for quantitation[c]	Mascot score[d]
3T3-L1 (d0)	TF1	0.9	-	3	99
	PRRX1	1.0	-	1	61
3T3-L1 (d15)	TF1	0.6	-	3	99
	PRRX1	1.3	-	1	61
HIB 1B (d9)	TF1	4.5	4.5×10^{-4}	11	707
	PRRX1	1.0	0.9	2	166

Table 19. Candidates for allele-specific binding proteins at the rs4684847 based on proteome analysis on a small scale.

[a]fold change was calculated as the mean ratio of normalized proteins abundance over the three experiments, [b]*P*-values were derived from unpaired *t*-tests, [c]Peptide count for quantitation refers to the number of peptides

uniquely assigned to one protein and therefore used for quantitation, [d]Mascot score is built as summed up single probability of identified peptides per protein and serves as indicator for the reliability of protein identification. LC-MS/MS data were quantitative analyzed by *Progenesis*.

7.2.1.2.2 Large scale affinity purification

Development and optimization of protein purification were first performed on a small scale (see above). After optimization of the protein purification, the process was scaled up to enhance the productivity of protein separations and subsequently improve the identification of proteins in samples by MS. The direct scale-up of the protein purification was achieved by transferring optimized conditions on a small-scale to a large production scale [394]. The *PPARG* gene is expressed abundantly and equally in white fat and brown fat, and is required as a master regulator for the development of both, white and brown adipocytes (reviewed in Ohno et al. 2012 [395]). Here, HIB 1B cell line was used as the source for further affinity chromatography that can be easily differentiated to high density and also provide advantage of relative short-term differentiation (9 days) compared to 3T3-L1 cell line (15 days). In the previous protein purification (see chapter 7.2.1.2.1), 500 µg nuclear extracts from HIB 1B adipocytes were used. A major problem of the previous purification was the low abundance of the proteins in eluates, which could be below the borderline for detection by LC-MS/MS. As the previous results indicated, simple and commonly method used for improving enrichment of proteins in affinity chromatography is to use increased amount of proteins. Hence, in further purification experiments on a larger scale 14 fold more amount of nuclear extracts (500 µg → 7 mg) were used in order to increase the enrichment of proteins of interest. According to the increased amount of nuclear extracts, the amount of biotinylated oligonucleotides (26 → 77 pmol) and magnetic beads (250 → 500 µg of Dynabeads® M-280 beads corresponding to the binding capacity of 100 pmol biotinylated oligonucleotides) was also increased. Previously, the washing with 50 mM NaCl resulted in faint EMSA signal of the allele-specific binding proteins at the C allele, suggesting that the target proteins could be eluted early in washing steps. To eliminate this possibility, the reaction solution was washed with 10 mM NaCL instead of 50 mM NaCl. In addition, the binding buffer used in EMSA (Fig. 5) contained 250 mM NaCl. According to the salt concentration in the EMSA binding buffer, the elution was here started with 200 mM NaCl instead of with 400 mM NaCl.

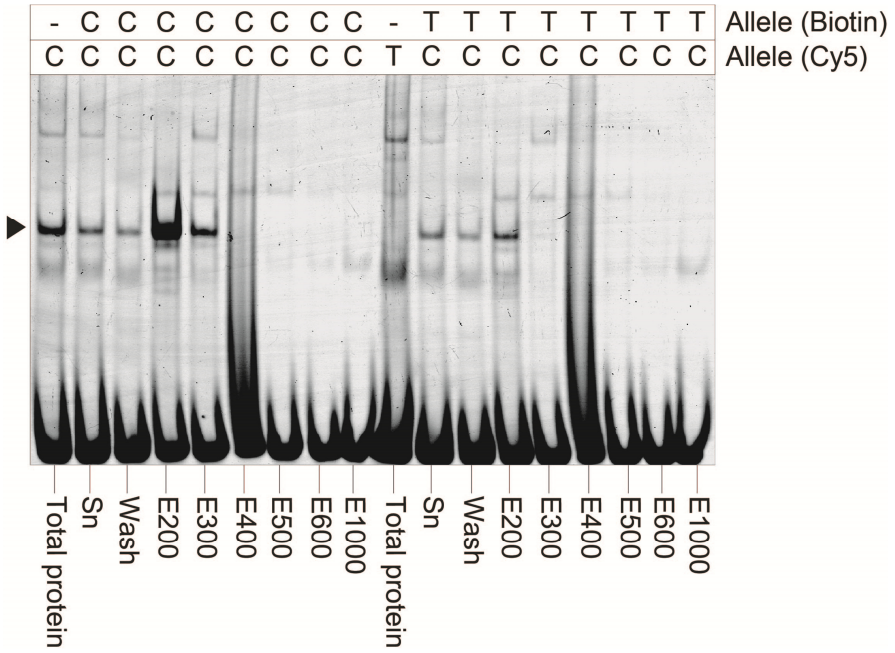


Figure 9. EMSA analysis of eluates obtained during affinity purification process for the rs4684847 using nuclear extracts from HIB 1B adipocytes on a large scale.

Nuclear extracts prepared from HIB 1B adipocytes were subjected to affinity purification as depicted in Methods (see 5.2.17). The initial nuclear extracts and specified chromatographic fractions (wash and eluates) were analyzed to confirm the enrichment of the allele-specific binding proteins using C allelic Cy5-labeled probe of the rs4684847. A black arrow indicates the allele-specific band. *Total protein:* nuclear extracts from HIB 1B adipocytes, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (10 mM). *E200-1000*: Elution of proteins with increasing concentration of NaCl (200 -1000 mM NaCl, respectively). Affinity chromatography using magnetic beads was performed using 7 mg from HIB 1B adipocytes for each allele, respectively. All experiments were performed in triplicates.

The affinity purification process was monitored by EMSA using the rs4684847 C allele as a Cy5-labeled probe, which showed more intensive allelic difference. There was obviously elevated enrichment in the allele-specific binding proteins when comparing the intensities for E200 eluate (like E300 eluate) from the purification on a large scale (Fig. 9) with E400 eluate from the previous purification on a small scale (Fig. 7C). Simultaneously, the allelic difference was seen more clearly in both eluates, E200 and E300 from the large scale purification which were analyzed by LC-MS/MS. In total, 828 proteins were identified by LC-MS/MS analysis for large-scale whereas a total of 667 proteins were identified for small-scale (Table 18). When only comparing the number of identified proteins, the use of 14 fold increased amount of protein extracts did not lead to 14-fold increase in number of proteins

identified by LC-MS/MS. However, TF1 was detected with an obviously improved fold change (4.5 → 8.0 at the C > T allele for 500µg → 7 mg protein nuclear extracts, respectively), further supporting TF1 as a candidate protein for the allele-specific binding at the rs4684847. PRRX1 was also detected with apparently increased fold change (1.0 → 2.6 at the C > T allele) and significant *P*-value (0.9 → 1.2 x 10^{-2}) as expected (Table 20). Thus, both TF1 and PRRX1 were included in further analysis as putative allele-specific binding proteins.

Cell line	NE amount applied for AC	Gene symbol	allelic fold change (C/T)[a]	*P-value*[b]	Peptide count for quantitation[c]	Mascot score[d]
HIB 1B	500 µg	TF1	4.5	4.5 x 10^{-4}	11	707
		PRRX1	1.0	0.9	2	166
	7 mg	TF1	8.0	1.4 x 10^{-3}	13	1316
		PRRX1	2.6	1.2 x 10^{-2}	5	752

Table 20. Candidates for allele-specific binding proteins at the rs4684847 based on proteome analysis on a large scale.

[a]fold change was calculated as the mean ratio of normalized proteins abundance over the three experiments, [b]*P*-values were derived from unpaired *t*-tests, [c]Peptide count for quantitation refers to the number of peptides uniquely assigned to one protein and therefore used for quantitation, [d]Mascot score is built as summed up single probability of identified peptides per protein and serves as indicator for the reliability of protein identification. LC-MS/MS data were quantitative analyzed by *Progenesis*. NE: nuclear extracts, AC: affinity chromatography.

7.2.1.3 Validation of PRRX1 and TF1 as allele-specific binding proteins using competition and supershift EMSA

Based on LC-MS/MS results, PRRX1 and TF1 were identified with significant fold change and *P*-value, as described in Table 20. To determine whether PRRX1 and TF1 bind at the rs4684847 in an allele-specific manner, competition EMSA and supershift were performed. First, a possible effect of the rs46848471 on transcription factor binding sites was evaluated by *in silico* analysis using MatBase (Genomatix, Munich, Germany). Based on the analysis, the rs4684847 does not overlap to the core of the predicted binding site for PRRX1. However, PRRX1 was predicted to bind in close proximity to the C/T allele of the rs4684847 (Fig. 10A). To determine the allele-specific binding of PRRX1 at the rs4684847 adjacent DNA sequence, two common approaches were used: competition with unlabeled (or cold) competitor DNA and antibody supershift. Competition and supershift assays using nuclear extracts from HIB 1B, 3T3-L1 adipocytes and human SGBS adipocytes did not provide any evidence for PRRX1 as an allele-specific binding protein to the rs4684847 (data not shown), which could be due to its low abundance in nuclear extracts and the quality / specificity of the

available antibodies. For this reason, competition and supershift assays for PRRX1 were performed using whole protein extracts isolated from 293T cells overexpressing Flag-PRRX1. As a negative control, empty vector containing the same backbone of Flag-PRRX1vector was used to control the expression of Flag-PRRX1 in 293T cells. To confirm the expression of Flag-PRRX1 protein, Western blot was performed using anti-PRRX1 polyclonal antibody from rabbit (Fig. 10B). The transfection with the Flag-PRRX1 vector resulted in a successful expression of Flag-PRRX1 protein comparing to control with the empty vector (Fig. 10B).

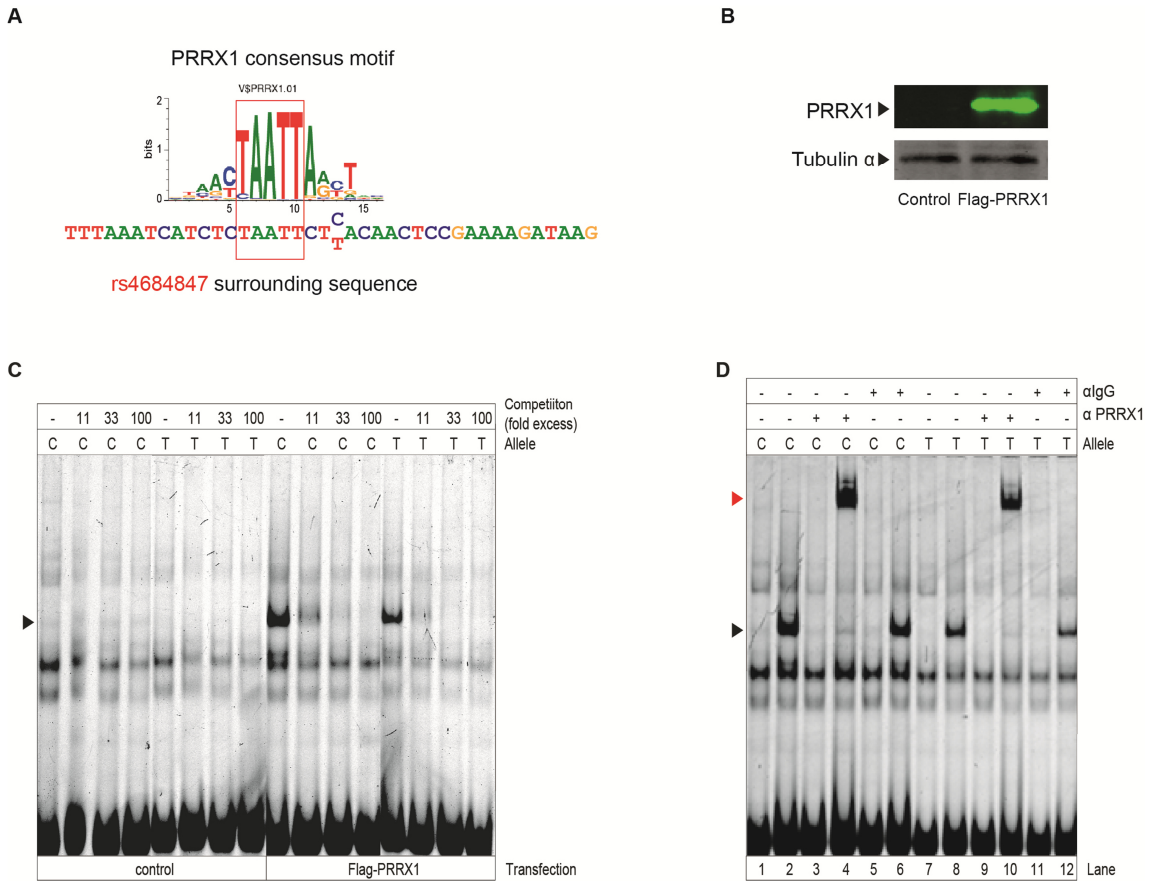


Figure 10. Allele-specific binding of the transcription factor PRRX1 at the C risk allele of the rs4684847.

(A) *In silico* analysis indicates that the C risk allele at the rs4684847 does not overlap the consensus binding site for the transcription factor PRRX1, but in close proximity of the rs4684847. (B) Overexpression of PRRX1 was evaluated by western blot analysis. Whole extracts from 293T cells overexpressing Flag-PRRX1 were incubated for 1h with anti-PPRX1 antibody (1:5000) and anti-tubulin α as internal control (1: 5000), respectively. (C-D) Competition and supershift EMSAs were performed to prove PRRX1 as an allele-specific binding protein to the rs4684847. Competition assay

was performed with excess of cold PPRX1 probe (11, 33 and 100 fold) using whole extracts from 293T overexpressing empty vector (control) or Flag-PRRX1 (C). For supershift, 1 µl of anti-PRRX1 antibody (M. Kern) or normal lgG isotope (Santa cruz) was pre-incubated with whole extracts from 293T cells overexpressing empty vector (control) or Flag-PRRX1 for 1 h at 4°C before adding Cy5-labeled probe (D). After addition of cold PRRX1 probe, the allele-specific band was disappeared. For supershift, the allele-specific band was migrated to the position with red arrow. Lane 1, 3, 5, 7, 9 and 11: whole extracts from 293T overexpressing empty vector (control) or lane 2, 4, 6, 8 and 10: whole proteins extracts from 293T overexpressing Flag-PRRX1. A black arrow indicates the allele-specific band difference. A red arrow indicates the supershift band.

Following the western blot analysis, competition and supershift assays were performed to prove the allele-specific binding of PRRX1 to the rs4684847 (Fig. 10C and Fig. 10D). First, overexpression of PRRX1 protein in 293T cells was also confirmed in EMSA by comparing to control vector (Fig. 10C), which was consistent with western blot result (Fig. 10B). The allele-specific band showed more intensive signal at the C allele, although it also appeared obviously at the T allele. Next, the allele-specific binding of PRXX1 was confirmed using 11, 33, 100 fold excess concentration of unlabeled PRRX1 consensus sequence. The allele-specific band was completely abrogated by the competition with 100 fold excess of the PRRX1 cold probes at both alleles during the non-specific bands seen in the same lanes seemed to be not dramatically affected (Fig. 10C, right panel). To further confirm the allele-specific binding of PRXX1 to the rs4684847, supershift assay was performed using antibody specific to PRRX1. Non-specific IgG polyclonal isotype was used as a negative control. The presence of the PRRX1 antibody resulted not only in disappearance of the allele-specific band, but also in an additional shifted band at both alleles (Figure 10D, lane 4 and 10). As a control, no supershift was apparent when an IgG antibody was used instead of the antibody against PRRX1 (Figure 10D, lane 6 and 12). Comparing the signal intensity of the supershift band at the C allele versus the T allele, more intensive signal was observed at the C allele than at the T allele (Figure 10D, lane 4 and 10). The competition and supershift EMSA results (presented in this thesis) verified PRRX1 as an allele-specific binding protein to the rs4684847 at the *PPARG* locus [98]. Further analyses demonstrated a novel activity of PRRX1 as a repressor of PPARγ2 expression and showed its adverse effect on lipid metabolism and insulin sensitivity in co-work with other colleagues [98].

Although PRRX1 was identified as an allele specific binding protein at the rs4684847, it is likely that PRRX1 is not the only one binding to the rs4684847 in an allele-specific manner.

As mentioned above, TF1 identified with significant fold change and *P*-value (Table 20) was considered studying further. An *in silico* analysis using MatBase (Genomatix, Munich, Germany) revealed that the rs4684847 alters the core of the predicted binding site for TF1, with the C allele creating the site and the T allele abrogating it (data not shown, not yet published). Similarly to the previous case, the competition and supershift EMSAs were performed to examine TF1 as an allele-specific binding protein to the rs4684847. To determine experimentally whether TF1 binds differentially to the rs4684847 C/T allele, the nuclear extracts from HIB 1B and 3T3-L1 adipocytes were used in competition and supershift EMSAs (Fig. 11). First, the protein binding in EMSA was competed with biotinylated probes to verify the allele-specific binding of proteins to the rs4684847. As shown in Fig. 11A, the allele-specific binding of proteins to the rs4684847 was readily blocked by a 1-fold molar excess of biotinylated probe containing the C allele while the band was blocked first by a 33-fold molar excess of biotinylated containing the T allele in HIB 1B adipocytes. Although there was a slight decrease in the signal intensity of the allele-specific band at the T allele, competition with a 100-fold molar excess of biotinylated probe containing the T allele did not completely abrogate it. To further confirm this, the same experiment was performed using nuclear extracts from 3T3-L1 adipocytes (Fig. 11C), showing a similar result from HIB 1B adipocytes (Fig. 11A). Next, in order to confirm the specificity of TF1 binding at the rs4684847, competition EMSA was performed with consensus oligonucleotides for SP1 (Sp/KLF family of transcription factor), MyoD (myogenic regulatory factors), CdxA (chicken homeodomain protein) and a non-specific scrambled control. Competition with a 33- and a 100-fold molar excess of those competitors did not affect the allele-specific binding of proteins whereas it was completely abolished by the addition of a 100-fold molar excess of biotinylated probe containing the C allele or TF1 consensus sequence (Fig. 11B). In supershift using anti-TF1 antibody the allele-specific band was disappeared and formed an additional shift band while the specific binding was unaffected in the presence of the isotope control antibody (Fig. 11D). Taken together, these results revealed that TF1 selectively binds to the rs4684847 in a genotype specific manner. Thus, future experiments would be necessary to uncover a *cis*-regulatory effect of TF1 on the *PPARG* gene expression and to demonstrate its biological relevance for T2D.

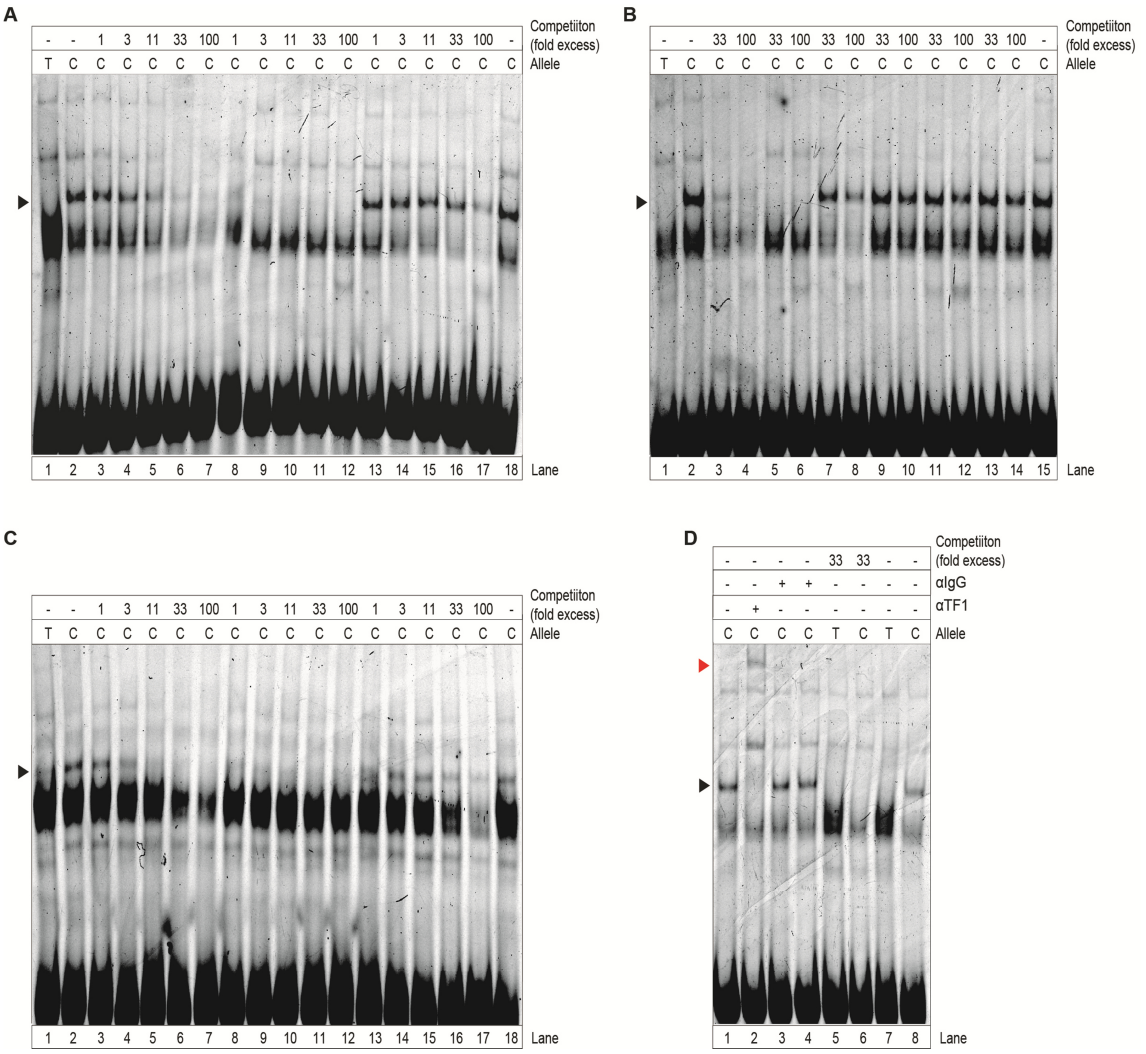


Figure 11. Allele-specific binding of the transcription factor TF1 at the C risk allele of the rs4684847.

Competition and supershift EMSAs indicated TF1 as an allele-specific binding protein to the rs4684847 C allele. (A) Competition assay was performed with a 1-, 3-, 11-, 33- and 100-fold molar excess of biotinylated oligonucleotides containing the C allele (lane 3-7), the T allele (lane 13-17) and unlabeled TF1 consensus probe (lane 8-12) using nuclear extracts from HIB 1B adipocytes. Lane 1, 2 and 18: without competition. (B) In competition EMSA assay, Cy5-labeled oligonucleotide probe of the rs4684847 was competed with a 33-, 100- fold molar-excess of biotinylated probes containing the C allele (lane 3-4), unlabeled consensus probes of TF1 (lane 5-6), SP1 (lane 7-8), MyoD (lane 9-10), CdxA (lane 11-12) and a non-specific scramble (lane 13-14), as indicated. Lane 1, 2 and 15: without competition. (C) Competition assay was performed with a 1-, 3-, 11-, 33- and 100-fold molar excess of biotinylated probes containing the C allele (lane 3-7), the T allele (lane 13-17) and unlabeled TF1 consensus probe (lane 8-12) using nuclear extracts from 3T3-L1 adipocytes. Lane 1, 2 and 18: without competition. (D) For superhsift, anti-TF1 antibody from rabbit (B. Kempkes, lane 2) and normal lgG

isotope from rat (Santa cruz, lane 3-4) was pre-incubated with nuclear extracts from HIB 1B adipocytes for 1 h at 4°C prior to adding Cy5-labeled probe. As control for the TF1-antibody supershift, for which concentration was unknown, normal IgG isotype was used in different concentrations (lane 3, 40 ng; lane 4, 200 ng). Competition assay was performed with a 33-fold molar excess of biotinylated probes containing the T allele (lane 5) or the C allele (lane 6). A black arrow indicates the allele-specific differences. A red arrow indicates the supershift band.

7.2.2 *FTO* locus

7.2.2.1 Optimization of conditions for allele-specific binding of proteins

The *FTO* gene is well known as the most robust and significant genetic contributor to obesity [396,397]. Since 2007, GWAS have revealed strong associations of *FTO* variants with BMI and risk of obesity, and subsequently T2D in multiple populations [107–109,164,198–202,205,206]. The rs1421085 located in the first intron of the *FTO* is in strong linkage disequilibrium (pairwise $r^2 > 0.97$) with the rs9939609, showing the strongest association with BMI in several studies. In a few studies was shown that the rs1421085 C risk allele was associated with increased body weight in different populations [199,204]. However, the molecular and pathophysiological mechanisms by which the rs1421085 might impact on weight gain, remain to be fully understood. Recently, two independent studies, the fine-mapping study [376] and the PMCA analysis [98] predicted the rs1421085 at the *FTO* locus as *cis*-regulatory. The following EMSA and reporter gene assays confirmed the allele-specific binding of proteins at the rs1421085 [98]. Based on these studies, one predicted *cis*-regulatory variant (rs1421085) was chosen for further study, showing the most significant allelic difference in EMSA experiments [98]. First, the previous EMSA result was confirmed using nuclear extracts from the same cell line, 293T under certain EMSA condition. The *FTO* gene is expressed in a variety of tissues relevant to metabolic diseases including adipose tissue and skeletal muscle with highest gene expression in hypothalamus [107,188,239]. The knowledge of allele-specific binding of proteins would be important in generating a more thorough understanding of how the *FTO* gene is regulated by *cis*-regulatory variants in specific cell or tissue types, and moreover its functional roles. Thus, in addition to nuclear extracts from 293T cells, extracts from adult mouse brain tissue, Huh7 cells, 3T3-L1 adipocytes and INS-1 cells were included for this purpose. In case of the rs1421085, the 62 bp Cy5-labeled probe surrounding the rs1421085 (T/C allele at position 42, (+) strand) was used in EMSA experiment for the reason that *cis*-regulatory module might be not recognized within 40 bp sequence unlike the rs4684847 at the *PPARG* locus [98]. As

indicated in Fig. 12A, the rs1421085 altered allele-specific binding of proteins to a *cis*-regulatory element. The upper allele-specific shift band (marked with number 1) was observed when incubated with nuclear extracts from 293T cells, adult mouse brain tissue and Huh7 cells. The upper allele-specific band (marked with number 1) was found in the presence of the C allele, but not with the T allele, suggesting that this is a C allele-specific complex. Otherwise, the lower band (marked with number 2) appeared at the T allele with intensiver signal compared to the C allele when incubated with nuclear extracts from 293T cells. Similar protein-DNA migration patterns were also obtained when using nuclear extracts from adult mouse brain tissue as well other cell lines analyzed (Fig. 12A). An obvious difference in signal intensity between both alleles was seen, with the T allele showing higher capacity for the protein binding compared to the C allele (Fig. 12A). Since binding conditions are specific to each protein-DNA interaction, appropriate binding reaction conditions should be established also for the rs1421085 at the *FTO* locus. First, to optimize amount of nuclear extracts in EMSA experiment, amount of nuclear extracts from 293T cells was titrated in the range from 156 ng to 10 μg using fixed oligonucleotide concentration (50 fM). As shown in Fig. 12B, decreasing amount of nuclear extracts led to the gradual diminution of the specific band. Additionally, a large excess of extracts resulted in high background signal, non-specific and smeared bands. The amount of the nuclear extracts (5 μg) resulted in the most clear specific band as darker than that found with less background and non-specific signals, which was consistent with those in other cell lines and tissue such as Huh7, INS-1 and mouse brain (data not shown). This ratio of protein/nucleic acid (5 μg/25 fM) was then included in all subsequent EMSA experiments and eventually adapted to protein purification.

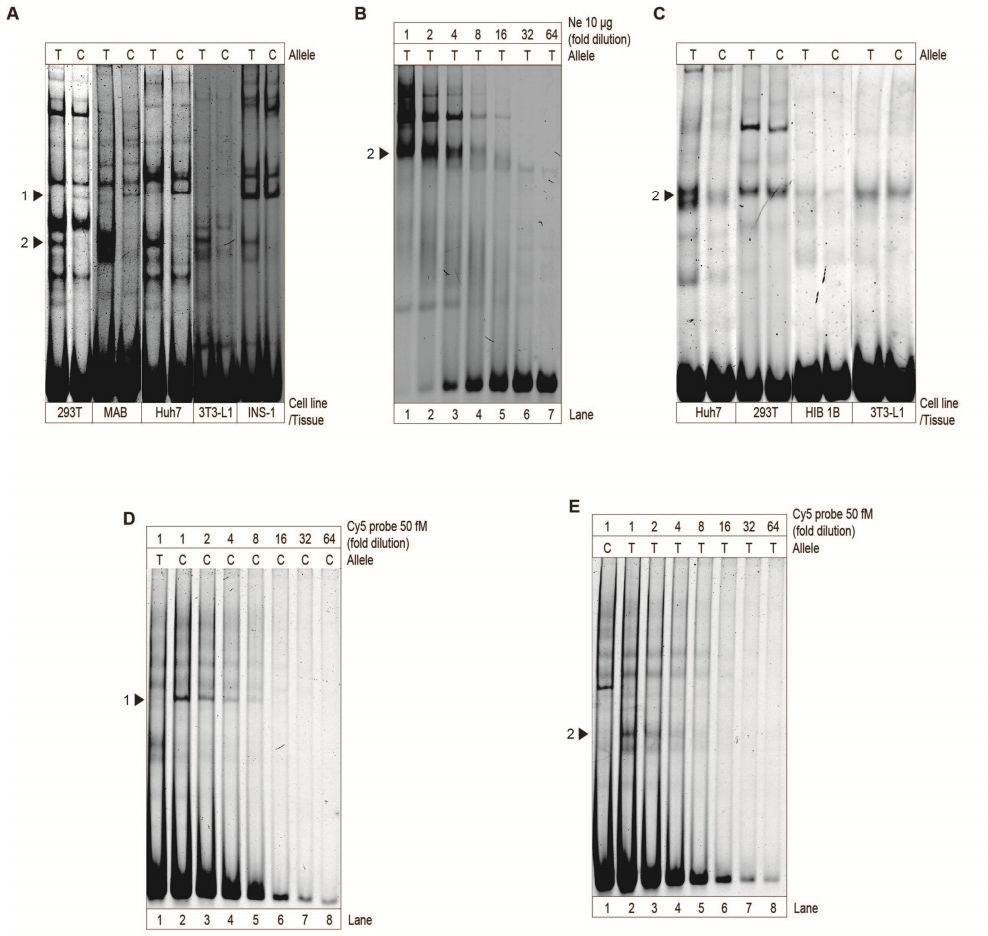


Figure 12. Analysis of allele-specific protein-DNA interaction at the predicted *cis*-regulatory variant, rs1421085 of the *FTO* locus.

To examine the allele-specific binding of proteins at the rs1421085 of the *FTO* locus, EMSA experiments were performed using allelic Cy5-labeled probes for the rs1421085 and nuclear extracts from different cell lines and tissue. (A) The predicted *cis*-regulatory variant, rs1421085 showed two allele-specific bands (marked with number 1 at the C allele and number 2 at the T allele) in 293T cells, mouse adult brain tissue (MAB), Huh7 cells, 3T3-L1 adipocytes (d15) and INS-1 cells. (B) Cy5-labeled T allele probe for the rs1421085 was incubated with decreasing amounts of nuclear extracts (in the range from 0.16 µg to 10 µg). Reduced amount of nuclear extract proteins resulted in decreased signal intensity, thereby increasing resolution of the second allele-specific band (marked with number 2) at the certain DNA concentration in 293T cells. (C) EMSA experiment with shortened length of Cy5-labeled probes (36 bp) was performed to examine dependence of the allele-specific binding properties of proteins on the length of oligonucleotides in Huh7 cells, 293T cells, 3T3-L1 adipocytes (d15) and HIB 1B adipocytes (d9). (D-E) To determine its optimal concentration (in the range from 0.78 pM to 50 fM), an oligonucleotide titration was performed using fixed amount of nuclear extracts

from MAB (10 µg/well) for the C allele (D) and the T allele (E), respectively. A black arrow indicates each allele-specific band.

Likewise, an *in silico* analysis was performed to assess the potential functional impact of the rs1421085 on predicted TFBSs using MatInspector (Genomatix, Munich, Germany). In addition, the putative bindings of TFs were also compared between 36 bp and 62 bp surrounding the rs1421085 to observe the improvement of the specific protein binding and the reduction of non-allele-specific signal depending on its length. As expected, the rs1421085 might alter the recognition / binding motifs of several TFs. Moreover, a few TFs were predicted to bind solely to the 62 bp compared to the 36 bp sequence (data not shown, not yet published), which could exert a determinative influence on transactivation of the target gene with neighbouring TFs in regulatory networks [382]. To characterize and compare the binding patterns of protein to the rs1421085, Cy5-labeled 36 bp oligonucleotides (T/C allele at position 16, (+) strand) were additionally generated for EMSA experiment. EMSA experiments were performed using 62 bp and 36 bp on different cell lines under the same conditions. EMSAs with the 36 bp probes demonstrated that non-allele-specific signal was reduced compared to using the 62 bp probes, however, the first allele-specific band (marked with number 1) disappeared in all cell lines analyzed. Of note, the other allele-specific binding (marked with number 2) was observed only for the Huh7 nuclear extracts in EMSA with the 36 bp probes (Fig. 12C), indicating that the shortened oligonucleotides led to the loss of specificity for the predicted motifs within certain cell lines. Based on the results, the use of 36-bp oligonucleotides was excluded for further analysis.

Next, to optimize the allele-specific binding of proteins to DNA, optimal concentration of oligonucleotides was determined in range from 0.78 to 50 fmol with excess amount of nuclear extracts (10 µg) from adult mouse brain tissue (Fig. 12D-E) as well 293T and Huh7 cells by EMSA assays (data not shown), respectively. Since one allele-specific band were seen at each allele (nr. 1: C allele; nr. 2: T allele), EMSA experiments were performed for each allele, respectively. As shown in Fig. 12D-E, the signal intensity of the allele-specific band diminished gradually with decreasing concentration of oligonucleotides, and non-allele-specific signals were also reduced, suggesting that the measured signal is to some extent proportional to the DNA concentration. Additionally, the concentration of poly (dI-dC) (3.5 µg/ml) was chosen for further EMSAs. It was the highest concentration that resulted not only in a dark allele-specific band as that found with no poly (dI-dC), but also obviously less background signals (data not shown). This concentration of poly (dI-dC) was adapted with

modification to affinity purification. The optimal reaction conditions for affinity purification were empirically determined by the best ratio of extracts amount to DNA concentration based on the EMSA results (Fig. 12). Finally, on the basis of the EMSA results and biological relevance to the function of *FTO* gene in brain, skeletal muscle and liver [110,188,190,191], mouse adult brain tissue, 293T and Huh7 cell lines were chosen for further analysis.

7.2.2.2 Affinity chromatography coupled to mass spectrometry: optimization

7.2.2.2.1 Small scale affinity purification using magnetic beads

First, streptavidin-biotin affinity chromatography system was used to isolate and enrich the allele-specific binding proteins. Initially, the purification was stared with nuclear extracts from 293T cell line for the reasons that this cell line can be obtained relative easily in large amount and showed a thick allele-specific band in the previous EMSA experiment (Fig. 12A). For the affinity purification, 26 pmol of each biotinylated oligonucleotide containing the identical sequence to the Cy5-labeled probes of the rs1421085 was coupled to the 250 µg of Dynabeads® M-280 beads (corresponding to the binding capacity of 50 pmol biotinylated oligonucleotides). All procedure was performed in the same manner as that used for the *PPARG* locus. The reaction solution contained 125 µg nuclear extracts from 293T cells and 0.01 % (v/v) detergent CHAPS for each allele. Initially, 2 µg/ml of poly (dI-dC) was added to the reaction solution. After incubation of beads with the reaction solution, the beads were washed three times with 1x binding buffer containing 50 mM NaCl by magnetic separation. Subsequently, the bound protein complexes were eluted by resuspending beads in elution buffer containing NaCl from 400 mM to 1000 mM. The steps along the affinity purification process were monitored by EMSA using the Cy5-labeled probes (containing T/C allele, respectively). To ensure the enrichment of the specific proteins, EMSA was first performed using the Cy5-labeled T allele probe (Fig. 13). As shown in Fig. 13A, very faint signal specific for the rs1421085 T allele was observed while the C allele exhibited almost no signal in eluate E400. Unfortunately, no visible allele-specific signals were observed in EMSA when using the Cy5-labeled C allele probe (data not shown). Subsequently, proteins from all samples of affinity purification were stained with silver nitrate (see Methods). Consistent with the EMSA results (Fig. 13A), the silver staining gel revealed an obvious allele-specific band in whole samples from *total protein* to eluates *E400-1000* (Fig. 13B). For all samples, especially eluate E1000, the major allele-specific band appeared obviously more intensive at

the T allele compared to the C allele with an estimated molecular mass between 50 and 75 kDa, which was different from EMSA results showing allelic difference only in eluates E400 (Fig. 13A). Finally, eluates E400 for each allele were subjected to LC-MS/MS analysis for protein identification.

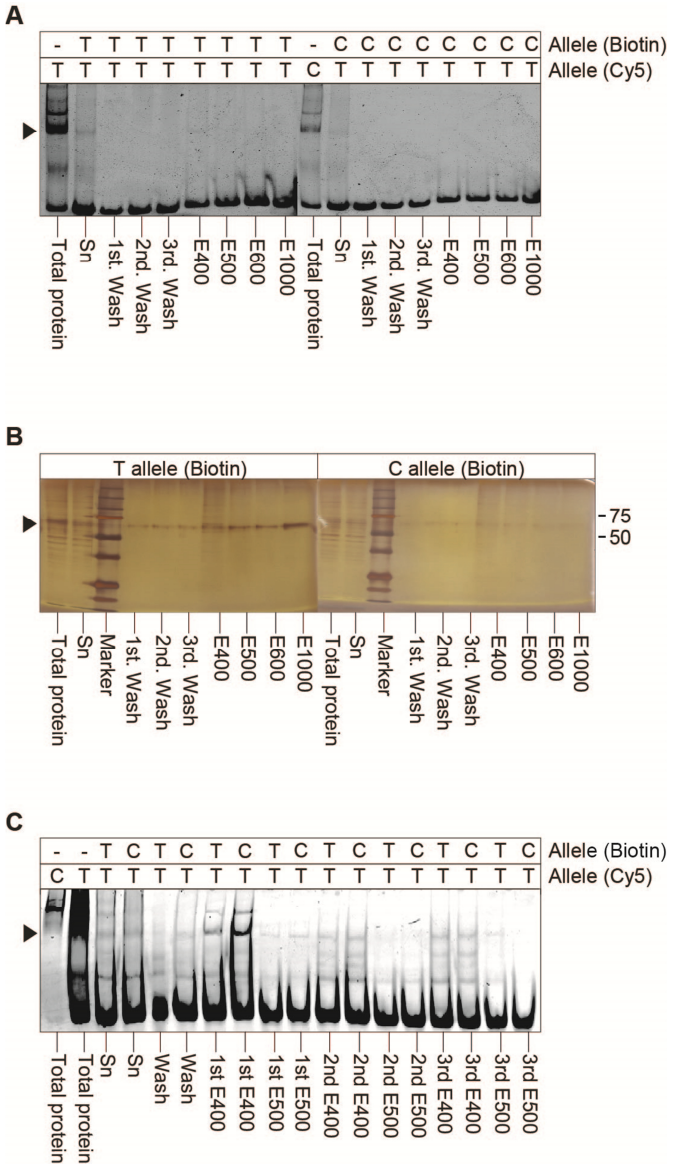


Figure 13. EMSA analysis of eluates obtained during affinity purification process for the rs1421085 using nuclear extracts from 293T cells.

Nuclear extracts prepared from 293T cells were subjected to affinity purification as depicted in Methods (see 5.2.17). The initial nuclear extracts and specified chromatographic fractions (wash and eluates) were analyzed to confirm the enrichment of the allele-specific binding proteins using the T

allelic Cy5-labeled probe of the rs1421085. Affinity chromatography using magnetic beads was performed using 125 µg (A and B) and 500 µg (C) nuclear extracts from 293T cells for each allele, respectively. (A) 2 µg/ml of poly (dI-dC) was added to the reaction. *Total protein:* 293T nuclear extracts, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (50 mM). *E400-1000*: Elution of proteins with increasing concentration of NaCl (400 -1000 mM NaCl, respectively). (B) Silver staining of eluates during affinity purification. The allele-specific binding protein was indicated between 50 and 75 kDa. (C) Poly (dI-dC) was used in the range 2-20 µg/ml (1st: 2, 2nd: 10, 3rd: 20 µg/ml). *Total protein*: 293T nuclear extracts, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (50 mM). *E400-1000*: Elution of proteins with increasing concentration of NaCl (400 -1000 mM NaCl, respectively). However, only eluates E400 and E500 were shown. A black arrow indicates the allele-specific band. All experiments were performed in triplicates.

Optimization of purification conditions was required to improve the enrichment of the proteins of interest in the eluates. The concentration of DNA and poly (dI-dC), and the amount of protein were considered as important factors for the optimization, as tested in the previous chapter. High concentration of DNA could result in increased undesirable non-allele-specific binding of proteins, causing possible inhibition of the specific protein binding to the target DNA. Conversely, the use of relative low concentration could enhance the selective binding of the protein of interest [342]. Moxley and his colleagues used in their study 1.34 nM DNA for 110 µg nuclear extract proteins to isolate C/EBP protein [342], which was apparently lower compared to the condition used in this study. For that reason, in further affinity purification the concentration of biotinylated oligonucleotides was varied over a range from 2 to 10 pmol with the same amount of nuclear extracts (125 µg), which was less than a half of that used previously (26 pmol). However, the affinity purification using relatively low concentration of biotinylated oligonucleotides did not achieve satisfactory enrichment of proteins as there was no detectable signal for any protein binding (data not shown), indicating the DNA concentration might be not enough for the efficient enrichment of the allele-specific binding proteins. Therefore, in next affinity purification, elevated amount of nuclear extracts from 293T cells (125 µg → 500 µg) was used with the appropriate concentration of biotinylated oligonucleotides (26 → 52 pmol). In addition, different concentrations of poly (dI-dC) were tested to observe the effect of competitor on the enrichment of target proteins during purification. Poly (dI-dC) was added to each reaction solution in the range 2-20 µg/ml (1^{st}: 2, 2^{nd}: 10, 3^{rd}: 20 µg/ml) followed by incubation for 20 min at 4 °C. After wash steps (three times in 1x binding buffer containing 50 mM NaCl), the proteins bound to the magnetic

beads were eluted by increasing concentration of NaCl (in range from 400 to 1000 mM NaCl). Following the purification procedure, EMSA was first performed using all samples collected with the Cy5-labeled T allele probe (Fig. 13A). The second allele-specific band (marked with number 2) (Fig. 12) appeared in most eluates, as assessed by EMSA, although abnormal migration of the band (smear-like signals or no clear band, or vertical streaks along the edges of the lanes) was observed in a few lanes including *Total protein* and eluate *1st E400*. However, there was no obvious difference in signal intensity between both alleles except of eluate *3rd E500* (Fig. 13C). Only the eluate *3rd E500* showed the differential allele-specific enrichment of proteins, with a stronger binding to the T allele than the C allele. Furthermore, poly (dI-dC) in range from 2 to 10 µg/ml showed slightly diminishing effect on the signal intensity of the allele-specific band. In two independent protein purifications using 10 and 20 µg/ml of poly (dI-dC), similar results were obtained. Thus, the concentration of poly (dI-dC) (10 µg/ml) was used in subsequent affinity purification for 293T cells for economic purpose. However, it still failed to enrich the specific proteins (Fig. 13C). Therefore, the samples from these purifications were excluded for further analysis by LC-MS/MS.

7.2.2.2.2 Small scale affinity purification using sepharose beads

The previous affinity purification using magnetic beads on a small scale failed to efficiently enrich the allele-specific binding proteins in eluates (Fig. 13). For further development of efficient purification method, other affinity chromatography was performed using sepharose. Sepharose is a cross-linked, beaded-form of agarose gel and the most widely used matrix since it is chemically versatile, making possible the stable attachment of ligands for purification of different enzymes, antibodies, and other proteins and peptides through the hydroxyl groups on the sugar residues. Due to the versatility and high mechanical stability of sepharose, its use has been greatly expanded in the number of potential applications i.e as an excellent matrix for high performance chromatographic procedures in affinity chromatography, ion exchange chromatography, and other modes of separation. Sepharose is often used in combination with chemistry, enzymes, antibodies, other proteins and peptides. Especially, heparin-sepharose is a widely used, successful and well-documented technique in affinity chromatography [342,398–400]. Heparin is a highly sulphated glycosaminoglycan and has widespread use as a general affinity ligand. Its high degree of sulfation mediates a strong acidic nature to the molecule, being able to bind to many substances by ionic interaction. Additionally, heparin contains unique carbohydrate sequences, acting as specific binding sites

for some proteins. Thus, heparin-sepharose has been extensively used for the purification of enzymes, coagulation proteins, steroid receptors and protein synthesis factors [399–401].

First, the oligonucleotides were modified with $(AC)^5$ as a linker and then coupled with sepharose beads (Fig. 14A). In total, 0.9 μmol of oligonucleotides for each allele was used, which was much higher than that for magnetic bead-based purification (52 pmol). Like magnetic bead-based purification, nuclear extracts from 293T cells were used also for sepharose bead-based purification. 500 μg of nuclear extracts for each allele was diluted in heparin affinity buffer without salt (HA-0), facilitating the binding of heparin to proteins. The protein mixture was subjected to the column containing sepharose beads to enrich the allele-specific binding proteins. After washing with HA buffer (100 mM KCl), proteins were eluted with increasing concentration of KCl (200-1000 mM KCl). The elution fractions were collected in 14 fractions. All flow through, wash and eluates (Fraction 1-14) were collected and analyzed by EMSA in order to monitor the enrichment of the allele-specific binding proteins during the procedure. EMSA was initially performed using the Cy5-labeled T allele as a probe. Faint, but visible bands were detected by EMSA in the three eluted fractions (Fraction 4-6) for the T allele (Fig. 14B), being the second allele-specific band (marked with number 2) (Fig. 12). In addition, most of the allelic binding activity appeared in the eluate fractions 4 to 6 at around 400-500 mM KCl, which was consistent with the results from magnetic beads (Fig. 13). However, no allele-specific binding appeared in the same fractions for the C allele (Fig. 14C). Finally, the fractions from 4 to 6 were pooled for each allele, respectively and subsequently analyzed by LC-MS/MS.

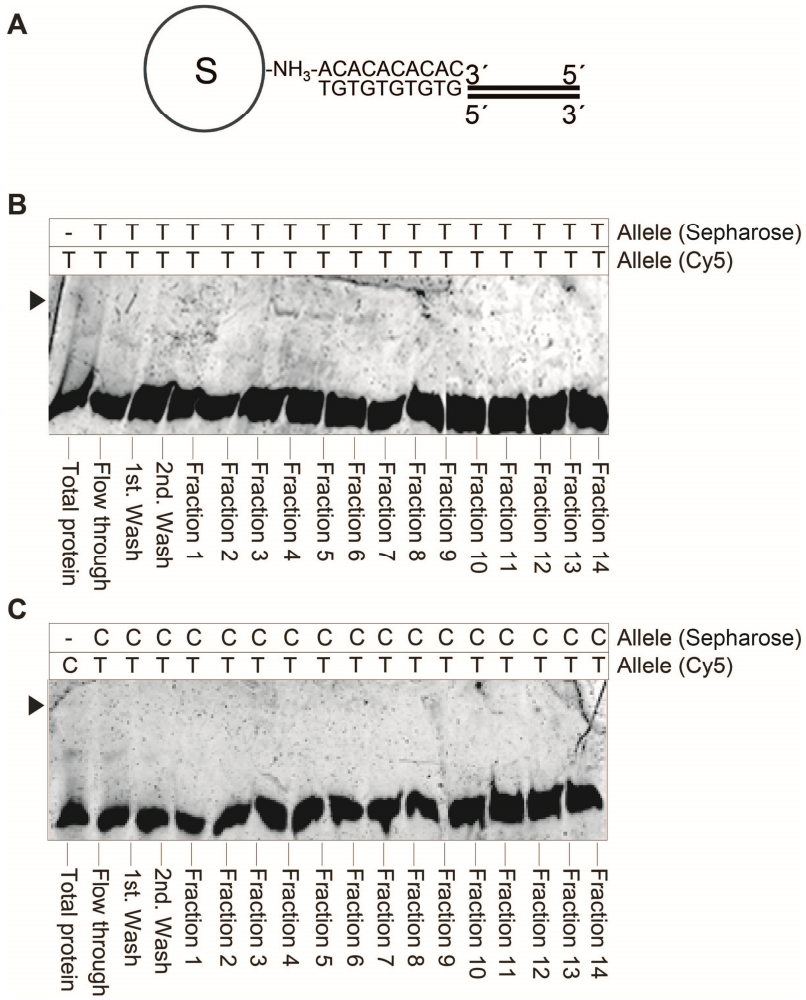


Figure 14. EMSA analysis of fractions obtained during affinity purification process for the rs1421085 using nuclear extracts from 293T cells.

(A) Oligonucleotides were first end-labeled with $(AC)^5$(5'-NH2-ACACACAC-3'). Then, the oligonucleotides were coupled to the CNBr-pre-activated sepharose beads. S: Sepharose. Total 0.9 µmol of oligonucleotides (50 nmol of DNA pro 1 g sepharose) was used in the affinity purification. (B-C) Affinity chromatography using sepharose beads was performed using 500 µg nuclear extracts from 293T cells for the T (B) and the C (C) allele as $(AC)^5$-labeled probe, respectively, as depicted in Methods (see 5.2.18). The specified chromatographic fractions (flow through, wash and fractions) were analyzed to confirm the enrichment of the allele-specific binding proteins using the T allelic Cy5-labeled probe of the rs1421085. *Total protein:* 293T nuclear extracts, *Flow through*: flow freely through the column, *Wash*: low concentration of KCl (100 mM). *Fraction 1-14*: Elution of proteins with increasing concentration of KCl (200 -1000 mM KCl, respectively). A black arrow indicates the allele-specific band. The experiment was performed only once.

Cell line	Allele-specific signal by EMSA	AC	NE amount applied for AC	No. of proteins identified[a]
293T	faint, but visible	magnetic	125 µg	145
	clear visible	sephaorse	500 µg	143

Table 21. Total number of allele-specific binding proteins identified by mass spectrometry at the predicted *cis*-regulatory, rs1421085.

[a]Number of all proteins identified by LC-MS/MS. LC-MS/MS data were quantitative analyzed by *Progenesis*. NE: nuclear extracts, AC: affinity chromatography.

LC-MS/MS analysis identified a total of 145 and 144 proteins for purification using magnetic and sepharose beads, respectively (Table 21). Although four times more amount of 293T nuclear extracts was applied for the sepharose-based purification compared to the magnetic-based purification, there was no difference in the number of the identified proteins between both protein purification strategies. In order to further assess the efficiency of the affinity chromatography strategies to isolate and enrich the allele-specific binding proteins, a comparison was made of the proteins isolated from affinity purification using magnetic beads with using sepharose beads. Among all identified proteins, a total of 77 and 75 proteins were detected solely in the eluates from the magnetic and sepharose bead-based purifications, respectively. Only 68 proteins were detected in both groups, although both affinity purifications were performed using nuclear extracts from the same cell line (Fig. 15). Initially, proteins were considered as a putative candidate only if the fold change of proteins between both alleles was more than two fold (T > C or C > T). Among 68 proteins found in both strategies, only 5 proteins fulfilled these criteria, however most of them were cytoplasmic proteins including keratin. Moreover, the proteins were not considered further, which were detected by only 1 peptide. Finally, proteins were only considered further based on biological relevance due to low abundance of proteins in eluates. Some proteins were found to be DNA binding proteins including transcription factors. Only very few proteins were detected to be involved in the adipocyte function (reference not given, not published). Thus, further purification was required on a large scale to facilitate the selection of candidate proteins.

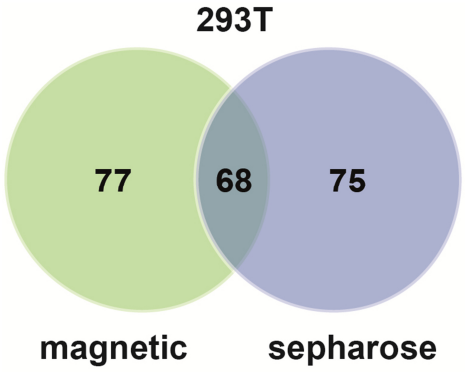


Figure 15. Venn diagrams showing the overlap between allele-specific binding proteins isolated through two affinity purification techniques.

Venn diagrams indicate the distribution of allele-specific binding proteins isolated from affinity purifications using magnetic or sepharose beads and then identified by subsequent LC-MS/MS analysis. Venn diagrams represent all proteins identified by LC-MS/MS. Interestingly, only 68 among 220 proteins shared among these two groups, despite both of these techniques used nuclear extracts from the same cell line, 293T.

7.2.2.2.3 Large scale affinity purification using magnetic beads

A pilot study of protein purification using sepharose beads was previously introduced (see above). However, there were some difficulties such as limited scalability, time consuming, costly and harmful procedures. Moreover, although four times more amount of nuclear extracts were applied for sepharose beads-based approach (500 µg for each allele) than magnetic beads-based one (125 µg for each allele) (Table. 21), the number of proteins identified by LC-MS/MS was not different from each other, indicating less efficient isolation of proteins using sepharose purification in this study. Thus, streptavidin-biotin system was applied for further purification, however on a large scale in order to improve the efficiency of enrichment of target proteins. To examine and further compare the specificity of the allele-specific binding proteins in different tissue or cell line, affinity purifications were performed using nuclear extracts from mouse adult brain and Huh7 cell line on a large scale. In the previous chapter (7.2.2.2.1), affinity purification using nuclear extracts from 293T cells (500 µg for each allele) was already performed. Thus, affinity purification using 293T cells was not considered in this chapter.

First, affinity purification started with nuclear extracts from mouse adult brain (2.5 mg for each allele). The initial starting material consisted of nuclear extracts prepared from the brains

of mice around 8 weeks of age. Corresponding to the large-scale concept, the affinity purification was performed with 52 pmol of biotinylated oligonucleotides and 500 µg magnetic beads of Dynabeads® M-280 beads (corresponding to the binding capacity of 100 pmol biotinylated oligonucleotides). During affinity purification, the reaction solution was washed with 50 mM NaCl, and the elution was performed using elution buffer containing in range from 200 to 1000 mM NaCl. Subsequently, EMSAs were performed to observe the enrichment of the allele-specific binding proteins using the Cy5-labeled C and T allele probes (C allele for the 1. band and T allele for the 2. band), respectively (Fig. 12). The results of the EMSA showed that only the second allele-specific band appeared in the eluate E300, whereas the first allele-specific band disappeared. In addition, the signal intensity of the second band was very faint (data not shown). These observations suggested that the elution conditions might be inappropriate for elution of protein-DNA complexes in the both allele-specific bands. Otherwise, it is also possible that the first allele-specific binding protein did not bind to the beads at all or inefficiently, or was already eluted during wash steps, or not eluted yet under the applied buffer conditions. To circumvent these problems, further affinity purifications were performed using wash buffer containing 10 mM NaCl. Also, the elution was performed using elution buffer containing a range from 50 to 500 mM NaCl instead of from 200 to 1000 mM NaCl. All purification steps were monitored by EMSA using Cy5-labeled probes containing each allele (T/C), respectively. In EMSA using the T allele probe, the second allele-specific band (Fig. 12) appeared with a more intense signal to the T allele than the C allele in eluate E300 (Fig. 16A). Conversely, there was no signal for the first allele-specific band in EMSA using the C allele probe (data not shown), suggesting that the first allele-specific binding protein still failed to be isolated during affinity purification, and other buffer condition would be required which is specialized for the first allele-specific binding proteins. In addition, there were several intensive signals of non-allele-specific bands, which could inhibit the binding of the allele-specific proteins to the DNA. To improve the selective binding of the allele-specific proteins and reduce non-allele-specific signals, the amount of the biotinylated oligonucleotides was reduced from 52 to 44.2 and 18.6 pmol in further affinity purifications, respectively. However, the purification using reduced amount of the biotinylated oligonucleotides (18.6 and 44.2 pmol, respectively) did not result in improved enrichment of the allele-specific binding proteins, as assessed by EMSA assays (data not shown).

Next, nuclear extracts from Huh7 cells (2.8 mg for each allele) were used for further purification on a large scale. In the case of Huh7 cells, the purification started with relatively

low amount of biotinylated oligonucleotides (18.6 pmol). Like in the purification using nuclear extracts from mouse adult brain, 50 mM NaCl in washing buffer was used for washing step. The elution was performed with increasing NaCl concentrations (in a range from 200 to 1000 mM NaCl). All steps along the affinity purification process were monitored by EMSAs using Cy5-labeled T- and C-allele probe, respectively. Compared to the previous EMSA results (Fig. 12), the first allele-specific band was not detected in any of the eluates, while the second allele-specific band was shown in eluate E200, as assessed by EMSA after protein purification (data not shown). This result suggests that the second allele-specific binding protein could be relatively weakly charged. A common strategy to isolate weakly charged proteins is lowering the salt concentration during elution, while the more strongly charged proteins are eluted at higher salt concentrations [402]. In addition, the protein-DNA complex was poorly resolved by EMSA, and the signals were apparently lower than backgrounds. Thus, to improve the enrichment of target proteins and further reduce the remaining contaminants to acceptable levels, the salt concentration of washing buffer was changed from 50 to 10 mM NaCl. Thereby, elution steps also started with lower salt concentration, 50 until 500 mM NaCl. Indeed, the low salt-based washing and elution steps led to the improved resolution of the second allele-specific band in eluate E200 (T > C), assessed by EMSA experiment using the Cy5-labeled T-allele probe (Fig. 16B). However, the purification under these conditions still failed to isolate the first allele-specific binding proteins (data not shown). Finally, the eluates E200, E300 and E400 obtained from both purifications (Fig. 16) were selected for further LC-MS/MS analysis, which showed signal of the allele-specific binding proteins in repeated experiments.

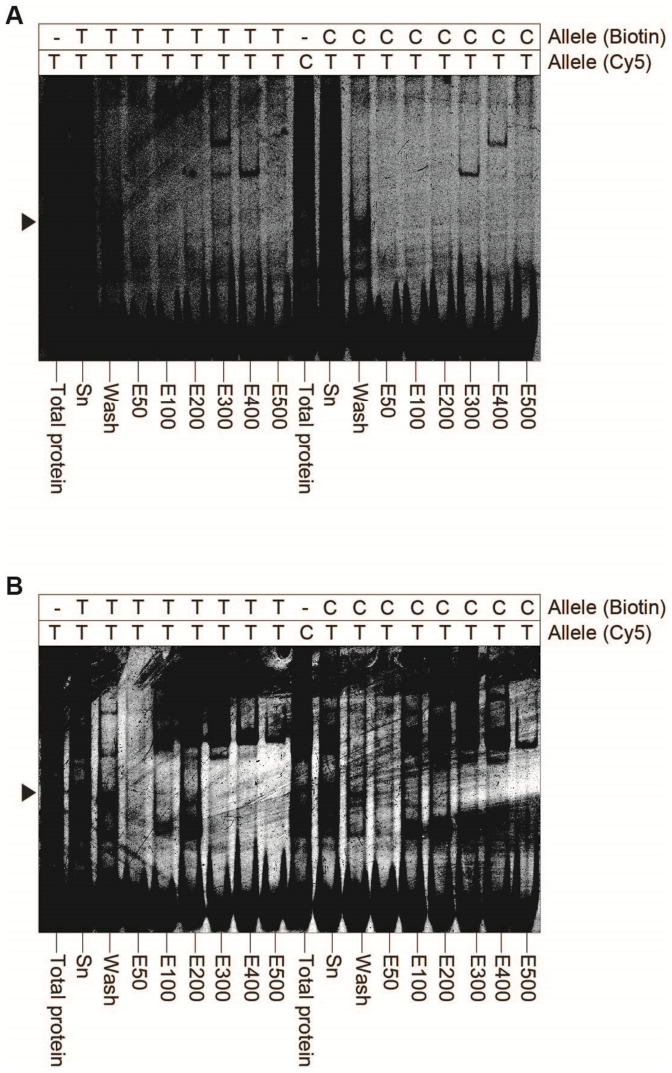


Figure 16. EMSA analysis of eluates obtained during affinity purification process for the rs1421085 using nuclear extracts from mouse adult brain and Huh7 on a large scale.

Nuclear extracts prepared from (A) mouse adult brain and (B) Huh7 cells were subjected to affinity purification as depicted in Methods (see 5.2.17), respectively. Affinity chromatography was performed using 2.5 mg nuclear extracts from mouse adult brain and 52 pmol biotin-labeled probes (A) and 2.8 mg nuclear extracts from Huh7 and 18.6 pmol biotin-labeled probes (B) for each allele. The initial nuclear extracts and specified chromatographic fractions (wash and eluates) were analyzed to confirm the enrichment of the allele-specific binding proteins using the T allelic Cy5-labeled probe of the rs1421085. A black arrow indicates the allele-specific band. *Total protein:* nuclear extracts, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (10 mM). *E50-500*: Elution of proteins with increasing concentration of NaCl (50 -1000 mM NaCl, respectively). All experiments were performed in triplicates.

Tissue/cell line	Allele-specific signal by EMSA	NE amount applied for AC	No. of proteins identified[a]
Mouse adult brain	faint, but visible	2.5 mg	703
Huh7	faint, but visible	2.8 mg	1039

Table 22.Total number of allele-specific binding proteins identified by mass spectrometry at the predicted *cis*-regulatory, rs1421085.

[a]Number of all proteins identified by LC-MS/MS. LC-MS/MS data were quantitative analyzed by *Progenesis*. NE: nuclear extracts, AC: affinity chromatography.

By LC-MS/MS analysis, in total 737 and 1,039 proteins were identified in mouse adult brain and Huh7, respectively (Table 22). Comparing the numbers of proteins identified by LC-MS/MS (Table 22) with those from the previous purification on a small scale (Table 21), there was an obvious increase in the number of proteins identified by LC-MS/MS after using a larger amount of nuclear extracts in the purification. These results are in line with the previous data at the *PPARG* locus (see chapter 7.2.1.2.2), indicating that the amount of protein is one of the critical parameters not only for EMSA, but also for protein purification. Moreover, it often requires large amounts of starting material for successful enrichment of target proteins since yield of purification is always less than 100% [403].

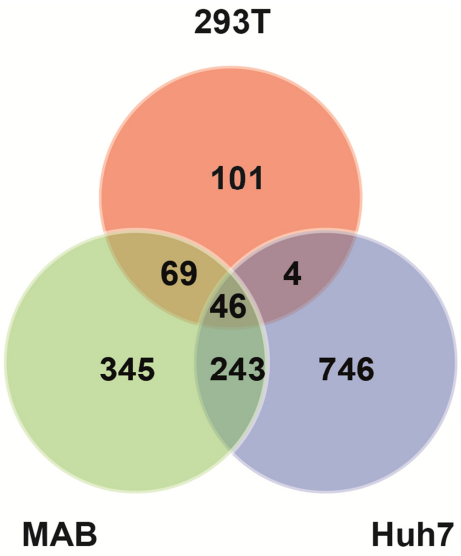


Figure 17. Venn diagrams depicting the overlap between the allele-specific binding proteins isolated by affinity purification using nuclear extracts from 293T, mouse adult brain and Huh7.

Venn diagrams indicate the distribution of allele-specific binding proteins isolated from affinity purifications using different nuclear extracts and then identified by subsequent LC-MS/MS. Venn diagrams represent all proteins identified by LC-MS/MS. Only 46 proteins are shared among these three groups.

To assess cell- or tissue-specific distribution of the proteins binding at the rs1421085, all proteins identified in 293T, Huh7 cells and mouse adult brain tissue were compared. Venn diagrams indicated shared and exclusive proteins among three groups (Fig. 17). Interestingly, the majority of proteins identified by LC-MS/MS was non-overlapping each other, while only 46 proteins were present in all cell lines and tissue analyzed. In total, 103, 345, 746 proteins were detected only in 293T cells, mouse adult brain tissue and Huh7 cells, respectively (Fig. 17). Note that protein contaminants including keratin, serum albumin [404] were not considered for further analysis. All other proteins identified were considered further. To understand the functional roles of the proteins identified in each cell line or tissue, each set of all identified proteins was functionally categorized based on universal gene ontology (GO) annotation terms by using the Genomatix GePS-tool (Genomatix, Munich, Germany) and were classified into cellular component, molecular functions, and biological process categories (data not shown, not yet published). Briefly, under the "molecular functions" category, of 703 and 1039 proteins identified, 41 (5.8 %) and 39 (3.8 %) proteins were annotated to the GO-term *sequence-specific DNA binding* in mouse adult brain ($P = 1.84 \times 10^{-6}$, Fisher´s exact test) and Huh7 cells ($P = 1.33 \times 10^{-12}$, Fisher´s exact test), respectively. Moreover, 43 (6.1 %) proteins in mouse adult brain showed a strong enrichment in the GO-term *transcription regulator activity* ($P = 2.13 \times 10^{-3}$, Fisher´s exact test). In Huh7 cells, 40 (8.4 %) proteins were annotated to the GO-term *protein binding transcription factor activity* ($P = 8.83 \times 10^{-5}$, Fisher´s exact test) and *transcription cofactor activity* ($P = 7.87 \times 10^{-5}$, Fisher´s exact test). Next, the enrichment of canonical signaling pathways was assessed by the GePS tool (Genomatix) within the set of all identified proteins. In total, 8 signaling pathways were found in both mouse adult brain tissue and Huh7 cells. Of the 8 signaling pathways, the most statistically significant association was seen for the *Signaling events mediated by HDAC Class I* ($P = 3.33 \times 10^{-5}$, $P = 1.32 \times 10^{-5}$, Fisher´s exact test) and the *mechanisms of transcriptional repression by dna methylation* ($P = 3.65 \times 10^{-4}$, $P = 9.84 \times 10^{-5}$, Fisher´s exact test) for mouse adult brain tissue and Huh7 cells, respectively. Further interesting pathways were included: the *ATR signaling pathway* and the *prc2 complex sets long-term gene silencing through modification of histone tails*.

Cell line	Gene symbol	allelic fold change (T/C)[a]	*P-value*[b]	Peptide count for quantitation [c]	Mascot score [d]
Mouse adult brain	TF2	0.7	0.3	11	491
	TF3	1.6	0.7	2	106
	TF4	0.9	1.0	41	2127
Huh7	TF4	0.3	1.2	2	47

Table 23. Candidate proteins from proteome analysis in mouse adult brain and Huh7.

[a]fold change was calculated as the mean ratio of normalized proteins abundance over the three experiments, [b]*P*-values were derived from unpaired *t*-tests, [c]Peptide count for quantitation refers to the number of peptides uniquely assigned to one protein and therefore used for quantitation, [d]Mascot score is built as summed up single probability of identified peptides per protein and serves as indicator for the reliability of protein identification. LC-MS/MS data were quantitative analyzed by *Progenesis*.

To select candidate proteins binding at the rs1421085 in an allele-specific manner, all proteins identified by LC-MS/MS in mouse adult brain tissue and Huh7 cells were first sorted and ranked according to the fold change, *P*-value and high accuracy (number of identified peptides for quantification $\geq$ 2, Mascot percolator score > 13, FDR < 1%, see Online Methods). However, no protein was found with either significant fold change (> 2 or < 0.5) or *P*-value (< 0.05, unpaired *t*-test, n = 3) in contrast to the previously case of the *PPARG* loci where several candidates were observed which fulfilled all the criteria (see chapter 7.3). This may be explained by the fact that the EMSA signals of the allele-specific bands after affinity purifications using mouse adult brain tissue or Huh7 cells were very faint (Fig. 16) compared to those using HIB 1B adipocytes at the *PPARG* locus (Fig. 9). Thus, the selection of the candidate proteins at the *FTO* locus was relied on *in silico* prediction (SNPInspector, Genomatix, Germany) and biological relevance based on published data (data not shown, not yet published). On the basis of those results, several proteins were selected as putative candidates including TF2, TF3 and TF4 (referred to as TF2, TF3 and TF4 in this study, respectively as not yet published) (Table 23). The transcription factor TF2 was identified with a fold change of 0.7 (*P* = 0.3) as well as the transcription factor TF3 with a fold change of 1.6 (*P* = 0.7), which were detected only in mouse adult brain tissue. Based on *in silico* analysis, TF2 was predicted as an allele-specific binding protein to the rs1421085 (with a stronger binding to the C allele than the T allele) with 0.877 matrix similarity (data not shown, not yet published). TF4 is the only one that was found in both mouse adult brain tissue (0.9-fold, *P* = 1.0) and Huh7 cells (0.3-fold, *P* = 1.2) among the candidates listed in Table 23. TF4 was previously reported to bind to the other SNP at the *FTO* locus in an allele-specific manner,

thereby to regulate the *FTO* gene expression (reference not given, not published). Interestingly, none of the candidates mentioned above was found in 293T cells (data not shown), which could be due to the small number of proteins identified (Table 21). Finally, these three candidates were included in further analysis as putative allele-specific binding proteins.

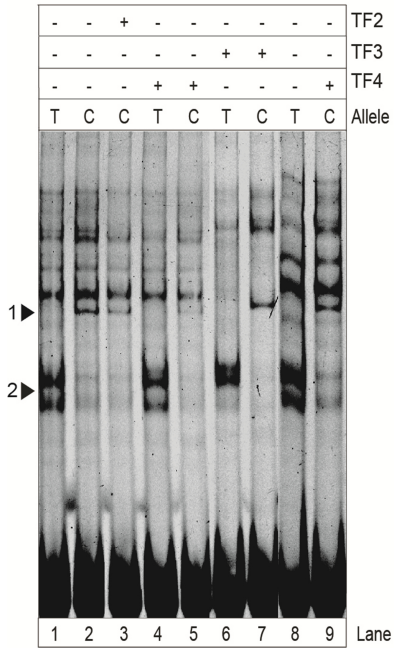


Figure 18. Allele-specific binding of the transcription factors TF2, TF3 and TF4 at the predicted *cis*-regulatory variant, rs1421085.

Competition EMSA was performed to prove TF2, TF3 and TF4 as allele-specific binding proteins to the rs1421085. Competition EMSAs were performed without (lane 1-2 and 8-9) or with 100 excess of unlabeled oligonucleotides for TF2 (lane 3), TF3 (lane 6-7) and TF4 (lane 5-6) using nuclear extracts from mouse adult brain. After addition of cold TF2, TF3 and TF4 probe the two allele-specific bands (marked with number 1 at the C allele and number 2 at the T allele) were disappeared or diminished. A black arrow indicates the allele-specific band difference.

To verify the binding specificity of those proteins to the rs1421085, competition assay was performed by adding unlabeled consensus sequences for TF2, TF3 and TF4 to the binding reaction, respectively (Fig. 18). Competition assay was first performed using nuclear extracts from mouse adult brain tissue with 100 fold excess of each competitor. In the presence of TF2 cold competitor, the first allele-specific band (marked with number 1) was visibly

diminished at the C allele (lane 3). Also, the first allele-specific band at the C allele was efficiently blocked in competition by unlabeled TF4 consensus binding sequence (Fig. 18, lane 4 and 5), while other protein-DNA complexes were not affected at all. The second allele-specific band at the T allele (marked with number 2) was seen as a thick band in mouse adult brain in the previous EMSA (Fig. 12), which was separated in two closely migrating bands (Fig. 18, lane 1 and 8). The use of TF3 cold competitor abolished the under part of the second allele-specific band (lane 6 and 7). Similar results were shown in EMSA assays using nuclear extracts from Huh7 cells (data not shown). These results suggested that TF2, TF3 and TF4 might bind to the rs1421085 directly in an allele-specific manner or involved in the allele-specific protein-DNA complex formations. However, these results are preliminary, and need to be confirmed by further experiments.

7.2.3 *TCF7L2* locus

7.2.3.1 Optimization of conditions for allele-specific binding of proteins

Associations between *TCF7L2* genotype and T2D have been evaluated extensively in a variety of studies [405–409]. The important role of the *TCF7L2* gene has been well established in pancreatic beta cell function in the context of insulin secretion and related metabolic phenotypes [407,410,411]. On the basis of numerous studies assessing epigenetic marks of regulatory regions [254,255], PMCA analysis [98] and further allele-specific regulatory properties [98,254,255,380], the rs7903146 was predicted as a *cis*-regulatory variant at the *TCF7L2* locus. The allele-dependent protein binding and transactivation by the rs7903146 was examined previously by EMSA and reporter gene assay [98]. Similarly to the approaches for other loci including *PPARG* and *FTO*, the allele-specific binding of proteins to the rs7903146 was confirmed by EMSA experiment. EMSA was performed with Cy5-labeled 45 bp oligonucleotides surrounding the rs7903146 region, which was predicted by PMCA corresponding to the length of recognition of CRM [98]. Here, INS-1 cell line (rat pancreatic beta cell line) was chosen as a major source in future experiments for two main reasons. First, the *TCF7L2* gene is dominantly expressed in pancreatic beta cells [235]. INS-1 is a suitable cell line for measuring glucose-stimulated insulin secretion [412] due to the difficulties in generating human pancreatic beta cells [413]. Second, INS-1 cell line has been well-studied in the literature, giving the opportunity to compare results with data from literatures in the same cell

line. In total, three allele-specific bands were observed in EMSA experiment (Fig. 19A). In consistent with the previous EMSA data [98], the most intense allele-specific shifted bands (marked with number 2 and 3) were observed at the T allele (Fig. 19A, lane 1-2). Like the mid-position band (marked with number 2), the lower shifted band (marked with number 3) showed more intensive binding of the proteins at the T allele than at the C allele. Conversely, the upper shifted band (marked with number 1) was observed with increased DNA-binding activity of proteins at the C allele compared to the T allele. To examine the possible influence of buffer components on the pattern of DNA binding activities of nuclear proteins, EMSA experiment was performed using different gel binding buffers including 5x GBB or 4x GBB. Unlike 5x GBB as a standard buffer for EMSA, 4x GBB contains different type of salt and higher salt concentration (250 mM NaCl → 800 mM KCl). Notably, the use of the 4x GBB resulted in the abrogation of the allele-specific shifted band at the mid-position (marked with number 2), which was probably due to inappropriate binding condition for the proteins caused by too high salt concentration (Fig. 19A, lane 3-4). Hence, the binding buffer 4x GBB was not considered for further EMSAs and protein purifications. Next, EMSA was performed using different amounts of nuclear extracts and concentrations of poly (dI-dC) in order to obtain optimal binding condition for protein-DNA complex formation. Compared to using 2.5 µg nuclear extracts (Fig. 19B, lane 1-6), EMSA using 5.0 µg nuclear extracts resulted, in general, not only in a high background smear of non-allele-specific proteins, but also in smeared allele-specific band with poor resolution (Fig. 19B, lane 7-12). Especially, the second allele-specific band (marked with number 2) was shifted upwards in mobility (i.e. slower migrating) (lane 7 and 8). To reduce non-allele-specific signals, the concentration of poly (dI-dC) was tested in range from 4.9 to 9.8 ng/µl in EMSA assay. Increasing concentration of poly (dI-dC) reduced obviously the signal intensities of non-allele-specific bands, as shown in Fig. 19B. Note that the second allele-specific band (marked with number 2) was partly abrogated with increased concentration of poly (dI-dC) during other allele-specific bands (marked with number 1 and 3) remained. Consequently, it abolished the allele-specific feature of the second allele-specific band (marked with number 2) (Fig. 19B, lane 3-6). Finally, the optimal binding condition including 2.5 µg nuclear extracts with 4.9 ng/µl poly (dI-dC) in 5x GBB binding buffer was chosen for further analysis based on EMSA results (Fig. 19).

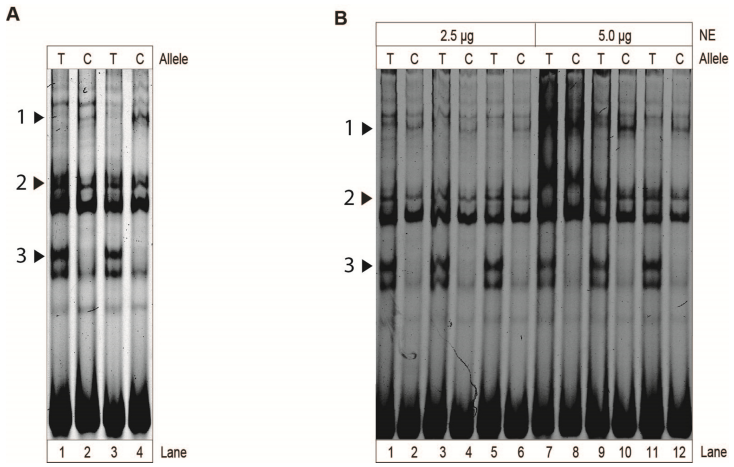


Figure 19. Analysis of allele-specific protein-DNA interaction at the predicted *cis*-regulatory variant, rs7903146 of the *TCF7L2* locus. .

To examine the allele-specific binding of proteins at the rs7903146 of the *TCF7L2* locus, EMSA experiments were performed using allelic Cy5-labeled probes for the rs7903146 and nuclear extracts from INS-1 cells. (A) EMSA was performed using 2.5 µg nuclear extracts and 4.9 ng/µg poly (dI-dC). The predicted *cis*-regulatory variant, rs7903146 showed three allele-specific shifted bands (marked with number 1, 2 and 3). lane 1-2: binding reaction in 5x GBB, lane 3-4: binding reaction in 4x GBB. (B) Cy5-labeled probe for the rs7903146 was incubated with different amounts of nuclear extracts. lane 1-6: 2.5 µg and lane 7-12: 5.0 µg nuclear extracts. In addition, different concentrations of poly (dI-dC) (in the range from 4.9-9.8 ng/µl) were used to reduce non-allele-specific signals. lane 1-2 and 7-8: 4.9 ng/µl, lane 3-4: 6.0 ng/µl , lane 5-6 and 9-0: 7.0 ng/µl , lane 11-12: 9.8 ng/µl. A black arrow indicates allele-specific band.

7.2.3.2 Affinity chromatography coupled to mass spectrometry using magnetic beads on a large scale

Next, affinity chromatography was performed using nuclear extracts from INS-1 cells to enrich the allele-specific binding proteins. In the previous studies, two approaches were used to isolate and enrich the allele-specific binding proteins: affinity purification using magnetic beads and sepharose beads. Finally, affinity purification using magnetic beads was used further for some reasons as described above (see chapter 7.2.2.2). Here, 1 mg of nuclear extracts for each allele was used for the purification, which was 400 times more increased than in EMSA experiments (2.5 µg) (Fig. 19). Using the described conditions, the affinity purification was performed with 44 pmol of biotin-labeled oligonucleotides. During the

affinity purification, the reaction mixture was washed three times with 10 mM NaCl, and the elution was started with 50 mM NaCl (50 – 500 mM NaCl). Subsequently, the steps along the affinity purification process were monitored by EMSA using the C or T allelic Cy5-labeled probe, respectively (C allele for 1. allele-specific band and T allele for 2. and 3. allele-specific bands, see Fig. 19). The second allele-specific band (marked with number 2) appeared more intensive at the T allele than at the C allele in eluates E200-400 (Fig. 20A), which was consistent with the previous EMSA results (Fig, 19). The allelic difference was also clearly seen in eluate E400 in EMSA experiment using the C allelic Cy5-labeled probe (Fig. 20B), further confirming the observation in Fig. 20A. The third allele-specific band (marked with number 3) also appeared as a larger smear at the T allele compared to the C allele in eluate E200 (Fig. 20A). In contrast, the first allele-specific binding proteins (marked with number 1) seemed to be either not isolated during affinity purification or present in very low abundance in eluates because there was a barely visible band for the first allele-specific binding proteins (Fig. 20B). These results suggest that the affinity purification conditions such as buffer components might be difficult to be optimal simultaneously for all three allele-specific binding proteins and may require further optimization to ensure the physical stability of protein-complexes during analysis. Finally, eluates E200 and E400 were selected for further LC-MS/MS analysis, which showed obvious enrichment of allele-specific binding proteins in the eluate. However, the eluate E300 was excluded for further analysis due to no obvious allelic difference in the second allele-specific band in repeated experiments (data not shown).

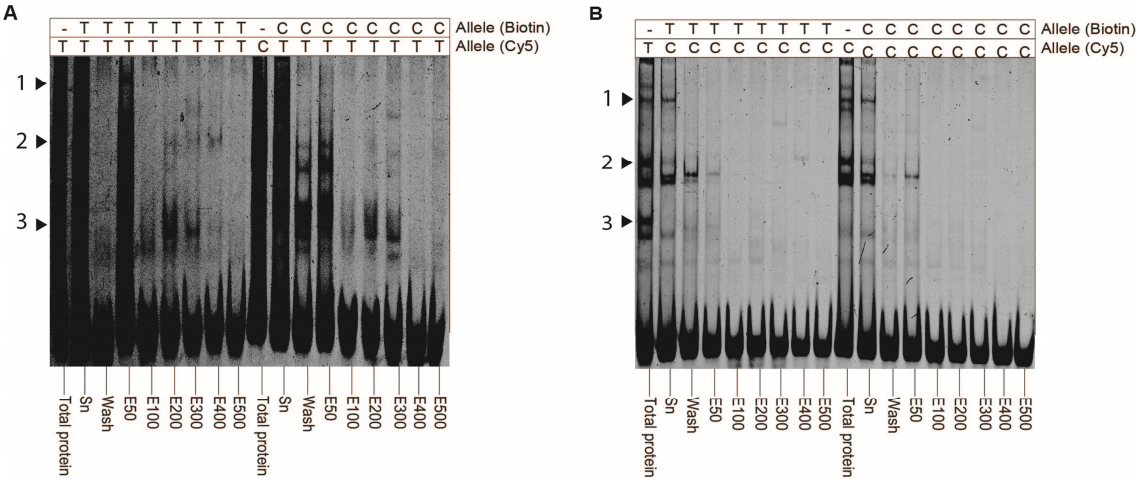


Figure 20. EMSA analysis of eluates obtained during affinity purification process for the rs7903146 using nuclear extracts from INS-1 cells on a large scale.

Nuclear extracts prepared from INS-1 cells were subjected to affinity purification as depicted in Methods (see 5.2.17). Affinity chromatography was performed using 1 mg nuclear extracts from INS-1 cells and 44 pmol biotin-labeled oligonucleotides for each allele. The initial nuclear extracts and specified chromatographic fractions (wash and eluates) were analyzed to confirm the enrichment of the allele-specific binding proteins using the T (A) or C (B) allelic Cy5-labeled probe of the rs7903146, respectively. A black arrow indicates the allele-specific band. *Total protein:* nuclear extracts from INS-1 cells, *Sn*: supernatant after incubation with magnetic beads, *Wash*: low concentration of NaCl (10 mM). *E50-500*: Elution of proteins with increasing concentration of NaCl (50 -500 mM NaCl, respectively). All experiments were performed in triplicates.

The MS analysis resulted in a total of 395 identified proteins. Proteins contaminants such as keratin, serum albumin were excluded for further analysis. All other proteins identified were considered further. To select candidate proteins involved in *cis*-regulatory activity, all proteins identified by LC-MS/MS were sorted according to annotation as transcription factor or their coregulators (MatBase tool, Genomatix), association to GO-terms DNA binding and transcription activity (GePS tool, Genomatix). GO terms were organized in several general categories, cellular component, molecular functions, and biological process categories (data not shown, not yet published). In brief, under the "molecular functions" category, 25 of 395 proteins (6.3 %) were annotated to the GO-term *sequence-specific DNA binding* (P = 1.68 x $10^{-2,}$, Fisher´s exact test). Moreover, 32 (8.1 %) and 31 (7.8 %) proteins showed a strong enrichment in the GO-term *protein binding transcription factor activity* (P = 1.41 x 10^{-11}, Fisher´s exact test) and *transcription cofactor activity* (P = 1.50 x 10^{-11}, Fisher´s exact test), respectively. The enrichment of canonical signaling pathways was also assessed by the GePS tool (Genomatix) within the set of all identified proteins. In total, 16 signaling pathways (P < 1.00 x 10^{-2}, Fisher´s exact test) were found. Of the 16 pathways analyzed, the most statistically significant association was seen for the *mechanisms of transcriptional repression by dna methylation* (P = 2.33 x 10^{-9}, BioCarta:mbdpathway, Fisher´s exact test) and the *Signaling events mediated by HDAC Class I* (P = 1.24 x 10^{-6}, NCI-nature:hdac_classi_pathway, Fisher´s exact test). Further pathways including the *prc2 complex sets long-term gene silencing through modification of histone tails* (P = 3.26 x 10^{-5}, BioCarta:prc2pathway, Fisher´s exact test), the *Hedgehog signaling events mediated by Gli proteins* (P = 6.10 x 10^{-5}, NCI-nature:hedgehog_glipathway, Fisher´s exact test), the *BARD1*

signaling events (P = 2.15 x 10^{-3}, NCI-nature:bard1pathway, Fisher´s exact test), the *regulation of eif2*, the *Validated targets of C-MYC transcriptional activation* (P = 3.66 x 10^{-3}, NCI-nature:myc_activpathway, Fisher´s exact test) and the *E2F transcription factor network* (P = 7.43 x 10^{-3}, NCI-nature:e2f_pathway, Fisher´s exact test) were also found with significant *P*-value.

Next, proteins were first sorted by high accuracy (number of identified peptides for quantification > 2, Mascot percolator score > 13, FDR < 1%, see Online Methods) and then selected according to a fold change (fold change > 2 or < 0.5) and *P*-value (P < 0.05; *t*-test, n = 3). In addition to these criteria, biological relevance was also considered. Several proteins met these criteria (data not shown, not yet published). The candidates were detected with significant fold change including TF3 (7.2), TF4 (5.8) and TF5 (8.4) (Table 24). Moreover, TF3 and TF5 were detected with significant *P*-value (1.8 x 10^{-2} and 3.9 x 10^{-2}, respectively). Of note that TF3 and TF4 were previously identified as an allele-specific binding protein at the rs1421085 of the *FTO* locus (see chapter 7.2.2.2.3). TF4 was detected with non-significant *P*-value (0.07), however considered as interesting candidate due to its allele-specific binding property and biological relevance; i.e. function in pancreas tissue (reference not given, TF not published). Thus, these three proteins were included in further analysis as putative allele-specific binding proteins.

Cell line	Gene symbol	allelic fold change (T/C)[a]	*P-value*[b]	Peptide count for quantitation [c]	Mascot score [d]
INS-1	TF3	7.2	1.8 x 10^{-2}	3	126
	TF4	5.8	0.07	18	166
	TF5	8.4	3.9 x 10^{-2}	34	2364

Table 24. Candidate proteins from proteome analysis in INS-1 cells.

[a]fold change was calculated as the mean ratio of normalized proteins abundance over the three experiments, [b]*P*-values were derived from unpaired *t*-tests, [c]Peptide count for quantitation refers to the number of peptides uniquely assigned to one protein and therefore used for quantitation, [d]Mascot score is built as summed up single probability of identified peptides per protein and serves as indicator for the reliability of protein identification. LC-MS/MS data were quantitative analyzed by *Progenesis*.

7.2.3.3 Validation of putative allele-specific binding proteins using competition and supershift EMSA

Based on the LC-MS/MS data, TF3, TF4 and TF5 were identified with significant fold change and in part with significant *P*-value, as described in Table 24. In an attempt to support the selection of these candidate TFs responsible for the allele-specific binding, an *in silico* analysis was performed to assess the potential functional impact of the rs7903146 on predicted TFBSs using SNPInsepector (Genomatix, Munich, Germany). The result of analysis revealed that TF3 might alter binding motifs of the rs7903146 and moreover preferentially bind at the T allele with 0.952 matrix similarity (data not shown, not yet published). Other proteins such as TF4 and TF5 were not predicted as binding proteins to the rs7903146.

Further EMSAs were performed in the presence of cold probes and antibodies raised against these transcription factors. Assays performed in the presence of anti-TF3 and anti-TF4 antibodies revealed some evidence that TF3 and TF4 might bind to rs7903146 at least in close proximity to the rs7903146 with the probe carrying the T allele compared to controls (data not shown). However, there were some technical issues to be resolved for optimal competition and supershift results.

7.3 Unbiased allele-specific quantitative proteomics unravels molecular mechanisms influenced by *cis*-regulatory genomic variations

GWAS identified thousands of loci associated with diverse diseases [55]. The majority of the identified variants are located in non-coding DNA regions and have been supposed to affect transcriptional regulation [55,100,259–262]. Advances of the ENCODE project [100,414–416] and novel bioinformatics approaches improved the identification of *cis*-regulatory variants at complex loci [98]. Moreover, deciphering allele-specific binding of transcription factors is essential to unravel the mechanisms ultimately affecting gene expression [98,266–269]. However, the identification of allele-specific coregulators remains elusive in most cases, despite the well-established importance of a coordinated interaction between transcription factors, coregulators, and the basal transcriptional machinery for regulation of gene expression [265]. Thus, in most cases the precise molecular mechanisms underlying associations between variants and disease risk remain unknown. Quantitative protein-DNA proteomics, coupling affinity chromatography with LC-MS/MS was reported for identification of enhancer-binding proteins [312,314] and to enable identification of allele-specific DNA binding proteins [98,312]. However, most studies require stable isotope labeling [312] or chemical labeling [314]. Despite

high-sensitivity and high-accuracy of such labeling approaches, they can be time-consuming, have limitations due to high-cost or inefficient labeling [329], may cause artifacts [330], and may be limited by missing data points due to under-sampling.

As shown in the previous chapter (7.2), several *cis*-regulatory variants predicted by PMCA could change their capacity of protein binding in an allele-specific manner. Moreover, to avoid such limitations of the labeling approaches, a label-free quantitative proteomics on salt eluted sub-fractions containing protein binding activity was developed to unravel allelic protein DNA interactions, which allows high coverage of quantified proteins and thereby identification of both, transcription factors and coregulators (Fig. 21A). Of note, it was found that candidate *cis*-regulatory variants at T2D associated *PPARG* locus were inferred by bioinformatics (PMCA) [98] and chromatin immune- precipitation sequencing (ChIPseq) data [353] analysis. Hence, the questions in this section were that *i)* if several *cis*-regulatory variants are present at the *PPARG* locus, which supports experimentally the prediction of PMCA analysis and *ii*) if the method of unbiased, label-free proteomics to identify transcription factors and transcription cofactors could further clarify the difference between *cis*- and non *cis*-regulatory variants.

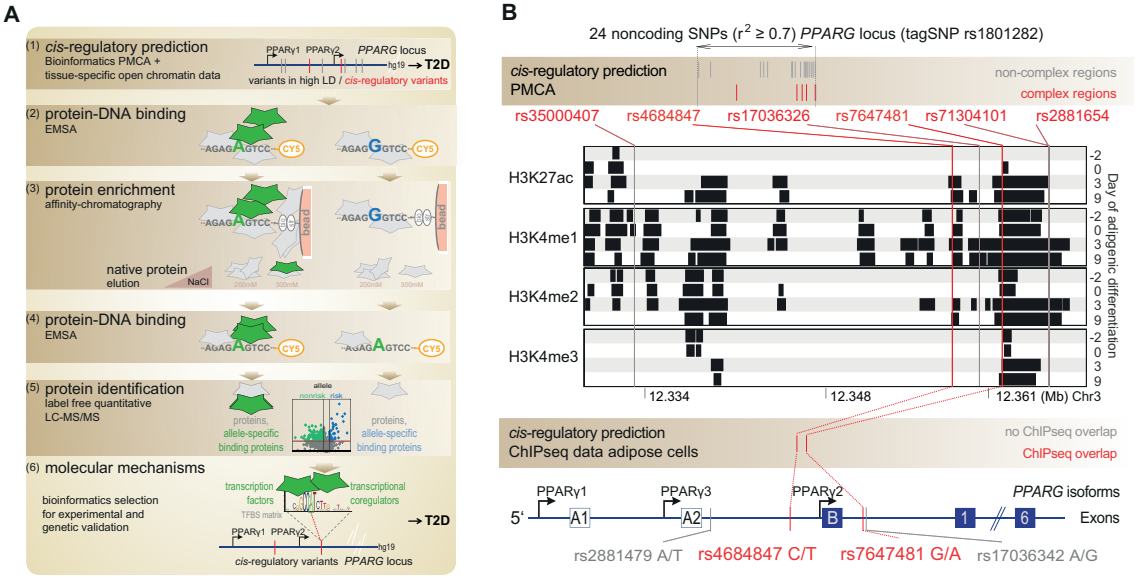


Figure 21. Discovery of allele-specific binding proteins at *cis*-regulatory variants.

(A) Workflow. (1) *cis-regulatory variant prediction* at disease associated variants (*PPARG*) in high LD ($r^2 \geq 0.7$. 1000G) by an integrative framework of bioinformatics phylogenetic TFBS module complexity analysis [98] and adipose tissue epigenetic mark data [353]; (2) *protein-DNA binding* assessed

by Cy5 labeled oligonucleotides matching the risk and nonrisk allele, respectively, in electrophoretic mobility shift assay (EMSA); (3) *protein enrichment* with biotin (bio) labeled oligonucleotides on streptavidin-beads (str) and elution of native protein complexes with increasing concentration of NaCl; (4) *protein-DNA binding* in eluted fractions; (5) *protein identification* and quantification by LC-MS/MS and subsequent label-free quantitative analysis; and (6) *molecular mechanisms.* transcription factor and coregulator selection based on GO-term and network analysis as well as experimental and genetics verification. (B) Bioinformatics and cell-type specific epigenetic mark analysis at the *PPARG* locus infers *cis*-regulatory variants rs4684847 and rs7647481. *Upper panel, PMCA prediction:* analysis of cross-species TFBS pattern conservation predicts the six indicated candidate *cis*-regulatory SNPs (complex regions [98], red) out of 24 proxy SNPs ($r^2 \geq 0.7$. 1000 Genomes Pilot 1 CEU) at the T2D associated *PPARG* locus (tagSNP rs1801282). *Mid panel, active chromatin:* overlap of complex region to H3K27ac, H3K4me1 and H3K4me2 during adipogenic differentiation predicts *cis*-regulatory candidate SNPs rs4684847 and rs7647481. H3K27ac (*histone H3-lysine* 27 acetylation) and H3K4me1-3 (histone H3-lysine 4 mono-. di-. tri-*methylation)* chromatin marks at the *PPARG* locus in primary human adipocyte stem cells during adipogenesis [353]. *Lower panel: l*ocalization of predicted *cis*-regulatory (red) and non *cis*-regulatory (grey) SNPs chosen for following analysis is indicated relative to transcriptional start site of the PPARγ1-3 mRNA isoforms (blue boxes = coding exons, dashed white boxes = untranslated exons, blue lines = introns, black arrows = promoters).

7.3.1 Integration of bioinformatics and epigenetic mark analysis predicts *cis*-regulatory variants at the *PPARG* locus

The *PPARG* locus has been robustly associated with T2D [163,164,170]. Here we used data from the recently reported PMCA approach which assesses phylogenetic conservation of TFBS modularity to predict *cis*-regulatory variants [98]. Six out of 24 non-coding variants in high linkage disequilibrium ($r^2 \geq 0.7$) with the *PPARG* tagSNP rs1801282 were classified as complex, potentially *cis*-regulatory (Fig. 21B). For the rs4684847, a direct overlap to a homeobox transcription factor binding site (TFBS) matrix was reported [98]. The risk allele binding homeobox transcription factor PRRX1 was demonstrated to represses *PPARG* expression, thereby contributing to the insulin resistance phenotype at the *PPARG* locus [98]. Hypothesizing that additional *cis*-regulatory variants contribute to the complex *PPARG* locus phenotype, we further ranked the remaining five PMCA-inferred SNPs according to the overlap of each variant with marks of epigenetic mark during adipogenic differentiation (Fig. 21B) [353]. The overlap of the rs4684847 and rs7647481 with H3K27ac, and rs7647481 with

H3K4me3 epigenetic marks of regulatory region in late stages of adipocyte differentiation indicates a contribution to the regulation of the adipocyte specific *PPARG2* isoform [98,134–136,161,417]. An overlap of H3K4me1 and H3K4me2 in all stages of differentiation was found solely for the rs7647481. This suggests a regulatory region which may contribute to the expression of the *PPARG1* isoform in pre-adipocytes and adipocytes and possibly the ubiquitous expression in most human cell types [134–136,161,417]. Previously, in reporter gene assays of all five variants the strongest effect was confirmatively reported for the rs7647481 [98]. Thus, the integrative analysis, combining bioinformatics PMCA and publically available epigenetic mark data lookups, indicates multiple variants at the *PPARG* locus, i.e. rs4684847 and rs7647481, as candidate *cis*-regulatory.

7.3.2 Allele-specific protein-DNA interaction at the rs4684847 risk and rs7647481 nonrisk allele

Identification of proteins such as transcription factors, differentially binding at *cis*-regulatory variants, is essential to uncover underlying pathophysiological disease mechanisms and design potential interventions. Unbiased, quantitative, label-free protein-DNA proteomics was used for identification of *cis*-regulatory proteins. Allele-specific protein-DNA interaction was analyzed at both predicted *cis*-regulatory variants and two variants predicted as non *cis*-regulatory (Fig. 22A). The rs4684847 variant additionally serves as a positive control for reproducibility of label-free protein-DNA identifications [98]. First, electrophoretic mobility shift assays (EMSA) using nuclear extracts of mouse brown adipocytes, revealed allele-specific protein-DNA interaction for both predicted *cis*-regulatory SNPs, whereas the predicted non *cis*-regulatory did not (Fig. 22A, upper panels). Quantification of protein-DNA complexes confirmed allele-specific binding at both predicted *cis*-regulatory SNPs ($P = 0.03$) in contrast to non *cis*-regulatory SNPs ($P = 0.82$ / 0.80 for rs17036342 / rs2881479, respectively) (Fig. 22A, lower panels). Notably, we found increased protein binding activity at the common risk allele of rs4684847 and at the rare nonrisk allele of rs7647481 (Fig. 22A). The differential protein-DNA interaction patterns were confirmed in EMSA experiments using nuclear extracts from primary human preadipocytes, human SGBS cell strain preadipocytes and *in vitro* differentiated SGBS adipocytes (Fig. 23).

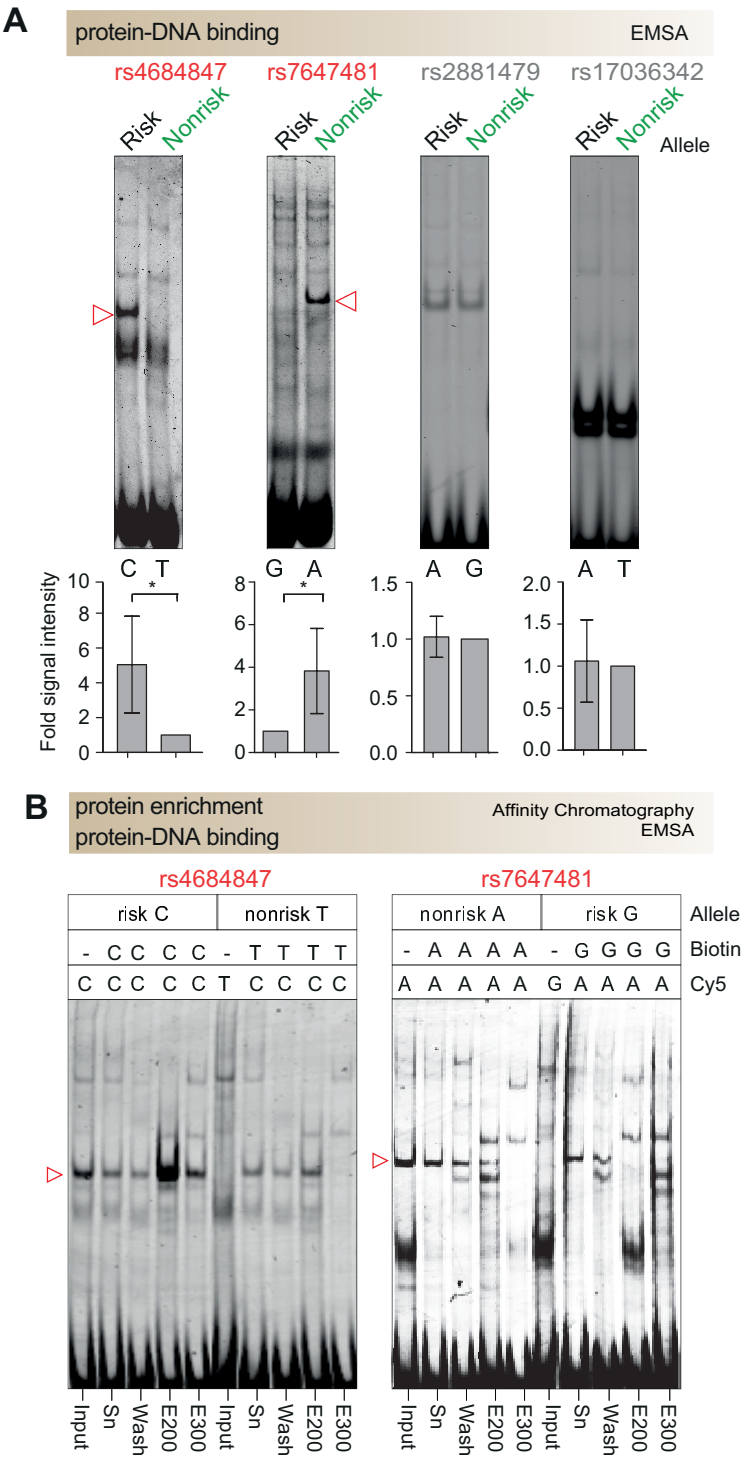


Figure 22. Enrichment of risk and nonrisk allele-specific binding proteins at predicted *cis*-regulatory variants.

(A) Representative EMSA experiments with allele-specific Cy5-labeled probes on nuclear extracts from HIB 1B cells (triangle = allele-specific band) demonstrated allele-specific differential binding affinity of proteins at the risk / nonrisk allele of predicted *cis*-regulatory rs4684847 / rs7647481

variants (red), respectively and no binding at predicted non *cis*-regulatory variants (grey). Bar charts illustrate signal intensity of protein-DNA complexes. Mean ± SD of five experiments. *$P < 0.05$. *P*-value by paired *t*-test. (B) Enrichment of allele-specific differential binding proteins. EMSA with binding-allele specific Cy5-labeled probes of predicted *cis*-regulatory SNPs using protein from affinity chromatography with the respective biotin-labeled risk / nonrisk allelic-probes. Triangle = allele-specific band; input: nuclear protein used for affinity chromatography; Sn: supernatant after incubation with biotin-labeled allelic-probe-magnetic beads conjugates; Wash: low NaCl concentration wash eluates; E200/E300: 200 and 300 mM NaCl protein eluates used for LC-MS/MS. Protein eluates E200 and E300 with differential protein-DNA binding contain the prioritized transcription factors YY1 at the rs7647481 nonrisk and PRRX1 at the rs4684847 risk allele, respectively (Table 26). All experiments were performed in triplicates. For enrichments at predicted non *cis*-regulatory SNPs see Figure 24.

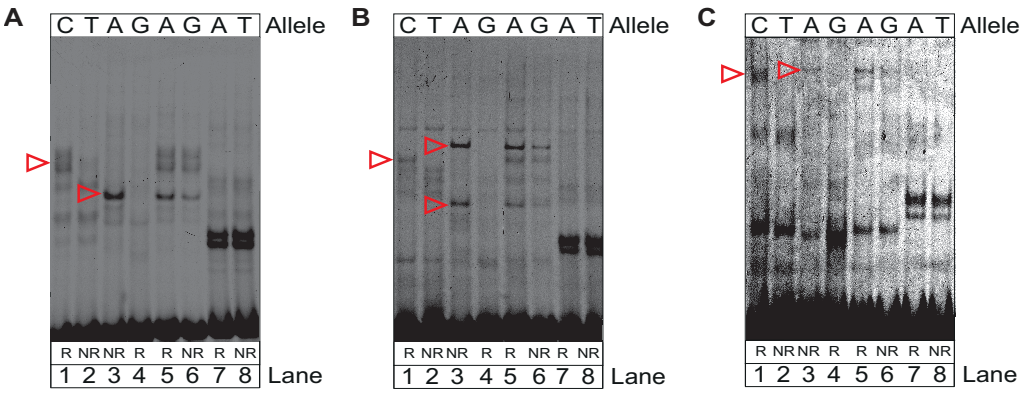


Figure 23. Analysis of risk and nonrisk allele-specific protein-DNA interaction at predicted *cis*-regulatory and non *cis*-regulatory variants in human preadipocytes and adipocytes.

EMSA with allelic Cy5-labeled probes for the two predicted *cis*-regulatory and three predicted non *cis*-regulatory variants using nuclear extracts from undifferentiated primary human preadipocytes (A), the human SGBS preadipocyte cell line (B) and SGBS cells *in vitro* differentiated to adipocytes for 14 days (C). Lanes 1-4: predicted *cis*-regulatory SNPs: rs4684847 (lane 1-2) and rs7647481 (lanes 3-4) showed the rs4684847 risk allele and the rs7647481 nonrisk allele specific protein binding. A red triangele indicates the allele-specific bands. Lanes 5-8: predicted non *cis*-regulatory SNPs rs17036342 (lanes 5-6) and rs2881479 (lanes 7-8) showed no allele-specific difference in protein binding. R = risk allele, NR = nonrisk allele.

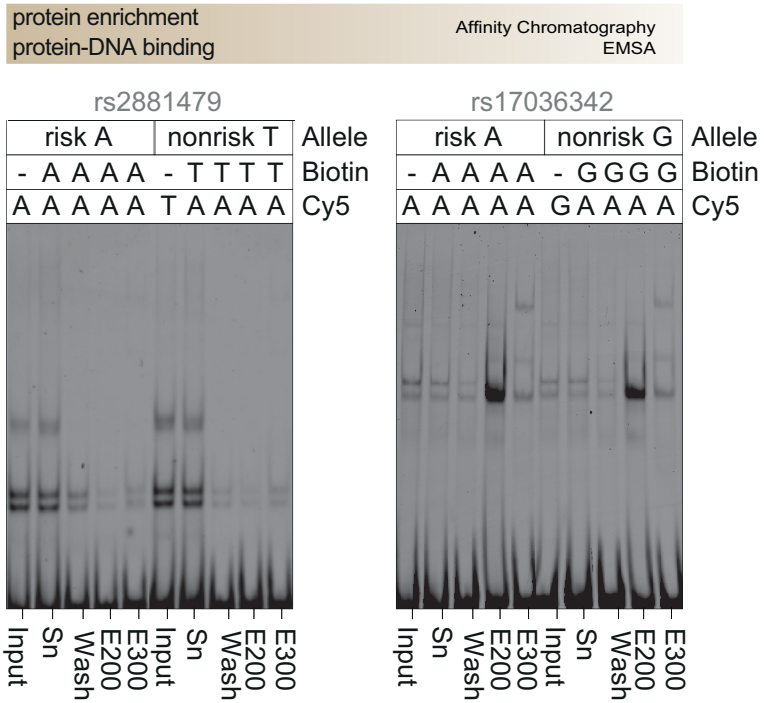


Figure 24. Enrichment of risk and nonrisk allele-specific binding proteins at predicted non *cis*-regulatory variants.

EMSA with binding-allele specific Cy5-labeled probes of predicted *cis*-regulatory variants using protein from affinity chromatography with the respective biotin-labeled risk / nonrisk allelic probes. Input: nuclear protein used for affinity chromatography; Sn: supernatant after incubation with biotin-labeled allelic-probe-magnetic beads conjugates; Wash: low NaCl concentration wash eluates; E200/E300: 200 and 300 mM NaCl protein eluates used for LC-MS/MS (Table 26).

7.3.3 Enrichment of DNA-binding proteins at predicted *cis*-regulatory and non *cis*-regulatory variants

To enrich the allele-specific binding proteins for identification by mass spectrometry, biotinylated oligonucleotides of 40 bp length with risk or nonrisk allele of each SNP at mid-position were incubated with nuclear extracts and concentrated the DNA-binding proteins by affinity chromatography with streptavidin coupled to magnetic beads. Proteomics has been used to identify allelic transcription factor binding with [314,418] or without prior sample fraction [312]. Here, to reduce the notoriously high background of unspecific protein binding to affinity beads and the DNA-backbone [419], an alternative fractionation approach and eluted the native protein complexes by step-wise increasing stringency were used. Thereby, direct control for

an enrichment of allele-specific binding proteins was possible in EMSA assays in the eluted fractions prior to mass spectrometric analysis (Fig. 22B). This focused mass spectrometric analysis to the relevant fractions, effectively reduced complexity and enabled highly specific detection of both transcription factors and transcriptional coregulators at the novel predicted *cis*-regulatory variant rs7647481 (see also chapter on YY1 and RYBP cofactor (7.3.7)) (Table 26). Optimization of binding conditions revealed the best oligonucleotide to bead ratio. Total protein concentration and the concentration of unspecific competitor are further critical experimental parameters (for details see online methods). EMSA experiments with bead-eluted proteins revealed an enrichment of allele-specific protein DNA-binding complexes for both predicted *cis*-regulatory SNPs, by increased binding to the risk C-allele of rs4684847 and to the nonrisk A-allele of rs7647481 (Fig. 22B, left and right panel, respectively). Protein eluates from predicted non *cis*-regulatory SNPs revealed no obvious allelic difference (Fig. 24). Next, protein eluates were subjected to LC-MS/MS and label-free quantification for identification of transcription factors and coregulators.

7.3.4 Label-free quantitative proteomics identifies risk and nonrisk allele-specific binding proteins at predicted *cis*-regulatory variants

Overall, LC-MS/MS detected up to 952 proteins in affinity chromatography eluates at *cis*-regulatory and non *cis*-regulatory SNP adjacent regions (Table 25). We performed label-free quantification based on peptide intensities in the extracted ion chromatograms to assess allele-specific binding of the identified proteins (Online methods). At predicted *cis*-regulatory SNPs we found 142 to 165 proteins, in contrast to only 44 to 82 proteins (20.0 - 16.3% versus 4.6 – 8.8% of LC-MS/MS identified proteins, respectively) at predicted non *cis*-regulatory SNPs with a significant allele-specific binding (fold change > 2 or < 0.5 , $P < 0.05$, n = 3, unpaired *t*-test, Table 25, Fig. 23). Comparing the numbers of allele-specific binding proteins in each set of identified proteins at predicted *cis*- versus non *cis*-regulatory variants, a significant enrichment of differentially binding proteins was found solely at *cis*-regulatory variants for each analyzed pair ($1.87 \times 10^{-25} \leq P \leq 1.72 \times 10^{-6}$), whereas comparing predicted *cis*- versus *cis*-regulatory and non *cis*- versus non *cis*-regulatory SNPs revealed no significant enrichment for most pairs ($4.26 \times 10^{-4} \leq P \leq 0.28$; two-sided, two-group binomial test for pairwise comparison of differentially binding proteins; see Online Methods). Thus, the highest numbers of allele-specific binding proteins were found at predicted *cis*-regulatory

SNPs supporting specific protein-DNA interaction (Fig. 22A). Moreover, when assessing GO-terms for allele-specific binding proteins (fold change > 2 or < 0.5, $P < 0.05$), a strong enrichment in the GO-terms *DNA binding proteins* ($P = 1.36 \times 10^{-6}$, $P = 1.44 \times 10^{-7}$) and *structure-specific DNA binding proteins* ($P = 2.11 \times 10^{-5}$, $P = 3.69 \times 10^{-8}$) was found at the predicted *cis*-regulatory variants (rs4684847, rs7647481) in contrast to low GO-term enrichment at predicted non *cis*-regulatory variants ($4.44 \times 10^{-3} \leq P \leq 0.04$, and $1.47 \times 10^{-3} \leq P \leq 9.43 \times 10^{-3}$, Fisher´s exact test, Supplementary table 1).

	SNP	Total proteins[a]	allele-specific proteins		
			% of total[b]	binding to	per allele[c]
cis-regulatory	rs4684847	824	20	risk	152
				nonrisk	13
	rs7647481	869	16.3	nonrisk	94
				risk	48
non *cis*-regulatory	rs17036342	951	4.6	risk	42
				nonrisk	2
	rs2881479	933	8.8	risk	79
				nonrisk	3

Table 25. Proteins identified by label-free proteomics at predicted *cis*-regulatory versus non *cis*-regulatory variants.

[a]Number of all proteins identified by LC-MS/MS, [b]Number of the allele-specific binding proteins (fold change > 2 or < 0.5 and *P*-value < 0.05) as percent of total proteins identified, [c]Number of the allele-specific binding proteins (fold change > 2 or < 0.5 and *P*-value < 0.05) at each allele of the respective SNP. Data for 300 mM NaCl elution are shown, for 200 mM NaCl elution see Supplementary table 2.

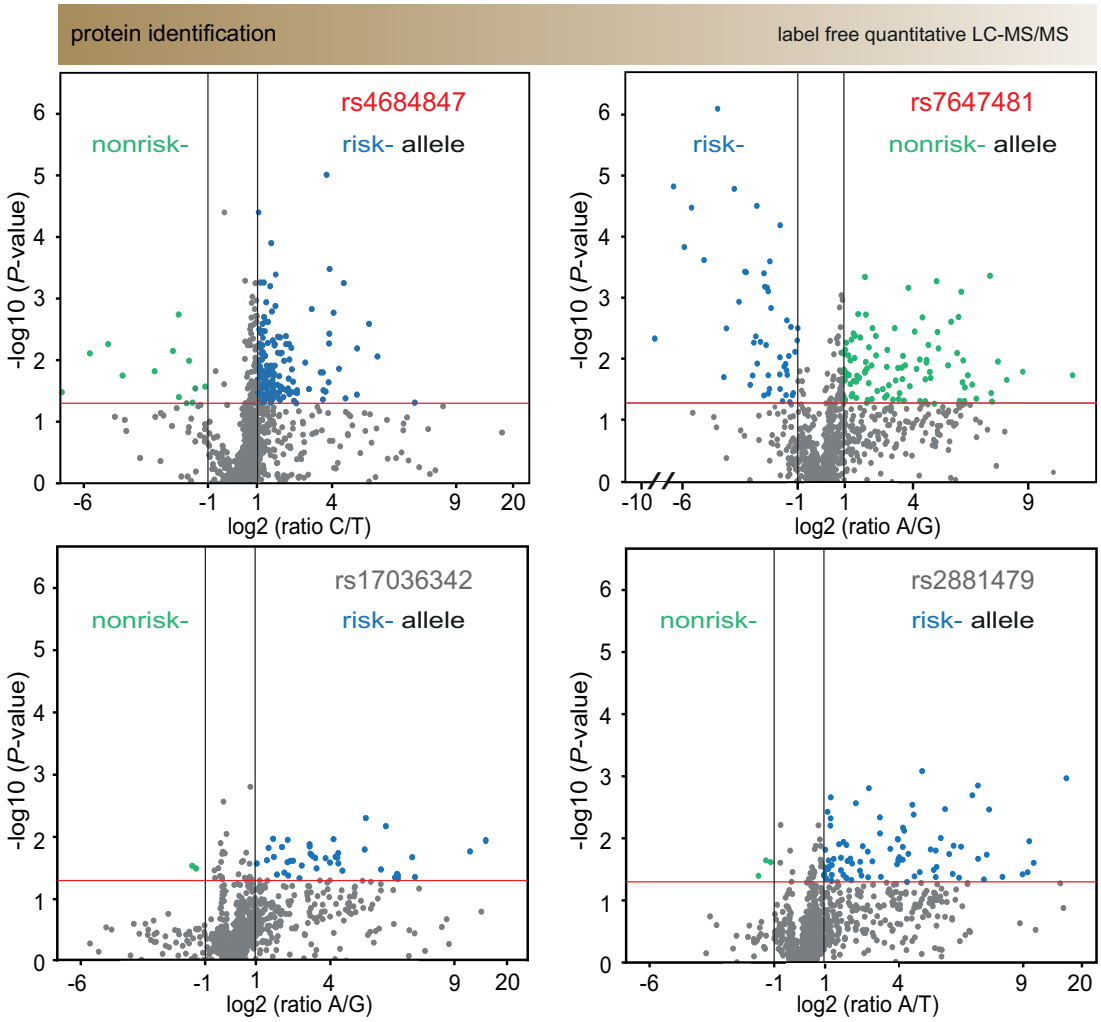


Figure 25. Label-free proteomics identified risk *versus* nonrisk allele-specific binding proteins at predicted *cis*-regulatory and non *cis*-regulatory variants (Eluate 300 mM NaCl).

Volcano plots for the indicated variants illustrate the distribution of risk (blue) and nonrisk (green) allele-specific binding proteins identified by LC-MS/MS (300 mM and 200 mM NaCl eluates see Figure 26) at predicted *cis*-regulatory (red) and non *cis*-regulatory SNPs (grey). Proteins with significant ($P < 0.05$, red line) allele-specific differential binding (allelic ratio < 0.5 or > 2) at the risk allele = blue dots, nonrisk allele = red dots; with no significant allele-specific binding = grey. Mean protein levels (log2 ratio of indicated alleles) and *P*-value from unpaired *t*-test of three independent experiments.

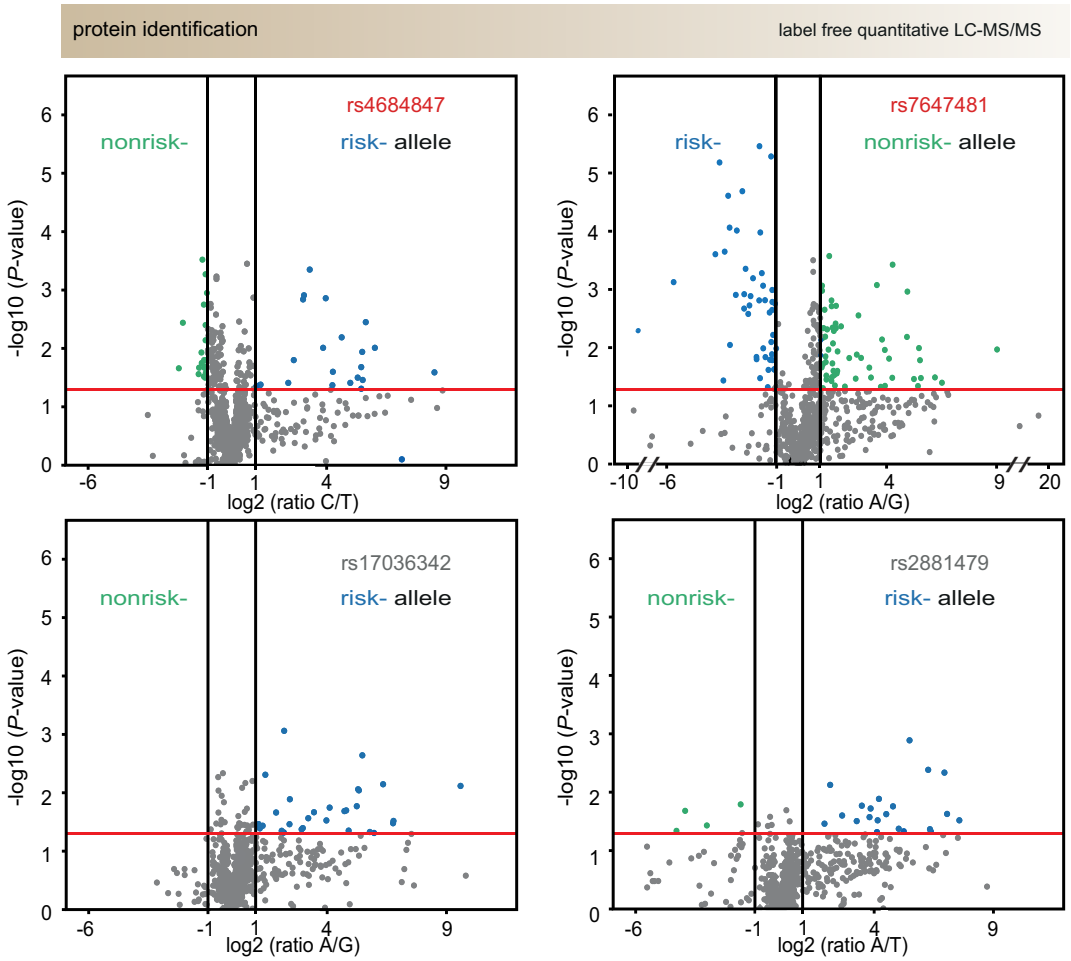


Figure 26. Label-free proteomics inferred risk *versus* nonrisk allele-specific binding proteins at predicted *cis*-regulatory and non *cis*-regulatory variants (Eluate 200 mM NaCl).

Volcano plots for the indicated variants illustrate the distribution of risk (blue) and nonrisk (green) allele-specific binding proteins identified by label-free LC-MS/MS at predicted *cis*-regulatory (red) and non *cis*-regulatory SNPs (grey). Results from 300 mM NaCl eluates (Figure 22 and 24) are shown. Proteins with significant ($P < 0.05$, red line) allele-specific differential binding (allelic ratio < 0.5 or > 2) at the risk allele = blue dots, nonrisk allele = red dots; with no significant allele-specific binding = grey. Mean protein levels (log ratio of indicated alleles) and *P*-value from unpaired *t*-test of three independent experiments.

7.3.5 Prioritizing *cis*-regulatory transcription factors from label-free quantitative proteomics

To select candidate proteins involved in *cis*-regulatory activity, proteins were ranked according to annotation as transcription factor (MatBase tool, Genomatix), association to

GO-terms *DNA binding* and *transcription activity* and identification by LC-MS/MS with significant allele-specific DNA-binding. Criteria for allele-specific DNA-binding were a fold change > 2 or < 0.05 ($P \leq 0.01$, unpaired *t*-test, n = 3, Supplementary table 2) and high accuracy (number of identified peptides for quantification $\geq$ 2, Mascot percolator score > 13, FDR < 1%, see Online Methods). At the *cis*-regulatory variant rs4684847, we identified the transcription factor Prrx1, confirming our previous results [98]. At the predicted *cis*-regulatory variant rs7647481, we identified the transcription factors Yy1 with the highest allelic fold-change (6.6-fold, $P = 2.94 \times 10^{-3}$) as well as Nfatc4 (2.6-fold, $P = 0.01$) (Table 26), while none of the proteins identified at predicted non *cis*-regulatory SNPs fulfilled our selection criteria. Next, we assessed the enrichment of canonical signaling pathways using the GePS tool (Genomatix) within the set of all identified allele-specific binding proteins (fold change > 2 or < 0.5, $P < 0.05$, unpaired *t*-test, Supplementary table 3) and subsequently the occurrence of candidate transcription factors in the identified pathways. Notably, the only transcription factor included in significantly enriched signaling pathways was YY1 ($P < 0.05$, Fisher´s exact test, *E2F transcription factor network*, *p53 pathway*, *prc2 complex sets long-term gene silencing through modification of histone tails*, and *Signaling events mediated by HDAC Class I*). Overall, our data suggest that binding of the transcription factor YY1 at the rs7647481 nonrisk allele may contribute to the *PPARG* locus phenotype, additional to the established role of PRRX1 binding at the *PPARG* rs4684847 risk allele [98].

SNP	Gene symbol	Allelic ratio	Allelic FC	*P*-value (FC)	Quantified peptides
rs4684847	Prrx1	C/T	2.6	0.01	5
rs7647481	Yy1	A/G	6.6	2.94×10^{-3}	9
	Nfatc4		2.6	0.01	2
rs17036342	n.d.				
rs2881479	n.d.				

Table 26. Prioritized *cis*-regulatory transcription factors.

LC-MS/MS identified transcription factors (gene symbol) binding at the rs4684847-C risk and rs7647481-A nonrisk allele prioritized by allelic fold change (FC) > 2.0, *P*-value < 0.05, GO-term annotation, and accuracy of mass spectrometry identification, i.e. number of peptides used for quantification > 2, Mascot percolator score > 13, FDR < 1% (see also Methods and Supplementary table 2). Note that the presented FC for Prrx1 was found in the rs4684847 300 mM elution, for Yy1 and Nfatc4 in the rs7647481 200 mM elution (Figure 22B), thus in the fractions with the clearest allelic protein-DNA binding after enrichment.

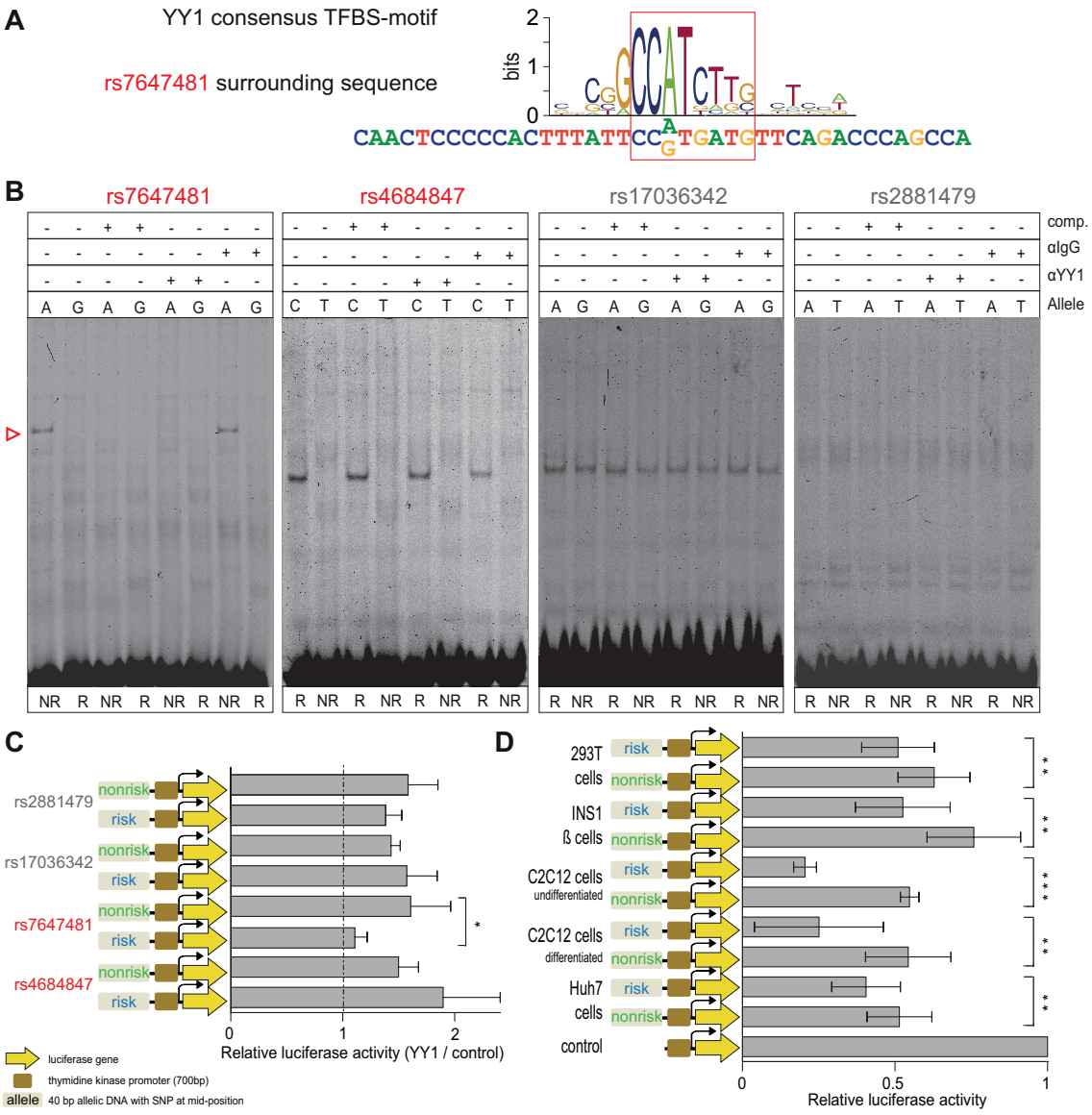


Figure 27. rs7647481 nonrisk allele-specific binding and transcriptional activity of the transcription factor YY1 inferred from proteomics analysis.

(A) The rs7647481G risk allele abrogates the core of a YY1 consensus binding site (MatBase Matrix Library 9.1, Genomatix. Munich, Germany). (B) Competition and supershift EMSA experiments using risk (R) and nonrisk (NR) allele-specific Cy5-labeled of the predicted *cis*-regulatory (red) and non *cis*-regulatory (grey) variants reveal a specific binding of YY1 at the rs7647481 nonrisk allele. Competition (comp.) assays using 33-fold excess of unlabeled YY1 probe and supershift assays by adding anti-YY1 (αYY1) or lgG (αlgG) isotype control antibody, respectively. (C) Reporter assays in 293T cells with constructs harbouring the risk and nonrisk allele of predicted *cis*-regulatory (red) and non *cis*-regulatory (grey) variants reveal specific activation from the rs7647481 nonrisk allele upon YY1 overexpression. (D) rs7647481 non-risk allele-specific activation of reporter gene activity in

293T-cells, INS-1 cells, C2C12 (undifferentiated. differentiated) myocytes and Huh7 hepatocytes. Reporter assays with luciferase constructs the respective allele at midposition as indicated. Mean ± SD from five to seven experiments. *P*-values from paired *t*-test with $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$.

7.3.6 YY1 drives transcriptional activity at the rs7647481 nonrisk allele of the *PPARG* locus

Confirming the mass spectrometric identifications and GO-term analysis, the common rs7647481G risk allele abrogates the core of a YY1 consensus transcription factor binding site (TFBS) (Fig. 27A). The protein-DNA interaction at the rs7647481-adjacent region was efficiently blocked in competition and supershift EMSA experiments by 33-fold molar excess of unlabeled YY1 consensus binding sequence or by pre-incubation with a YY1 specific antibody (Fig. 27B, left panels), while protein binding was not affected at all other tested SNP-adjacent regions, including the *cis*-regulatory variant rs4684847 (Fig. 27B, right panels). Further confirming the specificity of YY1 binding at the rs7647481-adjacent region, competition with unspecific competitor oligonucleotides (consensus MyoD myogenic regulatory factors, consensus CdxA chicken homeodomain protein, and scrambled control sequence) did not affect the allele-specific protein binding (Fig. 28. In reporter gene assays with luciferase constructs of all tested variants transfected into 293T cells, overexpression of the transcription factor YY1 revealed a significant activation of the rs7647481 nonrisk as compared to the risk allele ($P = 0.003$), whereas activity from the *cis*-regulatory variant rs4684847 and both non *cis*-regulatory variants was not affected (Fig. 27C). The rs7647481 nonrisk variant also increased transcriptional activity in different cell types significantly, i.e. by 1.2-fold in 293T cells, 1.4-fold in INS1 β-cells, 2.7-fold in C2C12 myoblasts, 2.2-fold in C2C12 myocytes, 1.3-fold in Huh7 hepatocytes ($P < 0.01$, Fig. 27D) and 1.5-fold in 3T3-L1 adipocytes [98]. Overall, our data establish the transcription factor YY1 to bind at the rs7647481 nonrisk allele and support a role in transcriptional gene regulation.

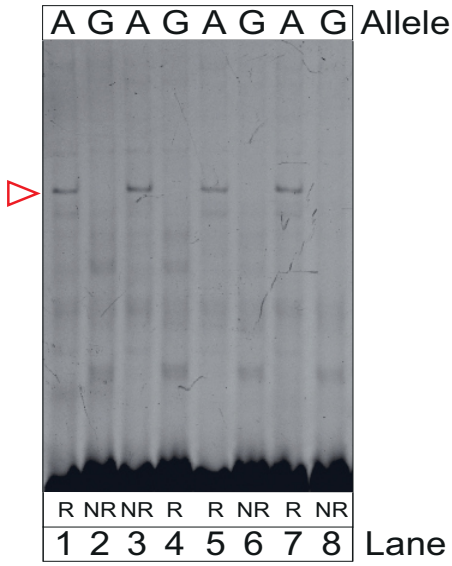


Figure 28. Competition EMSA using unspecific oligonucleotides.

In Competition EMSAs, Cy5-labeled oligonucleotide probes for the predicted *cis*-regulatory variant, rs7647481 were incubated with a 33 fold molar-excess of unlabeled non-allele-specific MyoD, CdxA, and a non-allele-specific scramble oligonucleotide competitor probes as indicated. Competition EMSA results support the specific binding of YY1 at the rs7647481. A red triangle indicates the rs7647481allele-specific band, which was not altered in signal intensity by addition of unlabeled probes MyoD, CdxA and scramble. Predicted *cis*-regulatory SNPs rs7647481 without competition (lanes 1-2), with MyoD (lanes 3-4), with CdxA (lanes 5-6) and with scramble (lanes 7-8) competitor probes. R = risk allele, NR = nonrisk allele.

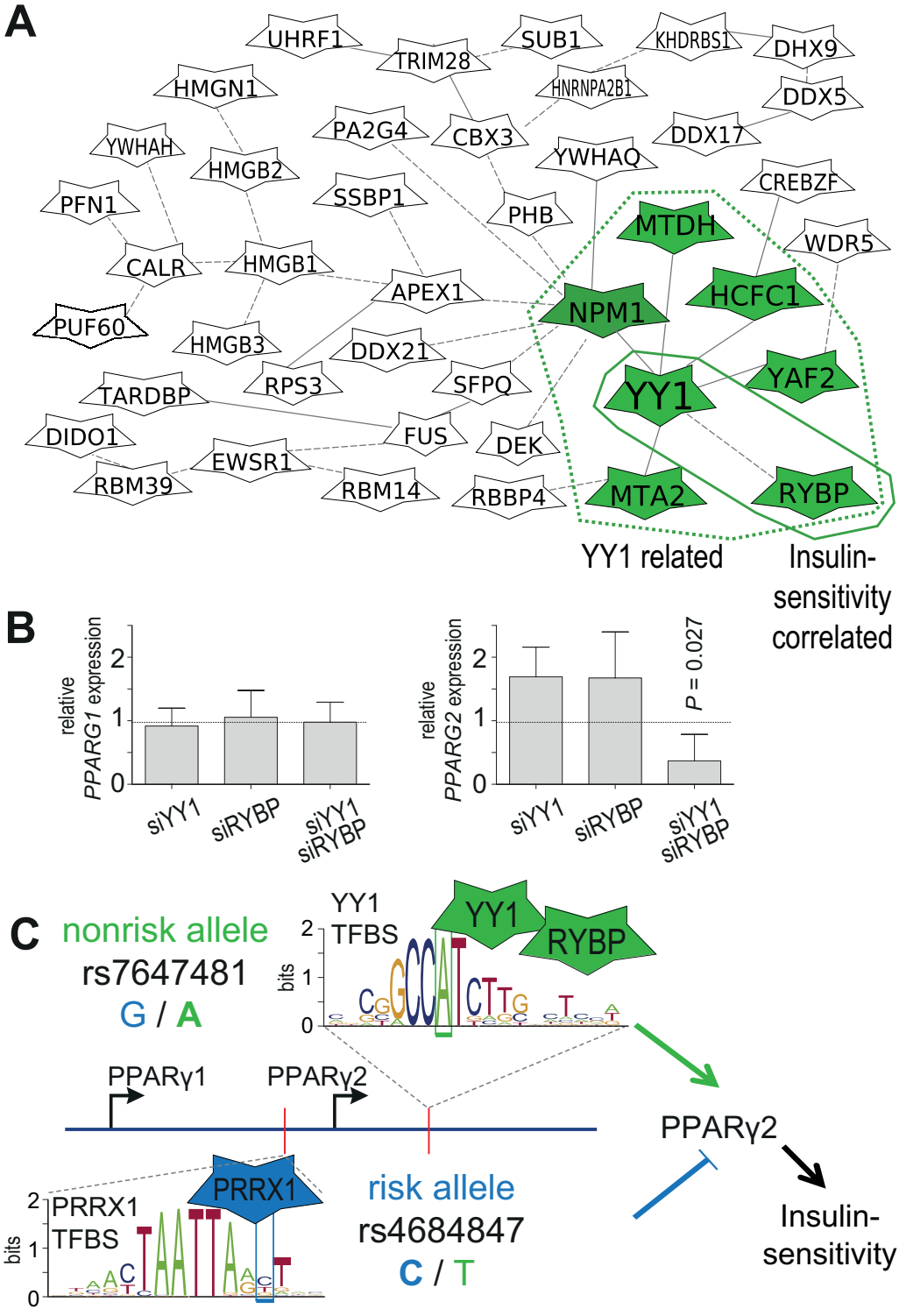
7.3.7 Cocitation interaction network reveals YY1 related coregulators

Metabolic homeostasis is largely regulated at the transcriptional level through the coordinated interaction between transcription factors, coregulators, and the basal transcriptional machinery [265]. Our pull down of functional protein-DNA binding complexes offers the opportunity to identify protein-DNA interactions, as demonstrated by the identification of transcription factors and additional numerous co-eluting transcriptional coregulators (Supplementary table 4). To gain insight into the underlying protein-protein interactions, we assessed literature co-citations of the prioritized transcription factors PRRX1, YY1, and NFATC4 with identified coregulators. We found a significant enrichment of transcriptional coregulators co-cited with YY1 (P = 1.56 x 10^{-5}, fishers exact test, Online Methods), i.e. RING1 and YY1 binding protein (*RYBP*), YY1 associated factor 2 (*YAF2*), prohibitin (*PHB*),

nucleophosmin (*NPM1*), host cell factor C1 (*HCFC1*), metastasis associated 1 family (*MTA2*), DEK oncogene (*DEK*), and high mobility group box 2 (*HMGB2*) (see Methods). No significant enrichment of co-cited proteins was discovered for NFATC4 or PRRX1. Visualizing GePS-tool (Genomatix, Munich) annotated gene-gene interactions of the transcription factor YY1 with all proteomics-inferred proteins annotated as cofactors reveals a network connecting YY1 with *RYBP, NPM1, YAF2, MTA2, HCFC1* and metadherin (*MTDH*) (Fig. 29A). *NFATC4* was found connected with calreticulin (*CALR*) and none of the identified cofactors was connected with *PRRX1* (Fig. 30).

Figure 29. Interaction network analysis of YY1 with cofactors infers RYBP contribution to nonrisk allele specific effect on insulin-resistance.

(A) Interaction network of the YY1 transcription factor identified at the rs7647481 non-risk with all transcriptional coregulators identified in the same label-free proteomics analysis. Associations by co-citation (—) or expert curation (---) from GePS tool analysis (Genomatix. See Methods). Proteins with direct interaction to the transcription factor YY1 (green ---) and with positive correlation of adipose mRNA levels to insulin-sensitivity (green —) are shown (Table 27). (B) *PPARG1* and *PPARG2* mRNA expression levels measured by qPCR (standardized to GAPDH) in SGBS preadipocytes treated with different siRNAs for 72h labeled as siYY1. siRYBP or siYY+siRYBP / siNT (non-targeting control). Mean ± SD from five to eight experiments. *P*-values from one sample *t*-test. (C) Impact of nonrisk and risk allele identified proteins on the *PPARG* locus phenotype insulin-resistance. The rs7647481 nonrisk A-allele promotes YY1 binding contributing to induced transcriptional activity and by interaction with RYBP to increased insulin sensitivity. The rs4684847 C-risk allele promotes binding of the *PPARG* suppressor PRRX1 and thereby increases insulin-resistance.


A
UHRF1
TRIM28
SUB1
KHDRBS1
DHX9
HMGN1
HNRNPA2B1
DDX5
PA2G4
CBX3
DDX17
YWHAH
HMGB2
YWHAQ
CREBZF
PFN1
SSBP1
PHB
MTDH
WDR5
CALR
HMGB1
APEX1
NPM1
HCFC1
PUF60
HMGB3
DDX21
YAF2
RPS3
SFPQ
YY1
TARDBP
DIDO1
FUS
DEK
EWSR1
RBM39
RBM14
RBBP4
MTA2
RYBP
YY1 related
Insulin-sensitivity correlated
B
relative PPARG1 expression
siYY1
siRYBP
siYY1 siRYBP
relative PPARG2 expression
P = 0.027
C
nonrisk allele
rs7647481
G / A
YY1 TFBS
bits
PPARγ1
PPARγ2
PRRX1
PRRX1 TFBS
risk allele
rs4684847
C / T
PPARγ2
Insulin-sensitivity

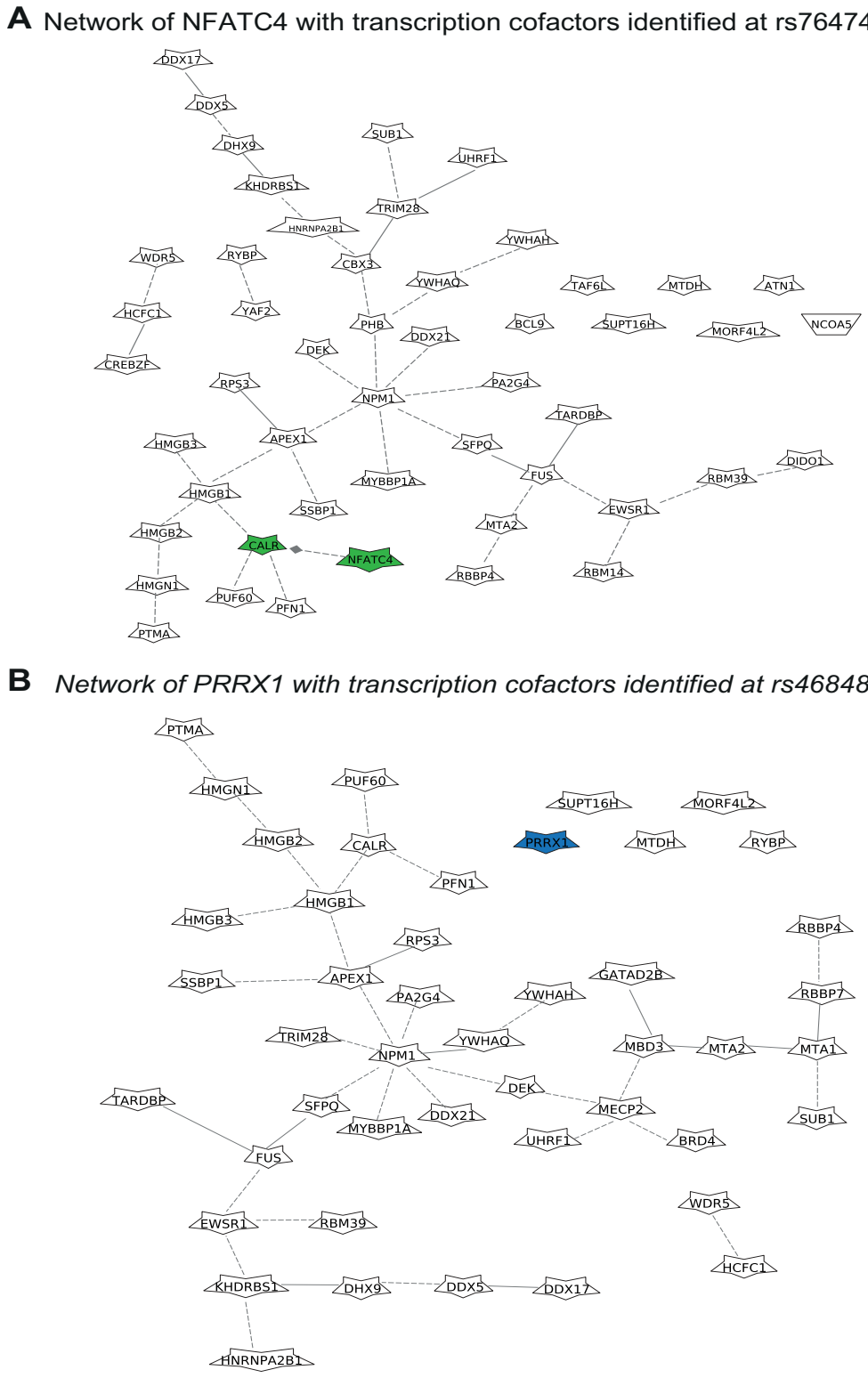


Figure 30. Interaction network analysis of NFATC4 and PRRX1 with LC-MS/MS identified cofactors.

Interaction network of the NFATC4 transcription factor identified at the rs7647481 nonrisk *(A)* and the PRRX1 transcription factor identified at the rs4684847 risk allele *(B)* with the respective set of transcriptional coregulators identified in the same label-free proteomics analysis. Associations by co-citation (—) or expert curation (---) from GePS tool analysis (Genomatix, Online Methods). Green = NFATC4 and proteins with direct interaction, blue = PRRX1 and proteins with direct interaction.

7.3.8 Allele-specific correlation of transcription factor and cofactor expression levels in adipose tissue with insulin resistance

Finally, we assessed if the risk and nonrisk allele-specific proteomics findings can be related to disease pathophysiology. The minor nonrisk allele of the *PPARG* locus (tagSNP rs1801282 Pro12Ala) was repeatedly associated with improved insulin-sensitivity in numerous studies [163,164]. A coordinated regulation of *PPARG* expression by YY1 and co-identified coregulators at the rs7647481 nonrisk allele may contribute to the phenotypes associated with the *PPARG* locus. Here, in adipose tissue samples of nonrisk allele carriers we observed both, a confirmative age- and age/BMI-independent negative correlation of total *PPARG* mRNA levels with the insulin resistance measure HOMA-IR (β = -6.25, P = 2.18 x 10-4; β = -3.26, P = 0.05, respectively) as compared to risk allele carriers (β = 0.23, P = 0.89; β = 0.11; P = 0.92, respectively, Table 27). For adipose tissue mRNA expression levels of the coregulator RYBP, identified by proteomics and reported to interact with the YY1 transcription factor [420], we found a negative age- and age/BMI-independent correlation with HOMA-IR in individuals carrying the nonrisk (β = -5.71, P = 1.15 x 10-3; β = -3.38, P = 7.04 x 10^{-3}; respectively) as compared to risk (β = 0.57, P = 0.67; β = 0.16, P = 0.85, respectively) allele. For none of the other proteins identified at the rs7647481 and co-cited with *YY1* (Fig. 29A) an allele-dependent correlation was observed (data not shown). While we found no significant correlation of *YY1* mRNA levels with HOMA-IR in the small available data set, a confirmative direction of beta values was observed (Table 27). Finally, we also found a positive correlation for adipose mRNA expression levels of both, *YY1* and *RYBP* with the insulin-sensitizing transcription factor PPARG from both alleles (data not shown). Assessing the effect on endogenous mRNA expression levels in SGBS preadipocytes, we found that knockdown of YY1 or RYBP alone was not sufficient to reveal a significant effect on *PPARG1* and *PPARG2* expression. Notably, simultaneous knockdown of *YY1* and *RYBP* revealed a significant two-fold reduction for endogenous mRNA expression levels of the

insulin-sensitizing *PPARG2* isoform (P = 0.027) (Fig. 29B), supporting the importance to encounter both, transcription factors and related cofactors. Insulin sensitivity may be increased by a rs7647481 nonrisk allele-specific, coordinated action of the transcription factor YY1 and the cofactor RYBP activating the expression of the insulin-sensitizing transcription factor PPARG, in addition to the previously reported rs4684847 risk allele specific inhibition of *PPARG* expression by PRRX1 [98] (Fig. 29C).

			HOMA_IR		
Gene	**allele**	**Adj**	***β***	**SE**	***P*-value**
PPARG	all	-	-3.47	1.19	**6.03×10^{-3}**
		a	-3.52	1.19	**5.48×10^{-3}**
		a.b	-1.57	0.87	0.79
	nonrisk	-	-6.27	1.33	**2.27×10^{-4}**
		a	-6.25	1.32	**2.18×10^{-4}**
		a.b	-3.26	1.50	**0.05**
	risk	-	0.25	1.68	0.88
		a	0.23	1.73	0.89
		a.b	0.11	1.10	0.92
YY1	all	-	0.07	1.59	0.96
		a	-0.01	1.69	1.00
		a.b	-0.25	1.08	0.82
	nonrisk	-	-2.87	2.63	0.29
		a	-3.97	2.79	0.17
		a.b	-2.33	1.70	0.19
	risk	-	2.18	1.77	0.23
		a	2.36	1.86	0.22
		a.b	1.15	1.23	0.36
RYBP	all	-	-1.98	1.11	0.084
		a	-1.98	1.11	0.083
		a.b	-1.14	0.73	0.13
	nonrisk	-	-5.52	1.49	**1.89×10^{-3}**
		a	-5.71	1.44	**1.15×10^{-3}**
		a.b	-3.38	1.09	**7.04×10^{-3}**
	risk	-	0.55	1.31	0.68
		a	0.57	1.31	0.67
		a.b	0.16	0.84	0.85

Table 27. Risk and nonrisk allele specific correlation of adipose tissue *PPARG*, *YY1* and *RYBP* mRNA expression with the T2D trait insulin-resistance.

Gene expression was measured in adipose tissue from a lean / obese patient cohort (38 subjects. mean ± SD 24.2 ± 9.1 kg/m^2). rs7647481 and rs4684847 risk allele and nonrisk allele genotypes were determined by Sequenom-assay. *Nonrisk.* subjects heterozygous or homozygous (n = 18) for the rs7647481A (YY1/RYBP binding) and rs4684847T nonrisk allele; *risk.* subjects homozygous (n = 20) for the rs7647481G and rs4684847C risk allele. *P*-values and *β*-estimates from linear regression analysis of *PPARG*. *YY1* and *RYBP* mRNA expression levels with insulin-resistance measure HOMA-IR (homeostasis model assessment of insulin resistance) are shown. Adj = correlations without adjustement (-). age (a) or age and BMI adjusted (a. b). The raw data was obtained from Prof. Peter Arner's laboratory (Department of Medicine, Huddinge, Karolinska Institutet, Stockholm, Sweden) and the data analysis was performed by Sophie Molnos (Institute of Epidemiology II, Helmholtz Zentrum München, Neuherberg, Germany). (

8. Discussion and conclusion

8.1. Requirement of a proteomics-based high sensitive approach for unraveling molecular mechanisms underlying genotype-phenotype associations

GWAS revealed numerous risk loci associated with common traits [18,85]. The majority of those variants are located in non-coding DNA regions and have been suggested to affect transcriptional regulation [55,100,259–262]. Recent technological advances such as ChIP-seq, DNase-seq, FAIRE-seq [100,254,414,421,422], fine mapping [423] or novel bioinformatics approaches [98] have enabled the large-scale identification of *cis*-regulatory, potentially diseases-causing variants within complex loci [98,415,421]. However, to demonstrate the specific functions of those variants remain as a formidable challenge in human genetics as such *cis*-regulatory elements are present at low abundance in genomic regions [424], and most of these genetic variants are located in non-coding regions [55,100,172], which make it difficult to unravel the mechanism linking the genetic variants to diseases or traits. Thus, signals emerging from GWAS have been rarely traced to the precise molecular mechanisms by which *cis*-regulatory variants may increase or decrease an individual's susceptibility to disease. Recently, evidence continues to accumulate linking gene regulation with *cis*-regulatory variants in a variety of diseases or traits by analyzing chromatin structure and binding of transcription factors [98,100,266,269,312,327,425], which provide a comprehensive view of human gene regulation.

The current field of human genetics has increasingly shifted its attention from disease gene identification to following through on next steps, most importantly pursuing the biological mechanisms linking genotype to phenotype [426]. Moreover, there is increasing evidence of studies demonstrating that gene-regulation depends on complex protein-DNA and protein-protein interactions [98,266–268,421]. So far, only few proteomic studies have carried out successful transcription factor identification at *cis*-regulatory variants [98,312,314,418,421], indicating that the allelic status of the DNA has a functional impact on gene expression. These results were recently confirmed and expanded by the ENCyclopedia Of DNA Elements (ENCODE) consortium [100]. However, in most cases identification of such complex *trans*-acting protein networks at *cis*-regulatory variants remains challenging. Recent bioinformatics approaches such as PMCA assessing the occurrence of conserved patterns of TFBS in *cis*-regulatory modules (CRMs) within the genomic region flanking a non-coding variant [98] is becoming increasingly important tool in combination with the epigenetic state (e.g., open

chromatin, histone marks) [353] for meeting this challenge. However, such TF motif based models to identify TF binding have not been sufficiently well calibrated to predict the functional impact of sequence on binding. Moreover, the sequence elements (motifs) and the epigenetic states such as open chromatin, DNA methylation state, histone marks represent a combinatorial regulatory code that remains still poorly understood in a global genomic scale [427]. Thus, unraveling the mechanisms underlying genetic variations influencing complex diseases might be even more challenging in human genetics because most gene variants associated with complex diseases such as T2D have only low or modest effects [408,428], and T2D susceptibility is closely associated with several lifestyle and environmental factors [429].

This study highlighted three points: *(i)* the development of a label-free quantitative DNA protein interaction approach enabling the identification of allele-specific protein complexes, *(ii)* the *in-depth* analysis of *cis*-regulatory variants at the *PPARG* locus and their biological role on gene expression, supported by the PMCA-based prediction of *cis*-regulatory activity [98] and the publically available data on epigenetic marks of regulatory regions [353], and *(iii)* further verification of the integrative framework analysis for *cis*-regulatory prioritization of non-coding variants [98] by applying the label-free quantitative proteomics experimental approach.

8.2 Development of a label-free quantitative protein-DNA proteomics, coupling affinity chromatography with LC-MS/MS

8.2.1 Analysis of allele-specific protein binding to *cis*-regulatory *PPARG*, *FTO*, *TCF7L2* T2D or obesity risk variants

It should be no surprise that only a small portion of genetic variants identified by GWAS is within protein-coding genes. A large number of variants associated with complex diseases are non-coding, which are expected to exert *cis*-regulatory effect on gene expression [55,100,416,430]. The most widely used approach to predict regulated genes is the expression quantitative trait loci (eQTL) mapping. This approach uses massive-scale parallel expression pattern to identify statistical associations between genotypes and gene expression in populations with a heterogenous genetic background. However, it is difficult to detect eQTLs with small effect sizes [431]. Most eQTL analyses are limited to gene expression measured in a single tissue type and not allowed to study more inaccessible tissues such as the brain [432]. Most current eQTL

mapping studies carry out each gene expression as one single trait, which ignore the trait-trait interaction completely [433]. Additionally, such studies often focus on the association between genetic variants and levels of whole-gene expression, without concerning e.g. isoforms resulting from alternative mRNA processing [434].

Of note, eQTL analysis allows to predicting the target genes of *cis*-regulatory variants and provides only indirect evidence of associations between genotype and gene transcription. In eQTL analysis, associations between alleles and target genes do not require knowledge of functional mechanisms (reviewed in Edwards et al. 2013 [263]). Moreover, mRNA expression-centric studies such as eQTL do not consider protein–DNA binding interactions, which are often not available for a given tissue and in the right biological context, making difficult to decipher a mechanism of how genetic variants confer a disease risk. Thus, more direct functional assays such as 3C and its derivatives are necessary for elucidating the mechanistic relevance to the disease or trait (see chapter 4.1). For example, Carbon-copy 3C (also known as 5C) 5C is widely applicable for identifying long-range chromatin interactions across large genomic regions [435]. Another variation of 3C, Chromatin interaction analysis by paired-end-tag sequencing (ChIA-PET) is useful for *de novo* detection of global chromatin interactions bound by a specific protein [436]. However, such methods are facing several disadvantages including time-consuming, cost-intensive procedure and limited amount of available cell material [264]. The approach presented here focuses not only on the detection of *cis*-regulatory variants, but also on the identification of allele-specific binding proteins, thereby understanding the mechanism underlying genetic variants. This study was inspired by the previous findings [98]. Claussnitzer et al. recently introduced a novel method for the prioritization of causal variants in LD regions detected by GWAS based on conserved co-occurring TFBS patterns within CRMs. Indeed, several variants were shown to occupy more or less allelic affinity to specific proteins, proven by EMSA and reporter gene assays under certain conditions [98]. These results helped to narrow down the list of possible loci and causal variants. Finally, several variants were selected at the *PPARG*, *FTO* and *TCF7L2* loci associated with T2D in this study.

A variety of studies have reported that the *PPARG* locus is robustly associated with T2D and insulin-sensitivity [163,164,170,180,364]. Sugii et al. reported that PPARG activation in adipocytes improved whole-body insulin sensitivity to a similar degree as with systemic TZD treatment. Additionally, PPARG activation enhanced adipokine profiles and reduced serum

lipids, high fat diet-induced inflammation and dramatically lowered circulating insulin levels [121,144]. The tagSNP rs1801282 (Pro12Ala C > G) in the *PPARG* gene is well known to be associated with BMI, fasting insulin and insulin sensitivity. The substitution of proline to alanine at the codon 12 results in reduced PPARγ2 function via decreased binding of the Ala variant to the PPAR response element and subsequent lower transcription activity of PPARγ [163,437]. Notably, the minor nonrisk G allele of the rs1801282 was repeatedly associated with improved insulin-sensitivity in several large-scale well-powered population studies and meta-analyses [108,163,164,168,170,438]. On the other hand, Heikkinen et al. demonstrated using a Pro12Ala knockin model that Ala/Ala mice on chow were leaner and more insulin sensitive than Pro/Pro mice, but in high-fat feeding such effects were eliminated and led to increased weight gain [180]. In line with this result, the Ala12 allele showed to gain more weight than the Pro12 allele in obese patients, which are associated with an increased risk of T2D [181,439], whereas the Ala12 allele is protective against T2D in nonobese subjects [163,170]. These results suggest that the Pro12Ala variant functions as an important modulator in metabolic control which strongly depends on the metabolic context [180]. Moreover, the minor 12Ala allele blunts the transcriptional activity of the insulin-sensitizing PPAR-γ2 transcription factor, however is paradoxically associated with enhanced insulin sensitivity in humans [108,163,164,168,170,437,438], suggesting the recently proposed 'multiple enhancer variant' hypothesis [440]. This hypothesis supposes that several causal variants within a given locus cooperatively affect gene expression and confer susceptibility to common traits. Thus, this study aimed the identification of further *cis*-regulatory variants at the *PPARG* locus affecting *PPARG* gene expression. Indeed, Claussnitzer and colleagues previously reported another *cis*-regulatory variant, rs4684847 at the *PPARG* locus and its allele-specific binding transcription factor PRRX1 (identified by AC-LC-MS/MS as described in this study) with their adverse effect on *PPARG*2 expression, lipid metabolism and systemic insulin sensitivity [98]. This finding would give a more comprehensive view of gene expression, DNA-protein interactions, protein-protein networks and signaling pathways in metabolic, genetic, or environmental context.

Among the 24 non-coding variants at the T2D associated *PPARG* locus ($r^2 \geq 0.7$ with *PPARG* tagSNP rs1801282, 1000 Genomes [172]), the PMCA approach predicted six variants as complex, i.e. predicted a *cis*-regulatory function. Out of the six variants, only the rs4684847 showed an overlap with cell stage-dependent histone H3-lysine 27 acetylation (H3K27ac). Moreover, only the rs4684847 showed direct overlap to a distinct homeobox

TFBS matrix, inferred from PMCA [98]. The rs4684847 (Intron C > T) is located 6.5 kb upstream of the PPARG2-specific promoter in complete LD with the rs1801282 [98]. The rs4684847 has not been well studied. A limited number of reports documented its significant association with all-cause mortality and cancer-related "mortality outcome" in a study of ~10,000 individuals [441,442]. Moreover, the rs4684847 showed statistically significant age-adjusted associations with both, baseline body mass and blood pressure. Of note, after adjustment for age, subjects carrying the nonrisk TT allele of rs4684847 were significantly more likely to have a higher BMI than individuals carrying the risk CC allele. Furthermore, subjects carrying the rs4684847 CT or TT genotypes were more likely to be prehypertensive or hypertensive at baseline compared to subjects carrying the CC genotype [167]. However, these reports may not be enough to uncover the mechanism underlying between the rs4684847 and T2D risk. Thus, EMSA assay was performed to observe allelic differential binding patterns of proteins at the rs4684847 in different cell lines including 293T, Huh7, INS-1, 3T3-L1 and HIB 1B (3T3-L1 and HIB 1B cell lines, published in the previous study [98]). The risk C allele of rs4684847 exhibited enhanced binding activity of proteins relative to the nonrisk T allele. The gene expression of *PPARG* is induced not only early during adipocyte differentiation, but also continues at a high level in mature adipocytes [130,443,444]. Since the importance of PPARG in adipocytes was demonstrated in a variety of studies, further experiments were performed mainly in adipocytes. There are some difficulties to study adipogenesis *in vivo* so that several cell line models have been established to study the cellular and molecular events in adipogenesis *in vitro* [336]. Mouse white pre-adipocyte cell line, 3T3-L1 [445], mouse brown pre-adipocyte cell line, HIB 1B [335] as well as human pre-adipocyte cell line, SGBS [446] have been well characterized in numerous studies and provide useful *in vitro* models for understanding the molecular basis during the adipogenic process. Notably, the allele-specific binding of proteins at the rs4684847 C risk allele was increased compared to the T nonrisk allele during differentiation of HIB 1B, which was also shown in 3T3-L1 in repeated EMSA experiments. White adipose tissue (WAT) stores excess energy as triglycerides [447], whereas the main role of brown adipose tissue (BAT) is to dissipate chemical energy as heat whereby the energy derived from fatty acid oxidation is used for the generation of heat due to mitochondrial uncoupling [448]. Also, it's well established that the *PPARG* gene is expressed abundantly and equally in white and brown adipocytes, and the *PPARG* gene expression is essential for the differentiation of both, white and brown adipocytes. Now, it is clear that several dominant transcriptional regulators control brown

adipocyte development and function, including peroxisome proliferator activated receptor gamma coactivator 1a (PGC-1a), Forkhead box C2 (FoxC2) and PRD1-BF-1-RIZ1 homologous domain containing protein-16 (PRDM16) (reviewed in Ohno et al. 2012 [395]). Genetic loss of PGC-1a in mice showed apparent disruption of cold-induced adaptive thermogenesis function in BAT [449]. Also, brown adipocytes lacking PGC-1α exhibited a blunted induction of thermogenic genes in response to cyclic adenosine monophosphate (cAMP) [450], suggesting that while PGC-1α is a crucial regulator of adaptive thermogenesis. A variety of research studies demonstrated, there exist two different types of brown adipocytes, "brown", and ''brite" or ''beige" adipocytes. After WAT of adult animals is exposed to chronic cold or β-adrenergic stimulation, distinct type of UCP1-positive adipocytes are found sporadically there, which are referred to as ''brite" or ''beige" adipocytes. These adipocytes are inducible brown-like adipocytes containing some biochemical and morphological characteristics of classical brown adipocytes such as presence of multilocular lipid droplets. Activation of PPARγ by synthetic ligands was shown to induce a brown fat-like gene program in WAT. Such ligands act directly by binding to and activating PPARγ and PPAR-response elements (PPREs) on the promoter and/or enhancer of BAT-selective genes. However, many questions remain to be answered. Of note, overexpression of PPARγ in white adipocytes does not lead to a white-to-brown adipose tissue conversion. The white-to-brown adipose tissue conversion takes multiple days by stimulation with PPARγ ligands, which was expected to be formed within hours. Taken together, the white-to-brown adipose tissue conversion seems to be achieved through PPARγ ligands in closer detail. In addition, the occurrence of inducible-brown adipocytes in WAT was shown to be associated with a protection against obesity and metabolic diseases in rodent models (reviewed in Ohno et al. 2012 [395]). These studies indicate an importance of understanding mechanisms shared by both, white and brown adipocytes, and by which environments stimulate the induction of brown adipocytes in WAT. In regard to this, the identification of allele-specific binding proteins at the variants of the *PPARG* locus may provide deep insights into not only the transcriptional regulation of *PPARG* gene, but also the understanding adipocyte differentiation.

GWAS has revealed the association between the *FTO* gene and obesity in 2007 [107–109,205]. Various independent follow-up studies confirmed these findings in different populations, and demonstrated that there are strong associations between *FTO* variants and BMI, and subsequently T2D [164,198–202,206]. The rs1421085 at the *FTO* locus was predicted as *cis*-

regulatory by two different analyses [98,376]. Subsequently, EMSA and reporter gene assays confirmed the allele-specific binding of proteins at the rs1421085 [98]. The rs1421085 is located in the first intron of the *FTO* and in strong linkage disequilibrium (pairwise $r^2 > 0.97$) with the rs9939609, which shows the strongest association with BMI in several studies. In a few studies, the rs1421085 risk C allele was shown to be associated with increased body weight in different populations [199,204]. However, none of these studies uncovers the molecular and pathophysiological mechanisms by which the rs1421085 might impact on weight gain. The *FTO* gene is expressed ubiquitously in a variety of tissues including adipose tissue, liver, pancreas and skeletal muscle, with the highest expression in hypothalamus [107,109,188,189,239,451,452]. The mRNA expression of *FTO* in adipose tissue was greater in obese than normal weight subjects, which was not influenced by the *FTO* rs9939609 genotype [190]. It is also in line with the observation that the expression of *FTO* was moderately increased in adipocytes compared with preadipocytes and was substantially reduced in white adipose tissues of obese, indicating that FTO might play a role in adipocyte function, but not in adipogenesis [452]. In contrast, Tews et al. demonstrated that the expression of *FTO* was decreased in adipocytes compared to preadipocytes [189]. In other study, FTO deficient mice resulted in a prominent reduction of adipocyte size [193,197]. Furthermore, FTO-deficient SGBS adipocytes led to the increased expression of uncoupling protein 1 (UCP-1), inducing a brown adipocyte phenotype [197]. The cellular functions of FTO may be cell-type specific. In the fasting state, *FTO* expression is increased in WAT and hypothalamic neurons, whereas it is decreased in BAT (reviewed in Pitman et al. 2012 [453]). Additionally, FTO knockdown in SH-SY5Y neuronal cells resulted in increased ATP concentrations, and decreased phosphorylation of AMP-activated protein kinase (AMPk) and Protein kinase B (Akt). In contrast, FTO knockdown in 3T3-L1 adipocytes exhibited decreased ATP concentration, and increased AMPk and Akt phosphorylation [453]. Otherwise, FTO functions in liver and pancreas are largely unknown. Recently, it was reported that FTO alters leptin action and glucose homeostasis in liver as a consequence of the dual effect on leptin-induced signal transducer and activator of transcription 3 (STAT3) phosphorylation [454]. FTO protein may play a hitherto unrecognized role in the control of first-phase insulin secretion in pancreatic β-cells [455]. Therefore, in this study the allele-specific binding patterns of proteins at the rs1421085 were first characterized in different cell lines including 293T, Huh7, 3T3-L1 adipocytes, INS-1 and adult mouse brain tissue in order to explore the mechanisms of its transcriptional regulation by the rs1421085. Interestingly, two allele-specific bands at each

allele (T/C) were observed in EMSA experiments. Consistent with the result from the previous study [98], the most intensive allele-specific signal comes from the T nonrisk allele. This result might be correlated with the fact that the rs1421085 risk C allele (P = 0.0015, effect size = 0.0056) was shown to be significantly associated with increased BMI in Korean population [204]. In addition, the first allele-specific band (with a stronger binding at the risk C allele) was observed only in 293T cells, adult mouse brain tissue and Huh7 cells, whereas the second allele-specific band (with a stronger binding at the nonrisk T allele) appeared in all cell lines and tissue analyzed including 3T3-L1 adipocytes. As mentioned above, the association between *FTO* expression and adipocyte differentiation was investigated in several studies, but with different results. Taken together, the second DNA-protein complex could be involved in adipogenesis. To explore this, the binding properties of proteins at the rs1421085 need to be further characterized during both, white and brown adipocyte differentiation in EMSA assays.

A number of recent studies have confirmed the findings by Grant et al. [76], demonstrating strong association between intronic variants in *TCF7L2* and T2D susceptibility [235,239–250]. However, only a few studies were undertaken to explore the functionality of these variants and the mechanism by which these may exert their effects, which still remain unknown in the vast majority of situations. The rs7903146 at the *TCF7L2* locus was predicted as a *cis*-regulatory variant, inferred from selective epigenetic marks state in human islets [254] and PMCA analysis [98]. In a variety of studies, the rs7903146 at the *TCF7L2* locus was shown to be associated with T2D [76,249,377,378]. Gaulton et al. reported that the rs7903146 risk T allele showed an islet-selective epigenetic marks state in human islets [254]. Moreover, the rs7903146 showed both, allele-specific binding of proteins and luciferase reporter activity in beta-cell line, whereby the risk T allele showed significantly greater enhancer activity than the C nonrisk allele [98,254,255]. The allele-specific regulatory properties for the rs7903146 have been largely limited to pancreatic beta cells [98,254,255], allowing to compare results with published data based on the use of same cell lines. However, it also was shown in other cell lines including myoblasts, neuronal cells [380]. In addition, the *TCF7L2* gene is dominantly expressed in pancreatic beta cells [235]. The rat pancreatic beta cell line, INS-1 is suitable for measuring glucose-stimulated insulin secretion [412] due to the difficulties in generating human pancreatic beta cells [413]. Thus, INS-1 cell line was used as a source for all experiments in this study. In EMSA assay, proteins preferred to bind at the rs7903146 T risk allele compared to

the C nonrisk allele with more intensive signal, confirming the previous data [98]. Here, in EMSA experiments using INS1 beta-cells in total three different allele-specific bands at both alleles were observed which was different from the previous results reported by Pang et al. with one allele-specific band at each allele [251], however using non-T2D related WiDr colon carcinoma cell line, indicating the allele-specific pattern could be varied dependent on conditions, interacting TFs and cell types. Several studies reported that gene regulatory regions in eukaryotes comprising *cis*-regulatory modules (CRMs) carry out their function by integrating the active TFs and their associated co-factors dependent on cell-type and stage of cell development [98,456,457]. In addition, gene regulatory programs are achieved in large part through the cell-type specific binding of TFs, which include direct DNA sequence preferences, DNA sequence preferences of cofactors and local cell-dependent chromatin context [458]. Also, after the discovery of the association between *TCF7L2* and T2D in 2006 [76], numerous studies has been made to demonstrate mechanisms underlying the function of TCF7L2 in Wnt pathway effector in pancreatic beta cells and indicate the beneficial effect of TCF7L2 in pancreatic beta cell for cell proliferation, insulin gene expression and insulin secretion [459], which makes much more sense to use INS-1 as a rat pancreatic beta cell line. However, it is unknown whether the here identified three protein-complexes participate in *TCF7L2* gene regulation, and if so, whether they contribute compensatory or synergistic to the *TCF7L2* expression each other.

8.2.2. Advantages and disadvantages of magnetic versus sepharose beads

Affinity purification is currently the most powerful method available to the downstream processing of proteins and peptides due to their selectivity and recovery [306]. Usually, matrix types used for affinity purification consist of porous support materials such as agarose, polymethacrylate, polyacrylamide, cellulose and silica, which may be available with common affinity ligands already immobilized (e.g. protein A, Cibacron Blue, heparin). Other types of matrix have been developed, such as nonporous supports, membranes, flow-through beads (perfusion media), monolithic supports and expanded-bed adsorbents. Among these, the most popular matrix is beaded agarose (e.g. Sepharose CL-4B; agarose crosslinked with 2,3-dibromopropanol and desulphated by alkaline hydrolysis under reductive conditions), polyacrylamide, and magnetic beads (reviewed in Magdeldin, S. and Moser, A. et al. 2012 [460]).

The type of beads used in affinity purification is a considerable factor as the efficiency and cleanliness of different types of beads may vary according to the cell type and the type of extract used. Dynabeads (Invitrogen) are suitable for nuclear extracts whereas sepharose and agarose beads (GE-Healthcare) have been shown to give lower backgrounds when used with cytoplasmic extracts and whole cell extracts [461]. In this study, two affinity purification strategies were tested for enrichment of allele-specific binding proteins. In the first approach, streptavidin-coated magnetic beads were used with biotinylated oligonucleotides. In the second approach, sepharose beads were employed as an alternative method to the magnetic beads. Since no information on prior work was available, conditions for the purification needed to be determined empirically. Thus, the work started with general assumptions based on the literature [316,342] and consideration of the type of protein being studied or the sequences containing putative binding proteins. During the method development and evaluation process, various conditions for affinity purification were tested by using several genetic variants at the previously introduced loci on a small scale. The purpose of such initial studies is to provide a preliminary analysis, and in particular to determine optimal conditions for further large scale experiments.

Sepharose-based system is a reliable and well-established technique for purification process. The porous beads provide a high surface area for interaction with proteins and allow a large molecular weight range. This technique is suitable for gravity-flow, low-speed-centrifugation, low-pressure procedures, and easy for scale-up and ideal for screening conditions, but can also result in lower the binding capacity and not be autoclaved [462] (www.lifetechnologies.com, www.labome.com). In this study, when using sepharose beads (see chapter 7.2.2.2.2), some difficulties appeared, such as limited scalability, relative high cost, and harmful procedures. Furthermore, sepharose-based purification required larger amount of proteins and relative complicated procedure compared to the magnetic beads-based purification. Apart from complicated procedure, the process was time consuming compared to magnetic beads and dangerous for workers in the labs since manipulation of the sepharose during CNBr activation is complicated due to many steps, and handling CNBr is highly toxic due to the poisonous vapors [463]. In case of the rs1421085 at the *FTO* locus, the affinity purifications were performed using magnetic beads and sepharose beads, respectively, followed by EMSA assays. Comparing the signal intensity of the allele-specific bands in EMSA assay, the use of magnetic beads exhibited more intensive signal than sepharose beads.

Based on this result and the advantages of magnetic beads (see below), further affinity purifications were performed using magnetic beads.

The use of magnetic beads to identify DNA-binding proteins has been already described in several studies [314,315]. After the incubation of proteins with magnetic beads, whole magnetic complex can be easily and rapidly removed from the sample using an appropriate magnetic separator. After washing out the contaminants, the isolated target protein complexes can be eluted and used for further work [306]. Such rapid procedure is very important because *i)* gene-regulatory DNA-binding proteins could be unstable in new physiological conditions, and *ii)* allows to work close to initial conditions established in the EMSA experiment [316]. Moreover, the streptavidin-biotin interaction is the strongest known non-covalent, biological interaction between a protein and ligand. The interaction is very rapid, and the formation is unaffected by wide extremes of pH, temperature, organic solvents and other denaturing agents (www.lifetechnologies.com), which has been exploited in many protein and DNA detection techniques. In addition, magnetic separation is usually very gentle to the target proteins or peptides. Large protein complexes may remain intact during the procedure, which tend to be broken up by traditional column chromatography techniques [306]. This advantage would be very crucial for study of protein-protein interaction networks. The abundance of the target protein complex was monitored in EMSA experiments with the input proteins, washes and purified fractions. However, the purification using magnetic beads on a small scale in this study was not successful to isolate sufficient amount of the target protein complex for further analysis. In case of the rs4684847 at the *PPARG* locus, there was almost no signal in EMSA experiment with eluates collected from protein purification using small amount of protein extracts (125 µg for each allele). In protein purification using four times increased amount of nuclear extracts (500 µg for each allele), approximately less than 20 % of the input proteins was retained in the eluate E400, as assessed by EMSA. These results indicate the importance of input protein amount for successful enrichment of proteins in protein purification. In case of the rs1421085 at the *FTO* locus, only faint trace of the target protein complex remained after the protein purification, which was impossible to measure due to the high background signal in EMSA experiment (125 µg for each allele). There could be several reasons for this, one might be accounted for small amount of proteins. Other possibility could be that the concentration of NaCl in the reaction mixture including input protein was higher than that in EMSA, which could yield inappropriate physiological condition for binding of target proteins.

Other causes for such loss could be the binding efficiency of biotinylated oligonucleotides to the streptavidin coated beads, or biotinylated linker to the original template sequence. Thus, further purification should lead to reduce sample loss, which would result in increased recovery rate.

On the basis of the results from the initial experiments, several factors were concerned for improvement of protein purification. First, the further purifications were performed on a large scale, i.e. using increased amounts of input protein and magnetic beads based on the fact that combination of both, the reduced shearing forces and the higher protein amount might positively influence the isolation process [306]. Second, the NaCl concentration in the input mixture was adjusted to approximately that used in EMSA in order to provide a more favourable environment for protein binding. In cases of *PPARG* and *FTO* loci, the proteins were eluted already with low salt concentration (e.g. in 200 mM NaCl), suggesting that the target proteins are relatively weakly charged, and proteins with weaker ion interactions will be released at lower salt strength [464]. In order to remove disrupting contaminants contributing to such early elution and to minimize loss of the target proteins, the washing and elution steps were started with lower salt concentration in the following purification experiments.

After the optimization, affinity purifications for the *PPARG* rs4684847, rs7647481 and the *FTO* rs1421085 exhibited markedly better enrichment of allele-specific binding proteins, as assessed by EMSAs. Even if the optimal conditions used in EMSA were applied for the affinity purification, there still remain many factors to be tested such as incubation time, beads concentration, temperature and pH which are essential for effective binding as well as elution of the target proteins. In pilot experiments, several conditions were already tested and taken into account for further purification, such as efficient incubation time for beads to the oligonucleotides, and concentration of beads and detergents (data not shown). Also, excessive concentration of oligonucleotides resulted usually in poor binding of proteins and relative high background caused from unspecific proteins. Conversely, in this study too low concentration of DNA led to the low efficient recovery rate of the target proteins. The approximate concentration of poly (dI-dC) supported as a DNA competitor the specific binding of proteins to the oligonucleotides [342], while the relative high concentration hampered the binding of the target proteins to the variants. However, such conditions were not defined in a systematic framework, but assumed based on the previous experience from small scale experiments and from literature for some reasons. First, there are many variable factors in the

development, in particular, the binding condition for unknown target proteins, which have to be considered for each variant and cell type. Second, the "best" combinations of parameters for each subset have to be first tested, and monitored by EMSA and LC-MS/MS. However, one set of such experiment is usually very labour-intensive, time and cost demanding along with the limited measurement capacities in the following LC-MS/MS procedure. Also, the successful adoption of EMSA binding-conditions to the affinity purification binding-conditions at different scales can be challenging. It was often shown that the optimal EMSA conditions were failed to adapt successfully to the binding reaction in affinity purification, which could be due to slight changed buffer composition caused from scale-up, or inevitable difference in binding conditions between both systems. The mixing efficiency of the magnetic beads within the reaction solution containing input proteins could largely influence the efficiency of the protein capturing in a fluidic system. In EMSA assay, proteins bind to the oligonucleotides under free-floating condition, whereas they have to bind to the fixed matrix immobilized with streptavidin in affinity purification. Moreover, coupling DNA to matrix may cause the modification of attached DNA and can affect the DNA interaction with the proteins of interest such as transcription factors [342]. On the other hand, there is a long spacer between the biotinylated probe and the immobilization tag, which reduces steric hindrance effects. In addition, the large surface area of magnetic beads might rather reduce steric hindrance of the targets access to oligonucleotides binding sites requiring less amount of DNA probe, which still maintains the same sensitivity [465]. To minimize the change of buffer composition in scaling-up, the proteins need to be more concentrated. It can also reduce the volume of reaction solution, facilitating the laboratory scale application of magnetic affinity separation techniques. Therefore, such parameters need to be considered in the further work for improvement.

8.2.3 Advantages of label-free quantitative proteomics

A few investigations have attempted to identify allele-specific binding proteins using proteomic technologies and have showed technological advances to some degrees [98,266,267,312]. However, there still remain unsolved problems. Application of the proteomics to whole organisms or specific tissues is valuable when *in-vivo* experiments are necessary to be performed. However, most of the proteomic techniques have been developed based on established cell-lines, and the application to tissue materials has been regarded as too

problematic since the chemical labeling methods such as ICAT [314], or silac labeling method [312] give a limited access to diseases-relevant human tissues. A quantitative proteomic method for the *in vivo* biological studies has been created in some cases, i.e in rats, mice [466] and some tissues such as mouse brain [467], however it still have some difficulties due to the great complexity of the tissue sample and the limitation on metabolic labeling of non-proliferating cells. Moreover, the labeling methods can be time-consuming, have limitations due to high-cost or inefficient labeling [329], may cause artifacts [330], and may be limited by missing data points due to under-sampling. These shortcomings and the lack of simple, but efficient approach suitable for the use of human material have encouraged development of alternative method.

In this study, a sensitive and robust proteomics workflow was developed, which utilises peptide intensity-based label-free quantification on co-registered peptide maps across samples instead of metabolic labeling strategies [312,418] or chemical labeling [314]. When metabolic stable isotope labeling is not suitable, or isotope labeling is insufficient, label-free approach is especially applicable [329]. Unlike the previous labeling methods, the label-free approach has no general dynamic range limitation, whereas isotope-labeled samples usually cause very large errors in ratio determination if fold changes of protein abundance is greater than 20:1 [331,468]. While previous label-free approaches have used so called "spectral counting" as an indirect measurement for peptide abundance, the recent label-free approach is based on direct comparison of peptide intensities across samples [329]. For the LC-MS/MS quantification, several software tools are available such as the *Progenesis* software [344] and the newly developed *Maxquant* software [469,470]. For the use of standard workflows without developing algorithms and pipelines themselves monolithic solutions, *Progenesis* or *MaxQuant* are very suitable tools for fast data analysis [471]. Merl et al. directly compared the label-free approach with SILAC labeling using two different programs (*Progenesis LC-MS* and *MaxQuant*), and verified its accuracy and robustness [329]. For these reasons, *progenesis* (Nonlinear Dynamics) was used in this study.

A robust quantification strategy in proteomics is necessary for identification of allele-specific DNA binding proteins since identification of sequence-specific DNA binding proteins is often interfered by their low abundance or the degeneration of their binding sites. In addition, non-specific binding proteins positively charged to the DNA backbone in high abundance as contaminants compete with low abundant specific DNA binding proteins [418].

Thus, the purification and characterization of proteins at low copy number (ranging between 10^3 and 10^5 molecules per cell) such as transcription factors, protein kinases, and regulatory proteins has been challenging in proteomics. These low-copy proteins will not be observed in the analysis of crude cell lysates without purification procedure. For example, if 1 pmol of pure protein is required for successful LC-MS/MS characterization, a transcription factor (10^3–10^5 molecules per cell) would require isolation from 2×10^7–10^9 cells. Because yield of purification is seldom 100%, it often needs large amounts of starting material, and sensitive analytical methods are essential for successful characterization of proteins [302,403]. Indeed, after scale-up it brought into the more significant fold change and *P*-value, facilitating the prioritization of the candidate proteins with allele-specific binding affinity. Using this strategy, the transcription factor PRXX1 was identified as one of the proteins binding directly at the rs4684847 [98]. Furthermore, another transcription factor YY1 binding at the rs7647481 variant was also found with their coregulators in networks, which will be further discussed later in the chapter (see 8.3). The results proved the high sensitivity of this approach and provide the enormous potential to study complex biological systems in their entirety.

LC-MS/MS analysis of four *PPARG* variants (2 predicted *cis*-regulatory and 2 predicted non *cis*-regulatory) resulted in up to 952 proteins in eluted fractions, which were dominated by non allele-specific binding proteins (see chapter 7.3.4). Similarly, other studies identified 904 total proteins [312] or up to 900 proteins [314], including very small number of proteins separated from the bulk of non-specifically binding proteins. Identification of allele-specific binding proteins against the high number of background proteins is limited in sensitivity of detection by the high complexity and dynamic range of the eluted proteins. Previous reports have reduced such complexity by off-line Strong Cation Exchange (SCX) chromatography [314] resulting in a total of 53 fractions per experiment, or by using SDS-PAGE gels [418] resulting in 6 fractions per experiment. Here, rather step-wise salt elution in protein purification was chosen to control complexity, which offers the opportunity to evaluate the presence of relevant differential allele-specific binders with EMSA assays directly in the eluted fractions. This approach helps to reduce analysis time at the mass spectrometers by pre-selection of relevant fractions and at the same time enables to increase biological replicates (n=3). Since transcription factors and other co-regulators could be among the lowest abundant proteins in cells, detection of relevant differential binders requires sufficient total protein input amounts. Thus, in line with previous reports, this study used 7 mg of nuclear extracts as a best suited

input amount for maximizing detection of differential binders. Himeda et al. have used 70 mg nuclear extract input, and identified three differential binders to the transcriptional regulatory element X (Trex) [314], while Mittler et al. identified 10 differential binders to a 26 bp promoter region of the *ESRRA* gene from 7 mg nuclear extracts [418]. In a recent study by Butter et al., between 1 and 7 differential binders to SNP containing regions of the *IL2RA* gene were identified from only 200 µg nuclear extract input per experiment, however, in this study, concatenated oligonucleotides of unspecified length were used for pull down, which was discussed as reason for the boost in sensitivity [312]. TFBS modularity is essential for protein-DNA binding [98] and using multiple copies of the same DNAs may affect the complexity of correlations between genomic sequence and predicted TF binding. Thus, here DNA oligonucleotides exactly representing the genomic regions of interest without concatenation were used in order not to introduce potential false positive interactions. In case of the rs4684847 and the rs7647481 at the *PPARG* locus, the optimal amount of input proteins was balanced to the maximal output of significantly differential binders in protein purification. Quantitative proteomic analyses of the eluted fractions which contain detectable allele-specific binding, resulted in a significant increase in identification of allele-specific binders. Between 142 and 165 allele-specific binding proteins (fold change > 2 and *P*-value < 0.05) were identified for the *cis*-regulatory variants containing regions from 7 mg input material, which is a considerable number as compared to those from the other previous proteomic studies, as mentioned above [312,314,418].

8.2.4. From enrichment to identification of allele-specific binding proteins

Using an enrichment process followed by the label-free quantitative proteomics, PRRX1 and TF1 were identified as TFs that bind at the rs4684847 of the *PPARG* locus in an allele-specific manner, as assessed by competition and supershift EMSA assays. The paired-related homeobox protein family includes Prrx1 and Prrx2. The Prrx1 gene is alternatively spliced to two proteins, Prrx1a and Prrx1b. Prrx1a and Prrx2 promote transcriptional activation, whereas Prrx1b acts as a transcriptional repressor [472]. Prrx1 expression is restricted to the mesoderm during embryonic development, and both Prrx1 and Prrx2 are expressed in mesenchymal tissues in adult mice [473–476]. Exogenous expression of *PRRX1* is associated with the promoted invasion of glioblastoma cells [477]. Recently, a repressive role in adipogenesis was implicated for PRRX1 by activating transforming growth factor-beta (TGF-beta)

signaling. PRRX1 acts downstream of tumor necrosis factor-alpha to inhibit osteoblast differentiation [478], however, its target genes remain elusive. Unless the *in silico* analysis did not demonstrate the binding site of PRRX1 to the rs4684847 of the *PPARG* locus, the homeobox overlap analysis inferred from PMCA indicated the binding of PRRX1 in close to the rs4684847 C/T. Indeed, the competition and supershift EMSA assays indicated that PRRX1 binds at the rs4684847 C risk allele with greater affinity than the T nonrisk allele. Its role on PPARG2 expression as a novel repressor was further evaluated *in vivo.* In addition, the correlations between *PRRX1* mRNA levels in human adipose tissue and the rs4684847 risk allele with HOMA-IR, BMI, TG/HDL ratio were demonstrated previously by Claussnitzer and colleagues [98]. Contrarily, other allele-specific binding protein, TF1 was predicted to bind to the rs4684847 in an allele-specific manner, inferred from *in silico* analysis (data not shown). This prediction was confirmed by competition and supershift assays. However, the functional roles of the TF1 on *PAPRG* expression, adipogenesis, adipocyte function, or its impact on metabolic phenotypes such as insulin-sensitivity were not evaluated further in this study and have not been reported in literature so far. Thus, TF1 will be interesting candidate to be considered in experimental follow-up studies.

In case of the *FTO* rs1421085, affinity purifications on a large scale were performed using nuclear extracts from adult mouse brain tissue and Huh7 cells. Following LC-MS/MS analyses, different candidate proteins were chosen for further analysis. This selection was mainly made from the literature and bioinformatics analysis because none of these proteins showed either fold change > 2 or significant *P*-value < 0.05, which might be due to a poor enrichment of the proteins possibly caused by the low abundance of proteins enriched in eluates. These results were in line with the result from EMSA assays after affinity purifications showing very faint signals of allele-specific bands. The initial competition EMSA experiment using nuclear extracts from mouse brain indicated that three transcription factors, TF2, TF3 and TF4 are capable to bind to the rs1421085 in an allele-specific manner. Moreover, a TF2 specific TFBS was predicted to be affected by the C risk allele of the rs1421085, however no TFBSs for TF3 and TF4 have been found in the DNA-probe sequence. Thus, TF2 is a good candidate to be involved in formation of the allele-specific protein complex. A second allele-specific protein-complex in EMSA assay (with a stronger binding to the T nonrisk allele than the C risk allele) could contain TF3 and TF4 based on the competition EMSA result. Furthermore, TF3 was previously reported to interact with TF4,

shown by glutathione S-transferase (GST) pull-down assays (note, reference is not indicated as TF2, TF3 and TF4 are unpublished data). Notably, it was also reported that TF4 binds allele-specific to another SNP at the *FTO* locus, which was shown to regulate the expression of the *FTO* gene (reference not indicated, not yet published). However, the preliminary competition EMSA result should be replicated using nuclear extracts from other tissues for which *FTO* gene function was supposed, such as brain [102,110,188], liver [454], or brown adipose tissue [197] and white adipose tissue [189,190,193,452]. Furthermore, further analyses such as supershift and functional assays are required. Identification of allele-specific binding proteins at the rs1421085 will be important in understanding the transcriptional regulation of *FTO* expression and thereby elucidating roles of FTO in the regulation of energy balance. The *FTO* risk alleles are actually located in the first intron of the *FTO* gene, that are close to the transcriptional start site of *RPGRIP1L* (the human orthologue of mouse Ftm), suggesting the possibility of co-regulatory mechanisms between *FTO* and *RPGRIP1L*. Indeed, Stratigopoulos et al. reported that putative overlapping regulatory region within intron 1 of *FTO* contains at least 2 putative transcription factor binding sites for CUTL1 (Cutl-like 1). One of which is included with another *FTO* variant rs8050136 [188], indicating that the association between *FTO* variants and body weight regulation is mediated through changing the expression of both, *FTO* and *RPGRIP1L* [479]. Thus, following the identification of the allele-specific binding proteins, the TF-dependent regulation of *FTO* gene and also its neighbor genes has to be determined, which might be co-regulated. Otherwise, Smemo et al. demonstrated that the obesity-associated variants at the *FTO* locus are not associated with expression of *FTO* gene, but *IRX3* gene in human brains, suggesting that IRX3 is a long-range target gene of obesity-associated *FTO* variants [102]. In turn, these data suggest the possible presence of other target genes of the rs1421085 at megabase distances in a cell or tissue specific manner such as brain tissue, liver, and adipose tissue.

At the *TCF7L2* locus, transcription factors TF3, TF4 and TF5 were chosen as candidates based on LC-MS/MS results. Notably, the *in silico* analysis predicted TF3 to bind more intensive at the T risk allele compared to the C nonrisk allele. Indeed, the preliminary competition and supershift assays indicated that the transcription factor TF3, TF4 and TF5 might be involved in the formation of allele-specific binding protein-DNA complexes (data not shown). Interestingly, TF3 and TF4 were found at both, *TCF7L2* (T2D risk locus) and *FTO* (obesity risk locus) loci, as allele-specific binding TFs, suggesting the common gene

regulatory pathways between T2D and obesity. It is unclear yet whether these three TFs participate in *TCF7L2* gene regulation and signaling pathways involved in insulin secretion. Thus, these experiments have to be replicated, and the functional roles of these TFs should be demonstrated in order to further elucidate the association between the *TCF7L2* rs7903146 variant and the risk to develop T2D. The regulation of *TCF7L2* expression levels have been well demonstrated in pancreas β-cell dysfunction [224,480,481]. Additionally, several studies attempted to address the physiological importance of splice isoforms, and the role of the rs7903146 in altering splicing on adipose tissue, in which the *TCF7L2* gene is well expressed. Mondal et al. reported that certain *TCF7L2* splice forms in subcutaneous adipose tissue are associated with reduced *TCF7L2* gene expression in the rs7903146 risk TT carriers, but overall *TCF7L2* gene expression was statistically not significantly associated with the rs7903146 genotype, which are consistent with the previous findings [235,482]. In contrast, Pang et al. demonstrated that individuals homozygous for the rs7903146 and the rs12255372 T2D risk alleles (TT/TT) expressed 2.6-fold greater levels of *TCF7L2* mRNA compared to individuals homozygous for the nonrisk alleles (CC/GG, $P = 0.006$), although differentially spliced *TCF7L2* transcripts did not differ by T2D risk-associated genotype in PBMC (peripheral blood mononuclear cells), suggesting the tissue-specific differences in enhancer usage between adipose and blood tissues [251]. Thus, future experiments are necessary to elucidate if the transcription factor TF3, TF4 and TF5 contribute to the described phenotypes, and regulation of *TCF7L2* gene or gene isoforms. For example, TF3, TF4 and TF5 siRNA knockdown in pancreas β-cell, myoblasts, neuronal cells in a genotype-dependent manner would give some hints about the *TCF7L2* gene regulation mediated by allele-specific binding transcription factors.

The variant rs4684847 at the *PPARG* locus demonstrated two allele-specific bands in EMSA assays. Indeed, the following LC-MS/MS analysis and further functional assays identified two allele-specific binding TFs, PRRX1 and TF1. Also, the rs1421085 at the *FTO* locus and the rs7903146 at the *TCF7L2* locus indicated two or three allele-specific bindings of proteins in EMSA assays. Considering the TFBS modularity at genomic regions surrounding *cis*-regulatory variants predicted by PMCA [98], such TFs are unlikely to act alone, which could have a modest impact and become more pronounced together in network. Moreover, the protein components of signaling pathways regulating gene activity could be additional targets for the development of personalized therapeutics. Thus, the importance of

combinatorial effects needs to be further elucidated across cell types and tissues. The second predicted regulatory variant, rs7647481 at the *PPARG* locus was identified as *cis*-regulatory with its allele-specific binding transcription factor, YY1. These results support the recently proposed 'multiple enhancer variant' hypothesis [440], which supposes that multiple causal variants in LD at GWAS inferred loci. To assess if also for the *FTO* and *TCF7L2* loci, additional to the variant rs1421085 and rs7903146, further variants in high LD may affect regulatory activity, further experiments will be required.

8.3 Unraveling molecular mechanisms influenced by *cis*-regulatory genomic variants at the *PPARG* locus using unbiased allele-specific quantitative proteomics

In the past decade, numerous genetic variants have been associated with common traits through GWAS, and recent advances of the ENCODE project and novel bioinformatics approaches [98,100,414,415] have facilitated identification of *cis*-regulatory, potentially diseases-causing variants within complex loci. However, the identification of genes that serve as the molecular basis of risk etiology, and the identification of causal variants have not been nearly as successful. For future implementation of molecular targeted therapies, the precise delineation of molecular mechanisms affected by causal *cis*-regulatory variants would to be essential, i.e. as a first step the efficient and precise identification of allele-specific binding proteins. Only a few studies successfully provided such protein identifications [312,314,418]. Identification of allele-specific binding proteins by TFBS matrix overlap [266,327] or ChIP-seq [353] faces limitations such as availability of TFBS matrix annotation and the complexity of TFBS modularity [98,483]. Moreover, the spatial, temporal expression patterns of TFs and their coregulators emphasize the need to consider cell-type specific open chromatin data [98,100,259,355,484] to prioritize candidate *cis*-regulatory variants.

Previously was shown that PMCA, a computational analysis of phylogenetic conservation with a complexity assessment of co-occurring TFBS, can identify *cis*-regulatory variants [98]. Supporting the recently proposed 'multiple enhancer variant' hypothesis [440], which supposes that multiple causal variants in LD at GWAS inferred loci, PMCA inferred multiple variants in high LD which may affect *cis*-regulatory activity at loci associated with T2D, Asthma and Crohn's disease [98]. Thus, to distinguish *cis*-regulatory from non *cis*-regulatory variants at the T2D associated *PPARG* locus, an integrated framework was applied, combining computational TFBS modularity PMCA analysis [98], assessment of inferences from publically

available functional cell type- and differentiation-specific data from human adipocytes [353], and the label-free proteomics used in this study.

At the *PPARG* locus, PMCA predicted six candidate *cis*-regulatory SNPs out of 24 SNPs in high LD ($r^2 \geq 0.7$, 1000 Genomes [172], with the *PPARG* tagSNP rs1801282) with conserved binding site modularity [98]. Subsequent inference of a specific clustering of homeobox TFBS at predicted T2D *cis*-regulatory SNPs unveiled the essential role the rs4684847C risk allele [98]. Here, by further integration of epigenetic marks on the regulatory region data [353], the second *cis*-regulatory variant, rs7647481 was found, which is in perfect linkage to the rs4684847 and the well-established coding Pro12Ala variant which blunts transcriptional activity of the insulin-sensitizing PPARγ2 [163] despite being associated with enhanced insulin sensitivity. The both, rs4684847 and rs7647481 variants might contribute to adipocyte specific *PPARG2* or ubiquitous *PPARG1* isoform expression, respectively [98,134–136,161,417]), which show overlap with H3K27ac in late stages of adipocyte differentiation and rs7647481 with H3K4me1 and H3K4me2 in all stages of differentiation. The 'thrifty gene' locus *PPARG* [485] may be an example where multiple *cis*-regulatory variants and may provide some selective advantage during evolution and now contribute to the observed T2D disease phenotype.

Using unbiased quantitative proteomics, two-times more allele-specific binding proteins were found at *cis*-regulatory variants as compared to non *cis*-regulatory, further supporting the integrative framework predictions. Moreover, the binding of the TF PRRX1 at the rs4684847 C risk allele inhibits expression of the adipocyte-specific *PPARG2*, but not the ubiquitously expressed *PPARG1* isoform [98]. Insulin resistance was previously discovered to be associated with adipose tissue *PRRX1* expression by Claussnitzer and colleagues [98], identified as rs4684847 risk C allele binding transcriptional inhibitor in this study. Here, proteomic analysis at the rs7647481 A nonrisk allele inferred binding of two transcription factors, namely YY1, reported to regulate metabolic, diabetes-related phenotypes in skeletal muscle [486] and liver [487–489], and NFATC4, reported to promote mouse adipocyte differentiation by direct regulation of *PPARG* [490]. Notably, recent evidence suggests that YY1 shows the allele-specific transcriptional activity in the estrogen receptor beta gene promotor [491]. The high sensitivity of the applied proteomics approach further enabled identification of the YY1 interacting coregulator RYBP, which recently has been found to infer with skeletal myogenesis [492], additional to its function as transcription repressor in cancer [493,494], embryogenesis [420] and central nervous system development [495].

Importance of protein-complexes has been often recognized in many studies to perform a specific cellular function [496–499]. A discrete biological function can only seldom be attributed to an individual protein, and most biological functions come from interactions among many components, e.g. in the signal transduction system in yeast [500,501]. Recently, development of high-throughput in biotechnology has led to the rapid generation of numerous biological data such as protein interactions. Protein interaction networks for many species have been used in order to support the elucidation of protein function [502]. Thus, the protein-protein-Interaction (P-P-I) networks for YY1 were further analyzed in order to understand the connectivity of their signaling routes associated with the SNPs. The data were curated from the GePS (Genomatix, Munich, Germany) which provides both predicted and experimental interaction information with confidence score. Strikingly, the YY1 is well known to function as transcriptional activator/repressor in context with other regulators [503]. In SGBS preadipocytes, knockdown of YY1 or RYBP alone was not sufficient to reveal a significant effect on *PPARG1* and *PPARG2* expression. However, combined action of the transcription factor YY1 and its coregulator RYBP led to a significant two-fold reduction for endogenous mRNA expression levels of the insulin-sensitizing *PPARG2* isoform crucial for maintaining insulin-sensitivity [504], supporting the importance to P-P-I network in molecular process. Thus, different *cis*-regulatory variants at the *PPARG* locus may contribute to diseases pathophysiology, strengthening the 'multiple enhancer hypothesis' [440] and suggesting varying numbers of regulatory SNPs per LD block [98,356,505]. Besides YY1, proteomics at the rs7647481 variant inferred allele-specific binding of the transcription factor Nfatc4. In fact, Nfatc4 was reported to promote mouse adipocyte differentiation, and *PPARG* was supposed to be a target gene of Nfatc4 [506]. Both YY1 and NFATC4 TFBSs in close proximity have been implicated in the regulation of the human IFNG promoter in T cells [507]. The likewise co-occurrence of YY1 and NFATC4 at the rs7647481 variant, further corroborates the importance of assessing co-occurring TFBS, an essential feature of the PMCA methodology, and moreover the power of label-free proteomics to find relevant allele-specific binding transcription factors at *cis*-regulatory variants.

To the end, using proteomics directly on eluted fractions containing allele-specific binding-proteins allows identification of allele-specific binding transcription factors and moreover for the first time, identification of related transcriptional coregulators which are the lowest abundant proteins in cells [484,508]. It also can be assigned to disease pathophysiology as

exemplified at the *PPARG* T2D risk locus. Moreover, integrative approaches combining computational and cell type specific epigenetic marks, *cis*-regulatory prediction with highly sensitive proteomics data of allele-specific binding proteins can help to clarify the role of inherited and somatic variability. Moreover, it is supposed that an integrative analysis combining computational predictions, NGS based epigenetic marks data and improved classical molecular tools like label-free quantitative proteomics supports achieving the common goal of defining the pathogenic potential of variants and ultimately the search for pathways and molecular targeted therapies.

8.4. *Cis* and *trans* regulation of gene expression by genetic variants

Most studies have been so far limited to detect *cis*-regulatory variants affecting gene expression and only a few *trans*-regulatory variants, which typically have weaker effect sizes than *cis*-regulatory variants [509,510]. *Cis*-acting regulatory mechanisms by genetic variants may affect different aspects of gene expression including transcription, alternative mRNA processing or mRNA stability [434]. eQTL analysis has been successfully used for providing a better understanding the functional impacts of variants associated with complex traits and diseases via changes in whole-gene expression levels [260,262,509–512]. Other mechanisms such as alternative mRNA processing, were much less often studied despite their known importance in a target gene. Moreover, the alternative mRNA processing and mRNA stability exert their functional regulation upon local regions within mRNAs and thus only affect variants in nearby regions [434]. Recently, Smemo et al. revealed evidence of long-range target gene of *cis*-regulatory variants. There were the genomic interactions with the promoters of genes located within a 1-megabase (Mb) window around the obesity-associated variants, including *FTO* and *RPGRIP1L*, and *IRX3* genes, recapitulating aspects of *IRX3* expression. It is noteworthy that the 47-kb obesity-associated interval is full of *cis*-regulatory elements, indicating an abundance of enhancer-associated chromatin marks, DNase hypersensitive sites, and TF binding events [102]. However, the understanding about long-range gene regulatory control has been largely missed. The ENCODE project discovered more than 1,000 long-range interactions between promoters and distal sites that are located 120 kb upstream of the transcription start sites (TSSs) and include elements resembling enhancers, promoters and CTCF bound sites [513]. Similar result was also obtained by Vadnais et al. that CUX1 regulates genes at a distance and also regulate more than one gene on certain genomic loci at a long-

range [514]. Likewise, it would be very interesting to examine whether the putative allele-specific binding transcription factors TF2. TF3 and TF4 at the rs1421085 of the *FTO* locus could affect regulation of IRX3 gene or other genes at a long-distance in brain or other tissues such as liver and adipose tissue.

In contrast to *cis*, *trans*-variants have more difficulties to be mapped because most studies were conducted on relatively small sample sizes, limiting the power to detect variants affecting gene expression in *cis* and to a greater extent in *trans*, as *trans*-variants typically have weaker effect sizes than *cis* [509,510,515]. Moreover, while local variants are likely to be *cis*-acting, variants at a distance are likely to be *trans*-acting. For these reasons, identification of the *trans*-variants is a challenge in human genetics, requiring large sample sizes owing to the number of comparisons because all genotyped variants in the genome should be considered for each association [516]. A few studies reported that the findings of disease-related *trans*-variants implicated their roles in regulating the expression of multiple genes [517,518]. Interestingly, a large sample size allows detecting variants acting both in *cis* and *trans*, suggesting that there might be a regulatory relationship between *cis* and *trans* regulated genes [509]. Bryois et al. found that variants associated with complex traits and common diseases are more likely to be *cis*- and *trans*-eQTLs than matched variants, further confirming that a significant fraction of trait associated variants are acting at the gene expression level. Moreover, a large portion of *trans*-effects of *cis*-eQTLs is concordant with the fact that about 65% of the heritability of gene expression is *trans* to the gene in lymphoblastoid cell line [509]. This observation is consistent with the previous study demonstrating that many *trans* variants are associated with multiple transcripts. It suggests that they are multigene regulators, predominantly in a tissue-dependent manner [510]. An advantage of allele-specific approach in this study is, that it could more effectively determine how genes are regulated by allele-specific binding transcription factors using siRNA knockdown [98,519] and CRISPR-Cas9 gene-editing system [98] which allow to find both, *cis*- and *trans*-regulated genes. Taken together, such studies emphasize the importance of studying long-range *cis*-regulatory variants as well as *trans*-variants in complex traits to extend understanding the architecture and regulation of gene expression in multiple ways in human diseases. As mentioned above, this study successfully identified the *cis*-regulatory variants, altering the *PPARG* gene expression at a given locus and making it optimal for detecting *cis*-acting difference. As a next challenge, this study also will allow to detect such variants through important relationships between

gene-protein, protein-protein, contributing to the new challenge in human genetics.

8.5. Conclusions

In the last decade, the number of risk alleles for complex diseases such as T2D has been identified by GWAS, however most of those studies have been limited to detect rare variants with stronger effects [520]. Recently, there has been increasing attention in the biological mechanisms underlying genotype-phenotype associations [426], however most of them remain unknown. One possible mechanism is that genetic variants may influence gene transcription via transcription factor binding. Notably, recently published bioinformatics PMCA analysis [98] and the functional cell type- and differentiation-specific epigenetic marks data [353] allowed the prioritization of non-coding variants to *cis*-regulatory potentially contributing to the T2D susceptibility. The integrative computational analysis of PMCA assessing co-occurring transcription factor binding sites (TFBS) predicted several variants at the T2D risk loci with potential disruption or enhancement of TFs binding [98].

This study started with the selection of several predicted *cis*-regulatory variants at the *PPARG*, *FTO* and *TCF7L2* loci, inferred from PMCA analysis [98]. EMSA results confirmed that the selected variants rs4684847 C/T (*PPARG*), rs7647481 A/G (*PPARG*), rs1421085 T/C (*FTO*) and rs7903146 T/C (*TCF7L2*) changed the differential protein complex binding affinity between the risk and nonrisk alleles. In order to further elucidate their mechanistic role in T2D development, this study aimed to develop a highly-sensitive proteomics approach, enabling identification of allele-specific protein complexes contributing to disease-pathophysiology. Unlike previously published approaches that used radioactive or stable isotope labeling, this approach is adaptable to the use of human materials and provides advantageous such as a simple handling and time-saving procedure.

Using the PMCA prediction [98] and the label-free proteomics approach established in the here presented study, the novel *cis*-regulatory variant rs4684847 at the *PPARG* locus and the PRRX1 transcription factor regulating PPARγ2 expression in adipocytes were found. Moreover, the prediction from PCMA and epigenetic marks of regulatory regions suggested that multiple variants at one locus may contribute to disease risk [98,440]. Indeed, the results from *in depth-analysis* at the *PPARG* locus revealed that the variant rs7647481 at the *PPARG* locus was identified as a *cis*-regulatory, and a YY1 (Ying Yang 1) transcription factor and its

coregulator RYBP (RING1 and YY1 binding protein) were identified as nonrisk allele-specific binding proteins. Pathophysiological relevance of these findings was supported by the fact that adipose mRNA levels of *RYBP* correlated with improved insulin sensitivity in subjects carrying the rs7647481 nonrisk allele. Moreover, the allele-specific proteins identified (fold change > 2, $P < 0.05$) were highly enriched at the predicted *cis*-regulatory variants compared to predicted non *cis* regulatory variants, validating the high sensitivity of the approach presented in this study.

For the first time, this study presented an approach to infer allele-specific protein-DNA interaction networks. The here presented approach constitutes a frame workflow combining an integrative analysis for prioritization of non-coding variants (here PMCA [98] and epigenetic marks data [353]) with highly efficient label-free proteomics methodology to identify TFs and their cofactors and provide the possibility of applying to any kind of variability, including somatic mutations in cancer, without loss of generality. This approach further supports the way towards the identification of both, *cis*-regulatory variants and affected disease mechanisms, and thereby towards personalized therapy. Moreover, for the first time this study provided proteome-wide experimental evidence for a significantly increased binding of transcription factors and related proteins to predict *cis*-regulatory versus non *cis*-regulatory variants. Of note, this finding indicated that this unbiased proteomics approach is so sensitive enough that supported the inferred prioritization independent of prior predictions. Furthermore, it also supported the power of an integrative framework analysis for *cis*-regulatory prioritization of non-coding variants, PMCA [98].

In summary, the label-free proteomics approach here presented was successfully applied to the identification of allele-specific binding proteins. And this study supported the recently supposed ‘multiple enhancer hypothesis’ and mighty help further uncover the contribution of diverse *cis*-regulatory variants to disease pathophysiology. Furthermore, the efficient identification of TFs and their coregulators will serve as a base for a prediction of how genetic variants in regulatory mechanisms change gene expression profiles and human phenotype, which represents the current quests in the human genetics.

9. Appendix (Supplementary tables)

9.1 Supplementary table S1: Overrepresentation of Molecular Function GO-terms related to DNA-binding activity in the set of significant allele-specific binding proteins at the predicted *cis*-regulatory variant and non *cis*-regulatory variants

Label-free quantitative proteomic analysis identified in total 824 -952 proteins binding at the predicted *cis*-regulatory variant and non *cis*-regulatory variants. 25-165 proteins with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments, see Supplementary table S2) were assessed for the GO-terms Molecular Function, “Structure-specific DNA binding” and “DNA-binding” using the GePS tool (Genomatix, Munich, Germany). Results for the overrepresentation analysis of GO-terms “Structure-specific DNA binding” and “DNA-binding” is given with the "GO-term ID", the respective *P*-value (Fisher´s Exact test), the number of "Input gene lists", the number of "Genes observed", the number of "Genes expected", the number of "Genes total" and the "List of observed genes".

	SNP	Elution	GO-Term	GO-Term ID	*P-value* [a]	No. of input genes for GO term Molecular Function	Genes (observed) [b]	Genes (expected) [c]	Genes (total) [d]	List of observed genes
predicted *cis*-regulatory	rs4684847	200	Structure-specific DNA binding	GO:0043566	9.43×10^{-3}	31	3	0.4395	217	TOP2A, SSBP1, YY1
		300	DNA binding	GO:0003677	1.36×10^{-6}	107	36	16.1835	2315	TRPS1, HNRNPA2B1, PA2G4, PRRX1, CHD4, TARDBP, HNRNPL, DDX1, ZNF800, TMPO, SFPQ, RYBP, NONO, PARP1, POLG2, TP53BP1, SAFB, UHRF2, FUS, NOLC1, HDAC2, KHDRBS1, POLG, PSIP1, TOP2A, MSH2, HSPD1, ILF3, BCLAF1, IMPDH2, KIF4A, MYBBP1A, ALB, ATM, HNRNPK, DDB1
			Structure-specific DNA binding	GO:0043566	2.11×10^{-5}	107	9	1.5170	217	HNRNPA2B1, TARDBP, POLG2, SAFB, PSIP1, TOP2A, MSH2, HSPD1, HNRNPK
	rs7647481	200	DNA binding	GO:0003677	1.58×10^{-7}	83	32	12.5536	2315	UBP1, HNRNPAB, TAF6L, HNRNPA2B1, HNRNPL, DNMT1, TMPO, SFPQ, SP3, TDP1, NONO, RFC3, RFC2, UBTF, HELLS, SAFB, FUS, DNAJC2, KHDRBS1, HMGN3, PNKP, MCM7, HNRNPD, SP1, IMPDH2, XRCC1, TAF15, KIF4A, NFATC4, RUVBL2, YY1, HNRPDL
			Structure-specific DNA binding	GO:0043566	2.60×10^{-6}	83	9	1.1767	217	HNRNPA2B1, SP3, TDP1, SAFB, PNKP, MCM7, SP1, YY1, HNRPDL
		300	Structure-specific DNA binding	GO:0043566	3.69×10^{-8}	107	12	1.5170	217	HNRNPA2B1, FEN1, TDP1, RBMS1, MCM6, PNKP, MCM7, PSIP1, TOP2A, MSH2, MSH6, HNRPDL
			DNA binding	GO:0003677	1.44×10^{-7}	107	38	16.1835	2315	POLB, HNRNPAB, UBP1, SMARCAL1, SMARCA4, KDM1A, HNRNPA2B1, HNRNPL, HNRNPU, DDX1, FEN1, TMPO, PRPF19, TDP1, RFC3, RFC2, UBTF, TP53BP1, RBMS1, RFC1, MCM6, KHDRBS1, PHB, KLF13, PNKP, TOP2B, MCM7, PSIP1, TOP2A, MSH2, HNRNPD, MAZ, MSH6, XRCC1, MYBBP1A, RUVBL2, HNRPDL, DDB1
predicted non-*cis*-regulatory	rs17036342	200	Structure-specific DNA binding	GO:0043566	4.04×10^{-3}	23	3	0.3261	217	MCM6, MCM7, AKAP8
			DNA binding	GO:0003677	4.44×10^{-3}	23	9	3.4787	2315	MCM6, UBP1, TEAD1, TAF6L, MCM7, KDM1A, EBF2, EBF1, AKAP8
		300	Structure-specific DNA binding	GO:0043566	9.43×10^{-3}	31	3	0.4395	217	MCM6, MCM7, SSBP1
			DNA binding	GO:0003677	0.04	31	9	4.6887	2315	MCM6, UBP1, TEAD1, SMARCA4, MCM2, MCM7, KDM1A, EBF1, SSBP1
	rs2881479	200	n.d.							
		300	Structure-specific DNA binding	GO:0043566	1.47×10^{-3}	59	5	0.8365	217	MCM6, MCM7, TOP2A, PCNA, MSH6

[a] *P*-values from Fisher´s Exact Test, [b] Genes observed refers to the number of genes within the input list associated with GO category, [c] Genes expected refers to the number of genes which are expected to be observed randomly, [d] Genes total refers to the total number of genes in the GO category, n.d.= not detected.

9.2 Supplementary table S2: Allele-specific binding proteins and GO-term analysis / transcription factor annotation at the predicted *cis*-regulatory rs4684847 (A, B), rs7647481 (C, D) and non *cis*-regulatory rs17036342 (E, F) and rs2881479 (G, H)

For all supplementary tables S2: [a]fold change was calculated as the mean ratio of normalized proteins abundance over the three experiments, [b]*P*-values were derived from unpaired *t*-tests, [c]selection criteria for candidate allele-specific binding proteins mediating *cis*-regulatory activity. [d]Peptide count refers to the total number of identified peptides per protein, [e]Peptide count for quantitation refers to the number of peptides uniquely assigned to one protein and therefore used for quantitation, [f]Mascot Percolator score is built as summed up single probability of identified peptides per protein and serves as indicator for the reliability of protein identification.

9.2.1 Supplementary table S2A: Classification of allele-specific binding proteins at the predicted *cis*-regulatory variant rs4684847 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 828 proteins binding at the rs4684847 surrounding genomic region (200 mM NaCl eluate of affinity chromatography). 41 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the C-allele / T-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms “DNA binding” and “transcription factor activity” were assessed for the total set of 828 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 41 proteins (fold-change > 2 or < 0.5, $P < 0.05$) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change C/T[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Increased binding at the rs4684847 C allele (risk)	Ubp1	18.93	0.0429	3 out of 3 criteria	X	X	transcription factor	1	1	66	ENSMUSP00000009885
		Yy1	2.09	0.0440		X	X	transcription factor	2	2	125	ENSMUSP00000021692
		Rbpj	8.03	0.0014	1 out of 3 criteria			transcription factor	13	13	1316	ENSMUSP00000040694
		Ddx21	44.82	0.0116				transcription cofactor	1	1	153	ENSMUSP00000042691
		Mcm3	31.86	0.0389		X			1	1	173	ENSMUSP00000059192
		Top2a	2.30	0.0430		X			13	7	1024	ENSMUSP00000068896
		Atm	9.78	0.0004	none				2	1	34	ENSMUSP00000113388
		Snrpf	8.28	0.0012					3	3	245	ENSMUSP00000020203
		Zc3h18	15.52	0.0014					1	1	19	ENSMUSP00000017622
		Dbn1	49.83	0.0035					1	1	136	ENSMUSP00000021950
		Snrpg	24.69	0.0064					1	1	64	ENSMUSP00000086987
		Tomm70a	64.75	0.0098					1	1	98	ENSMUSP00000129186
		Sf3a3	14.38	0.0098					1	1	238	ENSMUSP00000030734
		Eif4a1	6.16	0.0158					4	4	171	ENSMUSP00000127034
		Zfp646	43.79	0.0207					1	1	26	ENSMUSP00000052641
		Dnajc9	19.18	0.0253					2	2	76	ENSMUSP00000022345
		Mtap1b	366.72	0.0257					3	3	316	ENSMUSP00000068374
		Eftud2	39.49	0.0313					2	1	180	ENSMUSP00000021306
		Thoc2	45.60	0.0348					2	2	121	ENSMUSP00000044677
		Hsp90aa1	5.23	0.0390					9	4	528	ENSMUSP00000021698
		Olfr1395	2.35	0.0416					1	1	15	ENSMUSP00000050142
		Elavl1	43.80	0.0490					2	1	64	ENSMUSP00000096549
(B)	Decreased binding at the rs4684847 T allele (non-risk)	Ssbp1	0.39	0.0274	2 out of 3 criteria	X		transcription cofactor	5	5	426	ENSMUSP00000110427
		Gaa	0.43	0.0003	none				2	2	192	ENSMUSP00000026666
		Anp32b	0.47	0.0005					3	3	92	ENSMUSP00000099990
		Ero1l	0.49	0.0011					2	2	171	ENSMUSP00000022378
		Flnb	0.45	0.0018					10	8	407	ENSMUSP00000052020
		Acaa1b	0.24	0.0036					1	1	29	ENSMUSP00000010795
		Pck2	0.47	0.0040					4	4	226	ENSMUSP00000038555
		Acadl	0.47	0.0072					2	2	107	ENSMUSP00000027153
		Scpep1	0.42	0.0117					1	1	55	ENSMUSP00000000287
		Gm9755	0.45	0.0158					1	1	29	ENSMUSP00000091180
		Eci1	0.48	0.0160					1	1	60	ENSMUSP00000024946
		Aldh6a1	0.42	0.0177					1	1	76	ENSMUSP00000082288
		Hadh	0.47	0.0206					4	4	264	ENSMUSP00000029610
		Suclg2	0.46	0.0215					2	2	41	ENSMUSP00000078774
		Pitpnm2	0.39	0.0215					1	1	20	ENSMUSP00000083292
		Gstm7	0.22	0.0217					1	1	23	ENSMUSP00000004137
		BC046331	0.47	0.0263					1	1	19	ENSMUSP00000040762
		Oat	0.45	0.0303					3	2	139	ENSMUSP00000081544
		Gstp1	0.47	0.0315					2	2	245	ENSMUSP00000129565

9.2.2 Supplementary table S2B: Classification of allele-specific binding proteins at the predicted *cis*-regulatory variant rs4684847 using GO-term analysis and transcription factors annotation

Label-free quantitative proteomic analysis identified in total 824 proteins binding at the rs4684847 surrounding genomic region (300 mM NaCl eluate of affinity chromatography). 165 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the C-allele / T-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms “DNA binding” and “transcription factor activity” were assessed for the total set of 824 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 165 proteins (fold-change > 2 or < 0.5, $P < 0.05$) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change C/T[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Increased binding at the rs4684847 C allele (risk)	Rybp	2.28	0.0035	3 out of 3 criteria	X	X	transcription cofactor	4	4	235	ENSMUSP00000098677
		Prrxl	2.59	0.0122		X	X	transcription factor	6	5	752	ENSMUSP00000027878
		Ddxl	3.00	0.0142		X	X	transcription cofactor	5	5	189	ENSMUSP00000065987
		Trp53bp1	2.89	0.0244		X	X	transcription cofactor	4	4	219	ENSMUSP00000106277
		Psip1	2.36	0.0360		X	X	transcription cofactor	18	17	1875	ENSMUSP00000030207
		Hnrnpa2b1	2.94	0.0001	2 out of 3 criteria	X		transcription cofactor	14	12	1421	ENSMUSP00000087453
		Fus	2.99	0.0016		X		transcription cofactor	8	7	1009	ENSMUSP00000076801
		Sfpq	2.00	0.0062		X		transcription cofactor	21	20	1576	ENSMUSP00000030623
		Khdrbs1	3.54	0.0077		X		transcription cofactor	5	5	196	ENSMUSP00000066516
		Ddx17	2.16	0.0137			X	transcription cofactor	22	15	1633	ENSMUSP00000055535
		Pa2g4	2.27	0.0195		X		transcription cofactor	7	7	400	ENSMUSP00000026425
		Parp1	2.11	0.0241		X		transcription cofactor	27	26	2038	ENSMUSP00000027777
		Chd4	3.78	0.0281		X		transcription cofactor	8	8	330	ENSMUSP00000060054
		Trps1	3.09	0.0284		X		transcription factor	29	26	2251	ENSMUSP00000077089
		Ddx5	2.51	0.0375			X	transcription cofactor	18	11	842	ENSMUSP00000021062
		Mybbp1a	2.76	0.0405		X		transcription cofactor	3	3	97	ENSMUSP00000044827
		Tardbp	3.54	0.0410		X		transcription cofactor	7	7	384	ENSMUSP00000081142
		Hdac2	12.39	0.0433		X		transcription cofactor	4	2	285	ENSMUSP00000019911
		Rbpj	13.76	0.0000	1 out of 3 criteria			transcription factor	13	13	1316	ENSMUSP00000040694
		Hmgn1	3.29	0.0013				transcription cofactor	2	2	171	ENSMUSP00000061012
		Nolc1	2.49	0.0023		X			1	1	21	ENSMUSP00000128331
		Ddb1	2.17	0.0036		X			7	7	227	ENSMUSP00000025649
		Impdh2	2.13	0.0054		X			4	2	180	ENSMUSP00000079888
		Hnrnpl	14.76	0.0054		X			12	12	1440	ENSMUSP00000049407
		Rbbp4	4.43	0.0055				transcription cofactor	10	6	576	ENSMUSP00000099658
		Hnrnpul1	2.29	0.0057		X			8	6	338	ENSMUSP00000037268
		Sarnp	5.14	0.0063				involved in transcription regula	3	3	210	ENSMUSP00000100863
		Ilf3	4.20	0.0102				transcription factor	4	4	92	ENSMUSP00000065770
		Top2a	2.91	0.0122		X			13	7	1024	ENSMUSP00000068896
		Hnrnpk	3.17	0.0167		X			14	3	1525	ENSMUSP00000039269
		Tmpo	2.59	0.0169		X			8	3	675	ENSMUSP00000020123
		Zfp800	2.70	0.0173		X			2	1	37	ENSMUSP00000039222
		Polg	2.48	0.0198		X			22	20	1626	ENSMUSP00000073551
		Polg2	2.40	0.0241		X			10	8	848	ENSMUSP00000021060
		Uhrf2	3.08	0.0248		X			7	6	546	ENSMUSP00000025739
		Ebf3	4.70	0.0290				transcription factor	1	1	14	ENSMUSP00000033378
		Ybxl	4.10	0.0294				transcription factor	5	3	559	ENSMUSP00000078589
		Nono	4.46	0.0309		X			18	15	1662	ENSMUSP00000033673
		Rbbp7	3.11	0.0325				transcription cofactor	8	5	285	ENSMUSP00000033720
		Hoxb4	3.66	0.0329				transcription factor	2	1	91	ENSMUSP00000048002
		Safb	3.28	0.0337		X			3	3	302	ENSMUSP00000092849
		Bclaf1	2.27	0.0363		X			3	2	109	ENSMUSP00000043583
		Hspd1	4.10	0.0395		X			14	14	1866	ENSMUSP00000027123
		Ewsr1	3.00	0.0397				transcription cofactor	3	2	158	ENSMUSP00000073034
		Tcp1	23.33	0.0421				transcription factor	6	6	956	ENSMUSP00000086418
		Msh2	2.05	0.0425		X			10	9	577	ENSMUSP00000024967
		Ruvbl1	161.16	0.0493				transcription cofactor	1	1	14	ENSMUSP00000032165
		Rps8	2.04	0.0000	none				6	6	598	ENSMUSP00000099757
		Zc3h18	14.94	0.0003					1	1	19	ENSMUSP00000017622
		Caprin1	3.31	0.0004					6	6	401	ENSMUSP00000028607
		Hspa8	2.17	0.0005					21	17	1828	ENSMUSP00000015800
		Mrrf	2.39	0.0005					5	5	415	ENSMUSP00000028250
		Elavl1	22.09	0.0006					2	1	64	ENSMUSP00000096549
		Eif4b	2.86	0.0006					3	3	267	ENSMUSP00000127774
		Gm8991	2.55	0.0011					9	9	991	ENSMUSP00000072775
		Myadm	9.03	0.0015					1	1	78	ENSMUSP00000094505
		Snrpg	16.70	0.0017					1	1	64	ENSMUSP00000086987
		Hist1h3f	2.39	0.0020					4	4	234	ENSMUSP00000074994
		Hnrnpa1	2.68	0.0024					12	8	1186	ENSMUSP00000042658
		Rpl23	2.30	0.0025					4	4	348	ENSMUSP00000099435
		Crat	44.87	0.0026					2	2	51	ENSMUSP00000028207
		Rps26	2.10	0.0032					2	2	46	ENSMUSP00000026420
		Rplp0	2.45	0.0034					5	5	175	ENSMUSP00000083705
		Ppp1ca	2.28	0.0034					9	1	642	ENSMUSP00000039109
		Eif4a1	14.85	0.0037					4	4	171	ENSMUSP00000127034
		Timm13	4.41	0.0041					1	1	153	ENSMUSP00000020440
		Srrm2	3.53	0.0041					1	1	72	ENSMUSP00000085993
		Atm	3.76	0.0043					2	1	34	ENSMUSP00000113388
		Pcbp2	3.06	0.0047					4	2	178	ENSMUSP00000076294
		Tpx2	2.66	0.0052					9	9	227	ENSMUSP00000028969
		Actg1	2.09	0.0055					17	7	1693	ENSMUSP00000071486
		Eif5a	3.22	0.0055					6	6	667	ENSMUSP00000047008
		Pcolce	4.84	0.0055					8	8	328	ENSMUSP00000031731
		2310036O22Rik	3.21	0.0058					1	1	29	ENSMUSP00000044129
		Serbp1	2.13	0.0062					13	13	986	ENSMUSP00000039110
		Snrpa1	32.28	0.0064					2	1	46	ENSMUSP00000117947
		Srsf10	2.28	0.0073					1	1	24	ENSMUSP00000095455
		Nhp2	2.34	0.0076					2	2	306	ENSMUSP00000120014
		Rps23	3.93	0.0076					3	3	67	ENSMUSP00000054490
		Rpl18a	2.23	0.0077					3	3	76	ENSMUSP00000058368
		Mki67	3.80	0.0078					5	5	349	ENSMUSP00000033310
		Hist1h2bc	2.49	0.0084					5	5	355	ENSMUSP00000018246
		Cct7	56.85	0.0087					6	6	626	ENSMUSP00000032078
		Lrch3	4.85	0.0097					1	1	19	ENSMUSP00000023491
		Gml0036	2.26	0.0097					2	2	232	ENSMUSP00000078670
		Erh	2.21	0.0099					1	1	153	ENSMUSP00000021559
		Rpl13	2.05	0.0101					4	3	228	ENSMUSP00000000756

Continue

		Gene symbol	Fold change C/T[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Increased binding at the rs4684847 C allele (risk)	Nap1l1	7.56	0.0111	none				1	1	19	ENSMUSP00000070068
		Srp14	2.02	0.0115					2	2	63	ENSMUSP00000009693
		Hnrnpr	2.03	0.0121					4	1	259	ENSMUSP00000081239
		Efcab8	3.23	0.0124					1	1	14	ENSMUSP00000135811
		Pspc1	2.13	0.0127					9	9	646	ENSMUSP00000022507
		Eef1a1	19.43	0.0139					15	15	1401	ENSMUSP00000042457
		Myo15b	5.20	0.0141					1	1	14	ENSMUSP00000117804
		Znf512b	2.69	0.0143					13	13	546	ENSMUSP00000115601
		Plxnb1	2.12	0.0148					1	1	23	ENSMUSP00000071966
		Rpl36	2.34	0.0150					2	2	156	ENSMUSP00000079340
		Rpl14	2.10	0.0152					3	3	207	ENSMUSP00000131489
		Hbb-y	12.35	0.0157					1	1	153	ENSMUSP00000033229
		Plp2	2.36	0.0158					1	1	29	ENSMUSP00000033486
		Eef1g	11.63	0.0160					5	5	339	ENSMUSP00000093955
		Gm4987	2.99	0.0161					1	1	21	ENSMUSP00000077440
		Rps17	2.49	0.0180					3	3	252	ENSMUSP00000079628
		Snrpd3	3.12	0.0182					4	4	261	ENSMUSP00000020397
		Rpl21	3.66	0.0184					2	2	52	ENSMUSP00000041652
		Ppib	2.12	0.0188					14	13	1152	ENSMUSP00000034947
		Rbmxl1	4.77	0.0199					2	2	139	ENSMUSP00000048153
		AC239834.2	3.03	0.0216					11	9	951	ENSMUSP00000100550
		Snrpe	5.39	0.0223					2	2	153	ENSMUSP00000127164
		Sf1	2.65	0.0229					3	3	144	ENSMUSP00000109115
		Dbn1	14.51	0.0230					1	1	136	ENSMUSP00000021950
		Hsp90b1	2.41	0.0235					21	20	1804	ENSMUSP00000020238
		Nudt21	2.04	0.0246					2	2	39	ENSMUSP00000034204
		Cald1	2.37	0.0248					11	11	676	ENSMUSP00000110673
		Snrpn	2.95	0.0251					1	1	103	ENSMUSP00000055941
		Ddx39b	3.49	0.0253					5	4	355	ENSMUSP00000070682
		Fn1	2.90	0.0260					51	48	4778	ENSMUSP00000054499
		Mndal	2.44	0.0264					5	2	280	ENSMUSP00000106841
		Fam76b	3.81	0.0269					3	3	96	ENSMUSP00000062642
		BC046331	2.90	0.0271					1	1	19	ENSMUSP00000040762
		Srsf1	2.00	0.0282					8	7	391	ENSMUSP00000103553
		Tubb5	2.59	0.0284					7	3	517	ENSMUSP00000001566
		Marcks	2.56	0.0286					2	2	54	ENSMUSP00000090245
		Gm9396	2.03	0.0293					6	6	555	ENSMUSP00000084807
		Cct8	8.41	0.0294					5	5	224	ENSMUSP00000026704
		Serf2	2.18	0.0300					1	1	29	ENSMUSP00000097074
		Vim	6.03	0.0309					14	12	1247	ENSMUSP00000028062
		Hsp90aa1	12.57	0.0312					9	4	528	ENSMUSP00000021698
		Hist1h4j	2.16	0.0320					5	5	369	ENSMUSP00000085006
		Cct2	13.38	0.0322					10	10	747	ENSMUSP00000036288
		Hsp90ab1	5.22	0.0334					11	5	863	ENSMUSP00000024739
		Rpl3	6.01	0.0335					7	7	563	ENSMUSP00000080354
		Lrba	2.68	0.0337					1	1	14	ENSMUSP00000103261
		Eif2s3x	2.76	0.0338					8	3	489	ENSMUSP00000059395
		Prpf40a	5.96	0.0341					2	2	40	ENSMUSP00000075655
		Eif3c	4.87	0.0354					1	1	14	ENSMUSP00000032992
		Sf3a3	32.08	0.0364					1	1	238	ENSMUSP00000030734
		Acin1	2.45	0.0366					6	5	438	ENSMUSP00000022793
		Eef2	2.36	0.0388					13	12	715	ENSMUSP00000046101
		Tgfb1	2.89	0.0388					1	1	15	ENSMUSP00000002678
		Srsf9	2.06	0.0399					1	1	27	ENSMUSP00000031513
		Rpl24	2.92	0.0399					1	1	21	ENSMUSP00000023269
		Ubap2l	2.49	0.0416					4	4	125	ENSMUSP00000029553
		Lyar	2.48	0.0428					2	2	56	ENSMUSP00000084791
		Mmp12	3.01	0.0441					1	1	19	ENSMUSP00000005950
		Sart1	3.72	0.0442					2	2	182	ENSMUSP00000047397
		Fxr1	2.83	0.0451					3	2	238	ENSMUSP00000001620
		Hnrnpm	5.55	0.0462					6	5	261	ENSMUSP00000084864
		Hmmr	2.40	0.0476					1	1	19	ENSMUSP00000020579
		Canx	2.21	0.0477					3	3	198	ENSMUSP00000020637
		Dazap1	2.11	0.0481					5	5	420	ENSMUSP00000089958
		Fmr1	2.90	0.0489					2	1	156	ENSMUSP00000085906
(B)	Decreased binding at the rs4684847 T allele (non-risk)	Kif4	0.22	0.0018	out	X			5	5	88	ENSMUSP00000048383
		Alb	0.05	0.0179		X			3	1	256	ENSMUSP00000031314
		Csnk1a1	0.28	0.0496				transcription cofactor	1	1	19	ENSMUSP00000110901
		Exoc3l	0.03	0.0056	non				1	1	15	ENSMUSP00000053766
		Dsp	0.19	0.0070					6	5	319	ENSMUSP00000115062
		Pard3b	0.02	0.0077					1	1	14	ENSMUSP00000116912
		Idh3a	0.30	0.0102					1	1	14	ENSMUSP00000127526
		Sgce	0.11	0.0152					1	1	21	ENSMUSP00000004750
		Plec	0.46	0.0272					1	1	23	ENSMUSP00000023226
		1810035L17Rik	0.35	0.0289					1	1	29	ENSMUSP00000076673
		Olfr796	0.01	0.0328					1	1	14	ENSMUSP00000040207
		Try10	0.22	0.0396					3	1	88	ENSMUSP00000071976
		Ppig	0.33	0.0491					1	1	29	ENSMUSP00000114570

9.2.3 Supplementary table S2C: Classification of allele-specific binding proteins at the predicted *cis*-regulatory variant rs7647481 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 869 proteins binding at the rs7647481 surrounding genomic region (200 mM NaCl eluate of affinity chromatography). 108 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the A-allele / G-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms “DNA binding” and “transcription factor activity” were assessed for the total set of 869 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 108 proteins (fold-change > 2 or < 0.5, $P < 0.05$) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change A/G[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Increased binding at the rs7647481 A allele (non-risk)	Yy1	6.65	0.0029	3 out of 3 criteria	X	X	transcription factor	9	9	374	ENSMUSP00000021692
		Nfatc4	2.59	0.0114		X	X	transcription factor	2	2	58	ENSMUSP00000024179
		Ubp1	47.47	0.0350		X	X	transcription factor	1	1	73	ENSMUSP00000009885
		Taf6l	43.27	0.0481		X	X	transcription cofactor	1	1	28	ENSMUSP00000003777
		Ubtf	3.24	0.0041	2 out of 3 criteria	X		general transcription factor (P	13	13	637	ENSMUSP00000006754
		Ruvbl2	2.12	0.0011		X		transcription cofactor	3	3	88	ENSMUSP00000033087
		Sp3	2.99	0.0089		X		transcription factor	1	1	48	ENSMUSP00000065807
		Sp1	2.85	0.0394		X		transcription factor	1	1	46	ENSMUSP00000001326
		Hells	3.17	0.0471		X		transcription cofactor	1	1	25	ENSMUSP00000025965
		Mcm7	13.44	0.0497		X		transcription cofactor	1	1	24	ENSMUSP00000000505
		Tdp1	19.49	0.0004	1 out of 3 criteria	X			2	2	39	ENSMUSP00000021594
		Nfatc3	11.88	0.0009				transcription factor	1	1	19	ENSMUSP00000104931
		Ehmt1	31.01	0.0011				transcription cofactor	5	5	117	ENSMUSP00000088906
		Cbx3	3.85	0.0045				transcription cofactor	3	3	97	ENSMUSP00000031862
		Pnkp	2.95	0.0048		X			9	9	294	ENSMUSP00000003044
		Xrcc1	2.22	0.0049		X			8	7	228	ENSMUSP00000070995
		Rfc3	2.49	0.0051		X			5	4	217	ENSMUSP00000039621
		Tmpo	2.34	0.0068		X			14	14	632	ENSMUSP00000020123
		Dnajc2	2.15	0.0072		X			1	1	24	ENSMUSP00000030771
		Ehmt2	13.97	0.0076				transcription cofactor	4	4	162	ENSMUSP00000013931
		Kif4	3.45	0.0122		X			4	4	84	ENSMUSP00000048383
		Dnmt1	2.93	0.0196		X			29	29	1227	ENSMUSP00000004202
		Impdh2	2.18	0.0203		X			4	2	119	ENSMUSP00000079888
		Hmgn3	2.10	0.0230		X			2	2	61	ENSMUSP00000123932
		Rbbp4	3.02	0.0264				transcription cofactor	5	5	163	ENSMUSP00000099658
		Rfc2	3.43	0.0264		X			3	3	79	ENSMUSP00000023867
		Ddx21	19.29	0.0347				transcription cofactor	1	1	91	ENSMUSP00000042691
		Dhx9	15.33	0.0478				transcription cofactor	3	3	228	ENSMUSP00000038135
		Tmed10	2.66	0.0003	none				1	1	37	ENSMUSP00000037583
		Kif23	2.12	0.0009					1	1	23	ENSMUSP00000034815
		Eef1g	2.85	0.0016					5	5	177	ENSMUSP00000093955
		Abcf1	3.29	0.0020					9	8	349	ENSMUSP00000036881
		Dpy30	2.86	0.0020					1	1	24	ENSMUSP00000108190
		Nhp2	2.30	0.0023					2	2	184	ENSMUSP00000120014
		Mki67	2.00	0.0033					4	4	159	ENSMUSP00000033310
		Rrbp1	3.27	0.0049					43	43	2380	ENSMUSP00000016072
		Brd3	3.07	0.0054					8	7	230	ENSMUSP00000028282
		Lyar	2.00	0.0066					2	2	55	ENSMUSP00000084791
		Snrpg	30.58	0.0069					2	1	82	ENSMUSP00000086987
		Nop2	2.03	0.0078					14	14	572	ENSMUSP00000047123
		Dnajc9	44.63	0.0107					1	1	101	ENSMUSP00000022345
		Nup210	520.62	0.0113					1	1	23	ENSMUSP00000032179
		Prpf40a	15.25	0.0115					2	2	92	ENSMUSP00000075655
		Wiz	6.31	0.0139					12	11	608	ENSMUSP00000126253
		Nop58	2.03	0.0143					9	9	405	ENSMUSP00000027174
		Ssr1	2.36	0.0149					1	1	58	ENSMUSP00000021864
		Snrpf	7.23	0.0158					4	4	286	ENSMUSP00000020203
		Skiv2l2	4.98	0.0159					4	4	127	ENSMUSP00000022281
		Dbn1	17.42	0.0163					1	1	52	ENSMUSP00000021950
		Tomm70a	46.09	0.0173					1	1	38	ENSMUSP00000129186
		Ssb	2.38	0.0186					1	1	44	ENSMUSP00000088365
		Hspa9	2.00	0.0193					6	6	274	ENSMUSP00000025217
		Eif4a1	9.37	0.0232					4	4	201	ENSMUSP00000127034
		Slain2	3.36	0.0292					1	1	17	ENSMUSP00000115871
		Olfr1221	3.11	0.0311					1	1	18	ENSMUSP00000097383
		Hspa14	2.42	0.0318					1	1	29	ENSMUSP00000027961
		Sept8	2.89	0.0339					8	3	307	ENSMUSP00000104615
		Sf3b1	73.78	0.0342					1	1	37	ENSMUSP00000027127
		Eif5a	9.87	0.0344					5	5	468	ENSMUSP00000047008
		Hsp90ab1	6.10	0.0359					11	4	586	ENSMUSP00000024739
		Dhx36	38.13	0.0365					1	1	28	ENSMUSP00000029336
		Rbm28	2.42	0.0371					5	5	123	ENSMUSP00000007993
		Usp5	91.94	0.0422					1	1	20	ENSMUSP00000041299
		Cltc	4.34	0.0491					1	1	66	ENSMUSP00000050220

		Gene symbol	Fold change A/G[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(B)	Decreased binding at the rs7647481 G allele (risk)	Khdrbs1	0.12	0.0001	2 out of 3 criteria	X		transcription cofactor	7	7	355	ENSMUSP00000066516
		Ddx17	0.14	0.0001			X	transcription cofactor	28	20	1469	ENSMUSP00000055535
		Sfpq	0.30	0.0001		X		transcription cofactor	27	25	1437	ENSMUSP00000030623
		Fus	0.10	0.0002		X		transcription cofactor	13	11	1082	ENSMUSP00000076801
		Taf15	0.07	0.0003		X		transcription cofactor	5	3	386	ENSMUSP00000021018
		Hnrnpa2b1	0.18	0.0013		X		transcription cofactor	19	2	1182	ENSMUSP00000087453
		Hnrnpa2b1	0.14	0.0013		X		transcription cofactor	18	1	1114	ENSMUSP00000067491
		Dido1	0.22	0.0014				transcription cofactor	3	3	111	ENSMUSP00000084794
		Sf1	0.18	0.0022			X	transcription factor	7	7	339	ENSMUSP00000109115
		Rbm14	0.27	0.0152				transcription cofactor	6	6	248	ENSMUSP00000006625
		Ddx5	0.41	0.0170			X	transcription cofactor	24	15	1080	ENSMUSP00000021062
		Hnrnpl	0.08	0.0000	1 out of 3 criteria	X			27	1	2469	ENSMUSP00000134734
		Ewsr1	0.24	0.0007				transcription cofactor	8	7	419	ENSMUSP00000073034
		Nono	0.33	0.0009		X			18	16	1285	ENSMUSP00000033673
		Hnrnpd	0.44	0.0011		X			10	8	537	ENSMUSP00000019128
		Hnrnpab	0.48	0.0019		X			8	6	486	ENSMUSP00000074238
		Hnrpdl	0.43	0.0084		X			5	3	250	ENSMUSP00000084114
		Safb	0.50	0.0489		X			7	7	281	ENSMUSP00000092849
		Cpsf7	0.30	0.0000	none				1	1	35	ENSMUSP00000038958
		Rbmxl1	0.43	0.0000					11	3	326	ENSMUSP00000048153
		Xrn2	0.17	0.0000					16	16	641	ENSMUSP00000028921
		Aspg	0.11	0.0000					1	1	14	ENSMUSP00000078369
		Rbmx	0.19	0.0005					9	1	287	ENSMUSP00000110374
		Nudt21	0.32	0.0006					2	2	49	ENSMUSP00000034204
		Fam120c	0.02	0.0008					4	4	90	ENSMUSP00000073082
		Gm8991	0.35	0.0016					10	10	673	ENSMUSP00000072775
		Hnrnpa1	0.29	0.0016					15	13	1061	ENSMUSP00000084609
		Cirbp	0.44	0.0017					4	4	361	ENSMUSP00000101004
		Hnrpll	0.22	0.0020					8	8	335	ENSMUSP00000058308
		Bckdk	0.44	0.0024					1	1	25	ENSMUSP00000070345
		1810035L17Rik	0.41	0.0026					1	1	35	ENSMUSP00000076673
		Cdkn2aip	0.21	0.0028					7	6	271	ENSMUSP00000043713
		Pspc1	0.45	0.0064					12	12	526	ENSMUSP00000022507
		Rbm3	0.12	0.0095					9	9	554	ENSMUSP00000111277
		Tial1	0.33	0.0108					3	3	112	ENSMUSP00000033135
		Alkbh5	0.44	0.0137					4	4	165	ENSMUSP00000049116
		Rod1	0.36	0.0153					4	2	91	ENSMUSP00000030076
		U2af2	0.27	0.0166					9	9	497	ENSMUSP00000005041
		Hnrnpa0	0.44	0.0172					6	6	586	ENSMUSP00000007980
		Fam98a	0.47	0.0249					7	6	368	ENSMUSP00000108126
		Cacybp	0.47	0.0249					2	2	34	ENSMUSP00000014370
		Alyref	0.39	0.0254					9	9	763	ENSMUSP00000026125
		C330007P06Rik	0.30	0.0352					1	1	21	ENSMUSP00000040134
		Poldip3	0.10	0.0387					18	18	720	ENSMUSP00000054548

9.2.4 Supplementary table S2D: Classification of allele-specific binding proteins at the predicted *cis*-regulatory variant rs7647481 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 869 proteins binding at the rs7647481 surrounding genomic region (300 mM NaCl eluate of affinity chromatography). 142 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the A-allele / G-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms "DNA binding" and "transcription factor activity" were assessed for the total set of 869 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 142 proteins (fold-change > 2 or < 0.5, $P < 0.05$) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change A/G[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Increased binding at the rs7647481 A allele (non-risk)	Psip1	3.28	0.0044	3 out of 3 criteria	X	X	transcription cofactor	18	8	1066	ENSMUSP00000030207
		Ubp1	72.62	0.0104		X	X	transcription factor	1	1	73	ENSMUSP00000009885
		Trp53bp1	2.14	0.0165		X	X	transcription cofactor	3	2	118	ENSMUSP00000106277
		Kdm1a	168.01	0.0353		X	X	transcription cofactor	1	1	192	ENSMUSP00000035457
		Smarca4	75.10	0.0477		X	X	transcription cofactor	2	2	135	ENSMUSP00000034707
		Ruvbl2	3.08	0.0018	2 out of 3 criteria	X		transcription cofactor	3	3	88	ENSMUSP00000033087
		Mybbp1a	2.79	0.0042		X		transcription cofactor	10	10	349	ENSMUSP00000044827
		Maz	4.25	0.0064		X		transcription factor	9	8	331	ENSMUSP00000032916
		Ubtf	7.39	0.0070		X		general transcription factor (P	13	13	637	ENSMUSP00000006754
		Mcm7	12.62	0.0126		X		transcription cofactor	1	1	24	ENSMUSP00000000505
		Klf13	4.65	0.0443		X		transcription factor	1	1	40	ENSMUSP00000067680
		Phb	23.09	0.0481		X		transcription cofactor	1	1	51	ENSMUSP00000047536
		Xrcc1	4.74	0.0031	1 out of 3 criteria	X			8	7	228	ENSMUSP00000070995
		Pnkp	5.23	0.0042		X			9	9	294	ENSMUSP00000003044
		Dhx9	34.42	0.0067				transcription cofactor	3	3	228	ENSMUSP00000038135
		Mcm6	60.27	0.0079		X			2	2	102	ENSMUSP00000027601
		Top2a	5.27	0.0087		X			17	10	826	ENSMUSP00000068896
		Fen1	2.29	0.0087		X			8	8	449	ENSMUSP00000025651
		Tcp1	19.60	0.0101				transcription factor	11	11	751	ENSMUSP00000116108
		Polb	2.44	0.0105		X			14	14	719	ENSMUSP00000033938
		Top2b	2.01	0.0148		X			12	5	536	ENSMUSP00000017629
		Rfc2	3.81	0.0154		X			3	3	79	ENSMUSP00000023867
		Tmpo	2.98	0.0185		X			14	14	632	ENSMUSP00000020123
		Rfc3	3.19	0.0200		X			5	4	217	ENSMUSP00000039621
		Ddx21	27.16	0.0201				transcription cofactor	1	1	91	ENSMUSP00000042691
		Prpf19	11.63	0.0210		X			5	5	232	ENSMUSP00000025642
		Ruvbl1	268.46	0.0216				transcription cofactor	1	1	80	ENSMUSP00000032165
		Ddb1	2.06	0.0220		X			10	10	315	ENSMUSP00000025649
		Msh2	2.69	0.0225		X			22	20	1227	ENSMUSP00000024967
		Rfc1	2.90	0.0230		X			10	9	462	ENSMUSP00000031092
		Tdp1	70.90	0.0238		X			2	2	39	ENSMUSP00000021594
		Zfp148	2.51	0.0399				transcription factor	24	24	1609	ENSMUSP00000087106
		Msh6	4.08	0.0459		X			25	24	1098	ENSMUSP00000005503
		Dhx36	162.23	0.0004	none				1	1	28	ENSMUSP00000029336
		Rpl37	3.77	0.0005					1	1	13	ENSMUSP00000046506
		Snrpg	32.51	0.0005					2	1	82	ENSMUSP00000086987
		Snrpf	13.96	0.0007					4	4	286	ENSMUSP00000020203
		Dnajc9	68.42	0.0008					1	1	101	ENSMUSP00000022345
		Pds5a	3.88	0.0019					2	2	50	ENSMUSP00000031104
		Tomm70a	62.93	0.0021					1	1	38	ENSMUSP00000129186
		Eif4a1	21.24	0.0021					4	4	201	ENSMUSP00000127034
		Ddost	50.38	0.0025					1	1	16	ENSMUSP00000030538
		Atp5o	9.23	0.0031					1	1	16	ENSMUSP00000023677
		Dbn1	16.82	0.0035					1	1	52	ENSMUSP00000021950
		Fsd1	33.34	0.0036					1	1	24	ENSMUSP00000011733
		Olfr1395	2.67	0.0038					1	1	17	ENSMUSP00000050142
		Rtn4	2.52	0.0048					1	1	13	ENSMUSP00000077875
		Nop58	2.15	0.0054					9	9	405	ENSMUSP00000027174
		Luc7l	25.38	0.0059					1	1	18	ENSMUSP00000025023
		Cyb5r3	2.08	0.0070					4	4	99	ENSMUSP00000018186
		Rrbp1	2.02	0.0081					43	43	2380	ENSMUSP00000016072
		Nop2	4.29	0.0086					14	14	572	ENSMUSP00000047123
		Sf3a3	11.82	0.0088					2	2	203	ENSMUSP00000030734
		Rpn2	2.76	0.0095					2	2	65	ENSMUSP00000029171
		Eef1a1	24.10	0.0103					18	17	1066	ENSMUSP00000042457
		Usp5	205.48	0.0109					1	1	20	ENSMUSP00000041299
		Anp32e	23.96	0.0113					2	2	89	ENSMUSP00000015893
		Smu1	45.94	0.0125					2	2	219	ENSMUSP00000030117
		Nip7	4.60	0.0126					2	2	62	ENSMUSP00000034392
		Prpf40a	7.28	0.0132					2	2	92	ENSMUSP00000075655
		Hsp90ab1	11.42	0.0136					11	4	586	ENSMUSP00000024739
		Smc2	10.20	0.0138					3	3	90	ENSMUSP00000099979
		Slc25a3	4.02	0.0141					1	1	26	ENSMUSP00000075987
		Hnrnph3	17.80	0.0144					1	1	28	ENSMUSP00000020263
		Cnn3	434.12	0.0159					1	1	89	ENSMUSP00000029773
		Eftud2	26.54	0.0160					2	1	193	ENSMUSP00000021306
		Tmed10	3.78	0.0165					1	1	37	ENSMUSP00000037583
		Tuba1b	2.33	0.0181					14	1	861	ENSMUSP00000076777
		Thoc2	83.11	0.0182					2	2	176	ENSMUSP00000044677
		Eif3m	1944.28	0.0182					1	1	75	ENSMUSP00000028592
		Pycr2	2.21	0.0183					3	2	88	ENSMUSP00000027802
		Eif3h	23.20	0.0212					2	2	68	ENSMUSP00000022925
		Vim	9.48	0.0221					14	11	661	ENSMUSP00000028062
		Smc3	4.22	0.0245					8	8	207	ENSMUSP00000025930
		Snrpd1	2.43	0.0253					3	3	206	ENSMUSP00000002551
		Bzw1	95.06	0.0257					1	1	23	ENSMUSP00000051935
		Dcakd	7.74	0.0265					1	1	28	ENSMUSP00000021313
		Smchd1	14.51	0.0274					10	9	335	ENSMUSP00000121835
		Psmb7	76.26	0.0287					1	1	52	ENSMUSP00000028083
		Hsp90aa1	6.45	0.0306					11	5	433	ENSMUSP00000021698
		Rpl3	5.46	0.0327					7	7	333	ENSMUSP00000080354
		Eif5a	10.69	0.0356					5	5	468	ENSMUSP00000047008
		Exosc2	2.59	0.0366					1	1	13	ENSMUSP00000043519
		Rps23	7.30	0.0386					3	3	69	ENSMUSP00000054490
		Epb4.1	6.86	0.0418					1	1	37	ENSMUSP00000030739
		Rsl1d1	2.90	0.0423					1	1	46	ENSMUSP00000113431
		Mtap1b	107.73	0.0437					5	5	299	ENSMUSP00000068374
		Api5	10.41	0.0444					3	3	150	ENSMUSP00000028617
		Eef2	8.78	0.0453					13	12	499	ENSMUSP00000046101
		Tars	2.10	0.0459					4	4	169	ENSMUSP00000022849
		Nasp	12.17	0.0472					1	1	30	ENSMUSP00000030456
		Dync1h1	64.88	0.0476					5	5	355	ENSMUSP00000018851
		Fscn1	19.80	0.0478					1	1	103	ENSMUSP00000031565
		Nup205	172.62	0.0487					2	2	187	ENSMUSP00000039656

		Gene symbol	Fold change A/G[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(B)	**Decreased binding at the rs7647481 G allele (risk)**	Ddx1	0.20	0.0007	3 out of 3 criteria	X	X	transcription cofactor	23	23	1333	ENSMUSP00000065987
		Khdrbs1	0.21	0.0008	2 out o 3 criteria	X		transcription cofactor	7	7	355	ENSMUSP00000066516
		Smarcal1	0.19	0.0059		X		transcription cofactor	1	1	21	ENSMUSP00000047589
		Hnrnpa2b1	0.15	0.0117		X		transcription cofactor	19	2	1182	ENSMUSP00000087453
		Ddx17	0.36	0.0179			X	transcription cofactor	28	20	1469	ENSMUSP00000055535
		Sf1	0.43	0.0383			X	transcription factor	7	7	339	ENSMUSP00000109115
		Hnrnpl	0.01	0.0000	1 out o 3 criteria	X			27	1	2469	ENSMUSP00000134734
		Hnrnpd	0.15	0.0000		X			10	8	537	ENSMUSP00000019128
		Hnrnpl	0.02	0.0000		X			27	1	2448	ENSMUSP00000049407
		Hnrnpab	0.19	0.0007		X			8	6	486	ENSMUSP00000074238
		Hnrpdl	0.33	0.0131		X			5	3	250	ENSMUSP00000084114
		Rbms1	0.36	0.0152		X			2	2	78	ENSMUSP00000028347
		Rbm14	0.29	0.0176				transcription cofactor	6	6	248	ENSMUSP00000006625
		Dido1	0.12	0.0260				transcription cofactor	3	3	111	ENSMUSP00000084794
		Hnrnpu	0.29	0.0304		X			24	23	1201	ENSMUSP00000047571
		Cdkn2aip	0.04	0.0000	none				7	6	271	ENSMUSP00000043713
		Xrn2	0.07	0.0000					16	16	641	ENSMUSP00000028921
		Tial1	0.29	0.0001					3	3	112	ENSMUSP00000033135
		Fam120c	0.02	0.0002					4	4	90	ENSMUSP00000073082
		N4bp2l1	0.03	0.0002					4	4	120	ENSMUSP00000016279
		Sltm	0.22	0.0003					8	8	310	ENSMUSP00000049112
		Fam98a	0.10	0.0004					7	6	368	ENSMUSP00000108126
		Fam120a	0.11	0.0004					14	14	468	ENSMUSP00000053877
		D10Wsu52e	0.18	0.0004					18	18	993	ENSMUSP00000001834
		Rbmxl1	0.09	0.0012					11	3	326	ENSMUSP00000048153
		2700060E02Rik	0.22	0.0015					11	11	682	ENSMUSP00000022341
		Crat	0.36	0.0023					8	8	264	ENSMUSP00000028207
		Hnrnpr	0.41	0.0030					14	10	660	ENSMUSP00000081239
		Agap2	0.06	0.0031					1	1	13	ENSMUSP00000043466
		Rod1	0.14	0.0042					4	2	91	ENSMUSP00000030076
		Eefsec	0.01	0.0047					2	2	54	ENSMUSP00000131207
		Dnahc17	0.49	0.0049					1	1	15	ENSMUSP00000081864
		Cirbp	0.16	0.0052					4	4	361	ENSMUSP00000101004
		Rbm3	0.13	0.0054					9	9	554	ENSMUSP00000111277
		Syncrip	0.46	0.0076					15	11	753	ENSMUSP00000063744
		Wdr46	0.38	0.0088					3	3	88	ENSMUSP00000025170
		Arpc1b	0.29	0.0093					3	2	178	ENSMUSP00000082822
		U2af2	0.35	0.0119					9	9	497	ENSMUSP00000005041
		Aspg	0.21	0.0182					1	1	14	ENSMUSP00000078369
		Rpusd4	0.13	0.0183					1	1	13	ENSMUSP00000034543
		Asph	0.05	0.0195					3	3	74	ENSMUSP00000077273
		Jup	0.21	0.0304					5	5	155	ENSMUSP00000001592
		8430406I07Rik	0.44	0.0354					4	4	155	ENSMUSP00000028910
		Apoa1bp	0.21	0.0361					1	1	20	ENSMUSP00000029708
		Hnrpll	0.35	0.0383					8	8	335	ENSMUSP00000058308
		Srbd1	0.18	0.0387					24	23	918	ENSMUSP00000092810
		Acaa2	0.32	0.0481					1	1	41	ENSMUSP00000037348
		Hnrnpa0	0.39	0.0497					6	6	586	ENSMUSP00000007980

9.2.5 Supplementary table S2E: Classification of allele-specific binding proteins at the predicted non *cis*-regulatory variant rs17036342 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 952 proteins binding at the rs17036342 surrounding genomic region (200 mM NaCl eluate of affinity chromatography). 29 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the A-allele / G-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms "DNA binding" and "transcription factor activity" were assessed for the total set of 952 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 29 proteins (fold-change > 2 or < 0.5, *P* < 0.05) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change A/G[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Binding at the rs17036342 A allele (risk)	Ubp1	81.22	0.0072	3 out o 3 criteria	X	X	transcription factor	1	1	57	ENSMUSP00000009885
		Taf6l	26.40	0.0206		X	X	transcription cofactor	1	1	34	ENSMUSP00000003777
		Kdm1a	108.43	0.0332		X	X	transcription cofactor	1	1	243	ENSMUSP00000035457
		Ebf2	4.58	0.0009	2 out o 3 criteria	X		transcription factor	6	3	224	ENSMUSP00000022637
		Tead1	2.65	0.0049		X		transcription factor	3	2	125	ENSMUSP00000060671
		Ebf1	5.41	0.0129		X		transcription factor	4	1	202	ENSMUSP00000080020
		Mcm7	15.63	0.0297		X		transcription cofactor	1	1	50	ENSMUSP00000000505
		Mcm6	37.78	0.0170	1 out of 3 criteria	X			2	2	124	ENSMUSP00000027601
		Dhx9	17.18	0.0180				transcription cofactor	5	5	256	ENSMUSP00000038135
		Akap8	2.45	0.0367		X			1	1	33	ENSMUSP00000002699
		Snrpg	44.41	0.0023	none				1	1	46	ENSMUSP00000086987
		Sympk	771.51	0.0076					2	1	44	ENSMUSP00000023882
		Tomm70a	39.28	0.0087					1	1	29	ENSMUSP00000129186
		Dnajc9	40.21	0.0091					1	1	127	ENSMUSP00000022345
		Nup210	27.79	0.0201					1	1	15	ENSMUSP00000032179
		Dbn1	10.89	0.0215					1	1	65	ENSMUSP00000021950
		Snrpf	3.63	0.0218					4	4	265	ENSMUSP00000020203
		Sf3a3	9.23	0.0273					2	2	227	ENSMUSP00000030734
		Dhx36	109.93	0.0304					1	1	35	ENSMUSP00000029336
		mt-Co2	2.18	0.0344					1	1	23	ENSMUSP00000080994
		Prpf40a	5.35	0.0346					2	2	66	ENSMUSP00000075655
		Sf3b4	7.87	0.0403					1	1	41	ENSMUSP00000075709
		Gnb1	2.26	0.0407					6	2	250	ENSMUSP00000030940
		Nup160	7.60	0.0422					1	1	21	ENSMUSP00000059289
		Thoc2	29.83	0.0445					2	2	148	ENSMUSP00000044677
		Hsp90aa1	4.26	0.0448					7	4	361	ENSMUSP00000021698
		Ina	55.21	0.0469					2	1	64	ENSMUSP00000041347
		Fam98b	4.53	0.0479					1	1	49	ENSMUSP00000028825
		Naa38	62.51	0.0488					1	1	93	ENSMUSP00000057238
(B)		none										

9.2.6 Supplementary table S2F: Classification of allele-specific binding proteins at the predicted non *cis*-regulatory variant rs17036342 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 951 proteins binding at the rs17036342 surrounding genomic region (300 mM NaCl eluate of affinity chromatography). 44 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the A-allele / G-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms "DNA binding" and "transcription factor activity" were assessed for the total set of 951 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 44 proteins (fold-change > 2 or < 0.5, $P < 0.05$) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change A/G[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Binding at the rs17036342 A allele (risk)	Smarca4	103.82	0.0392	3 out of 3 criteria	X	X	transcription cofactor	2	2	108	ENSMUSP00000034707
		Kdm1a	97.54	0.0448		X	X	transcription cofactor	1	1	243	ENSMUSP00000035457
		Ubp1	16.01	0.0500		X	X	transcription factor	1	1	57	ENSMUSP00000009885
		Tead1	9.81	0.0200	2 out of 3 criteria	X		transcription factor	3	2	125	ENSMUSP00000060671
		Hcfc1	5.74	0.0241			X	transcription cofactor	9	9	219	ENSMUSP00000033761
		Mcm7	22.71	0.0348		X		transcription coactor	1	1	50	ENSMUSP00000000505
		Mcm6	42.08	0.0159	1 out o 3 criteria	X			2	2	124	ENSMUSP00000027601
		Ddx21	20.17	0.0182				transcription cofactor	2	1	93	ENSMUSP00000042691
		Ruvbl1	158.53	0.0211				transcription cofactor	1	1	73	ENSMUSP00000032165
		Dhx9	19.71	0.0214				transcription cofactor	5	5	256	ENSMUSP00000038135
		Tcp1	10.67	0.0218				transcription factor	10	9	721	ENSMUSP00000086418
		Bud31	3.67	0.0403				involved in transcription regula	2	2	81	ENSMUSP00000124999
		Mcm2	106.41	0.0413		X			1	1	75	ENSMUSP00000061923
		Ebf1	6.86	0.0462				transcription factor	4	1	202	ENSMUSP00000080020
		Tomm70a	76.12	0.0067	none				1	1	29	ENSMUSP00000129186
		mt-Co2	3.30	0.0107					1	1	23	ENSMUSP00000080994
		Snrpg	17.72	0.0109					1	1	46	ENSMUSP00000086987
		Prpf40a	4.94	0.0111					2	2	66	ENSMUSP00000075655
		Eif3m	6731.04	0.0114					1	1	69	ENSMUSP00000028592
		Hsp90aa1	9.09	0.0131					7	4	361	ENSMUSP00000021698
		Dbn1	9.11	0.0142					1	1	65	ENSMUSP00000021950
		Slain2	4.22	0.0145					1	1	39	ENSMUSP00000115871
		Slc25a3	2.90	0.0150					1	1	27	ENSMUSP00000075987
		Nup98	790.36	0.0171					1	1	31	ENSMUSP00000068530
		Eif4a1	9.32	0.0186					7	7	412	ENSMUSP00000127034
		Sf3a3	18.73	0.0209					2	2	227	ENSMUSP00000030734
		Flna	3.37	0.0209					11	10	360	ENSMUSP00000033699
		Fscn1	14.19	0.0220					1	1	52	ENSMUSP00000031565
		Lima1	2.77	0.0237					1	1	25	ENSMUSP00000073371
		Hsp90ab1	5.31	0.0243					10	6	581	ENSMUSP00000024739
		Snrpf	4.91	0.0255					4	4	265	ENSMUSP00000020203
		Eef1a1	16.09	0.0258					14	14	1117	ENSMUSP00000042457
		Eftud2	20.26	0.0267					3	2	287	ENSMUSP00000021306
		Ldha	2.09	0.0267					10	10	284	ENSMUSP00000036386
		Eif5a	7.31	0.0288					5	5	535	ENSMUSP00000047008
		Anp32e	11.04	0.0321					2	2	112	ENSMUSP00000015893
		Dhx36	66.34	0.0333					1	1	35	ENSMUSP00000029336
		Eef1g	4.51	0.0374					8	8	439	ENSMUSP00000093955
		Eef2	5.02	0.0418					15	14	694	ENSMUSP00000046101
		Ina	170.95	0.0441					2	1	64	ENSMUSP00000041347
		Sympk	105.99	0.0471					2	1	44	ENSMUSP00000023882
		Vcp	2.26	0.0499					13	13	580	ENSMUSP00000030164
(B)	Binding at the rs17036342 G allele (non-risk)	Ssbp1	0.35	0.0291	2 out of 3 criteria	X		transcription cofactor	5	5	296	ENSMUSP00000110427
		Ddx47	0.39	0.0321	none				1	1	15	ENSMUSP00000032326

9.2.7 Supplementary table S2G: Classification of allele-specific binding proteins at the predicted non *cis*-regulatory variant rs2881479 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 932 proteins binding at the rs2881479 surrounding genomic region (200 mM NaCl eluate of affinity chromatography). 25 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the A-allele / T-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms "DNA binding" and "transcription factor activity" were assessed for the total set of 932 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 25 proteins (fold-change > 2 or < 0.5, *P* < 0.05) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change A/T[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Binding at the rs2881479 A allele (risk)	Ubp1	76.94	0.0041	3 out of 3 criteria	X	X	transcription factor	1	1	61	ENSMUSP00000009885
		Trrap	83.05	0.0470	2 out o 3 criteria		X	transcription cofactor	1	1	20	ENSMUSP00000042544
		Ddx21	13.97	0.0267	1 out of 3 criteria			transcription cofactor	1	1	96	ENSMUSP00000042691
		Mcm3	37.81	0.0468		X			1	1	101	ENSMUSP00000059192
		Snrpg	44.70	0.0013	none				1	1	42	ENSMUSP00000086987
		Tomm70a	123.21	0.0046					1	1	36	ENSMUSP00000129186
		Snrpf	4.45	0.0075					4	4	251	ENSMUSP00000020203
		Dbn1	18.37	0.0130					1	1	121	ENSMUSP00000021950
		Sf3a3	11.17	0.0170					2	2	231	ENSMUSP00000030734
		Cap1	27.64	0.0175					1	1	30	ENSMUSP00000068260
		Fsd1	14.43	0.0190					1	1	21	ENSMUSP00000011733
		Ina	133.18	0.0236					3	1	126	ENSMUSP00000041347
		Dnajc9	22.73	0.0237					1	1	109	ENSMUSP00000022345
		Hsp90aa1	6.30	0.0251					10	5	452	ENSMUSP00000021698
		Cnn3	17.69	0.0303					1	1	125	ENSMUSP00000029773
		Asf1a	190.09	0.0303					1	1	25	ENSMUSP00000020004
		Eftud2	9.63	0.0312					3	2	153	ENSMUSP00000021306
		Prpf40a	3.78	0.0346					3	3	148	ENSMUSP00000075655
		Thoc2	32.65	0.0423					2	2	211	ENSMUSP00000044677
		Atic	80.89	0.0430					1	1	33	ENSMUSP00000027384
		Kpna3	17.36	0.0482					1	1	90	ENSMUSP00000022496
(B)	Binding at the rs2881479 T allele	Sik3	0.33	0.0162	none				1	1	14	ENSMUSP00000112749
		Dsg1c	0.07	0.0208					1	1	18	ENSMUSP00000054799
		Jakmip2	0.12	0.0371					1	1	16	ENSMUSP00000080881
		Jup	0.05	0.0458					9	9	264	ENSMUSP00000001592

9.2.8 Supplementary table S2H: Classification of allele-specific binding proteins at the predicted non *cis*-regulatory variant rs2881479 using GO-term analysis and transcription factor annotation

Label-free quantitative proteomic analysis identified in total 933 proteins binding at the rs2881479 surrounding genomic region (300 mM NaCl eluate of affinity chromatography). 82 proteins with a significance allelic fold-change > 2.0 or < 0.5 ((A) and (B), respectively; normalized mean protein abundance from three independent experiments, comparing the ratio of the A-allele / T-allele, *P*-value < 0.05, unpaired t-test) are shown. GO-terms "DNA binding" and "transcription factor activity" were assessed for the total set of 933 identified proteins FDR < 1% using the GePS tool (Genomatix, Munich, Germany). Proteins found in both, the respective GO-term output-lists and the list of 82 proteins (fold-change > 2 or < 0.5, $P < 0.05$) are indicated. Moreover, proteins were analyzed for transcription factor and cofactor annotation using MatBase tool (Genomatix, Munich, Germany). Further, the total number "Peptide count" of peptides identified or the number of uniquely "Peptide count for quantitation" identified peptides per protein, and the summed up "Mascot Percolator score" as indicator for the reliability of protein identification are displayed. Based on fold-change and *P*-value ranking, on the selection criteria GO-term overlap and TF-annotation proteins were categorized to assign candidates to mediate allele-specific *cis*-regulatory activity.

		Gene symbol	Fold change A/T[a]	*P-value*[b]	Selection criteria[c]	GO DNA binding	GO transcription factor activity	transcription factor and cofactor annotation (Genomatix)	Peptide count[d]	Peptide count for quantitation[e]	Mascot Percolator score[f]	protein accession number
(A)	Binding at the rs2881479 A allele (risk)	Ubp1	198.52	0.0034	3 out of 3 criteria	X	X	transcription factor	1	1	61	ENSMUSP00000009885
		Smarcc2	502.79	0.0381		X	X	transcription cofactor	2	1	155	ENSMUSP00000026433
		Kdm1a	287.29	0.0417		X	X	transcription cofactor	1	1	213	ENSMUSP00000035457
		Mcm7	20.94	0.0178	2 out of 3 criteria	X		transcription cofactor	1	1	42	ENSMUSP00000000505
		Thoc1	9.50	0.0083	1 out o 3 criteria	X			1	1	33	ENSMUSP00000025137
		Tcea1	3.44	0.0114				general transcription factor (P	7	7	269	ENSMUSP00000080266
		Pfn1	3.73	0.0128				transcription cofactor	2	2	35	ENSMUSP00000018437
		Ddx21	38.91	0.0151				transcription cofactor	1	1	96	ENSMUSP00000042691
		Mcm3	65.33	0.0179		X			1	1	101	ENSMUSP00000059192
		Top2a	2.62	0.0214		X			20	13	929	ENSMUSP00000068896
		Hnrnpl	15.79	0.0222		X			18	18	1439	ENSMUSP00000049407
		Pcna	2.28	0.0227		X			5	5	218	ENSMUSP00000028817
		Tcp1	16.17	0.0228				transcription factor	9	8	729	ENSMUSP00000116108
		Mcm6	685.96	0.0248		X			1	1	74	ENSMUSP00000027601
		Ddx3x	3.20	0.0327		X			15	13	651	ENSMUSP00000000804
		Msh6	3.44	0.0390		X			25	25	1132	ENSMUSP00000005503
		Dhx9	9.74	0.0417				transcription cofactor	4	4	323	ENSMUSP00000038135
		Snrpg	30.73	0.0008	none				1	1	42	ENSMUSP00000086987
		Atic	7795.06	0.0010					1	1	33	ENSMUSP00000027384
		Cnn3	145.21	0.0014					1	1	125	ENSMUSP00000029773
		Snrpf	6.98	0.0016					4	4	251	ENSMUSP00000020203
		Tomm70a	124.34	0.0020					1	1	36	ENSMUSP00000129186
		Msn	2.41	0.0022					8	5	339	ENSMUSP00000113071
		Dhx15	4.85	0.0027					3	3	91	ENSMUSP00000031061
		Fsd1	23.42	0.0029					1	1	21	ENSMUSP00000011733
		Dnajc9	58.14	0.0034					1	1	109	ENSMUSP00000022345
		Pkm	2.21	0.0037					14	13	812	ENSMUSP00000034834
		Dbn1	24.27	0.0041					1	1	121	ENSMUSP00000021950
		Eif4a1	9.49	0.0046					6	6	219	ENSMUSP00000127034
		Hspa9	2.41	0.0048					8	8	466	ENSMUSP00000025217
		Vcp	2.39	0.0062					10	10	494	ENSMUSP00000030164
		Anp32e	18.03	0.0067					2	2	81	ENSMUSP00000015893
		Hsp90aa1	18.66	0.0075					10	5	452	ENSMUSP00000021698
		Eftud2	51.68	0.0099					3	2	153	ENSMUSP00000021306
		Sf3a3	15.58	0.0103					2	2	231	ENSMUSP00000030734
		Ddost	15.68	0.0104					1	1	22	ENSMUSP00000030538
		Cap1	606.08	0.0111					1	1	30	ENSMUSP00000068260
		Etfa	3.04	0.0125					4	4	202	ENSMUSP00000034866
		Nasp	73.78	0.0131					1	1	123	ENSMUSP00000030456
		Pgaml	5.89	0.0134					4	4	115	ENSMUSP00000011896
		Thoc2	90.90	0.0136					2	2	211	ENSMUSP00000044677
		Eef1a1	17.94	0.0138					14	13	1076	ENSMUSP00000042457
		Flna	12.67	0.0147					4	4	212	ENSMUSP00000033699
		Shmt2	2.08	0.0152					6	6	361	ENSMUSP00000026470
		Fscn1	44.85	0.0157					1	1	115	ENSMUSP00000031565
		Gm5506	6.75	0.0163					8	8	560	ENSMUSP00000075513
		Hnrnph3	185.51	0.0183					1	1	18	ENSMUSP00000020263
		Eef2	16.25	0.0202					10	9	492	ENSMUSP00000046101
		Coro1b	3.21	0.0206					5	5	227	ENSMUSP00000008893
		Oxct1	4.17	0.0212					2	2	139	ENSMUSP00000106318
		Ina	145.70	0.0213					3	1	126	ENSMUSP00000041347
		Hnrnph1	17.92	0.0220					4	2	171	ENSMUSP00000070503
		Kdelc1	3.84	0.0232					2	2	102	ENSMUSP00000027213
		Hnrnph2	7.69	0.0235					2	1	62	ENSMUSP00000050838
		Cyp2j6	5.49	0.0238					1	1	13	ENSMUSP00000030303
		Farsb	4.39	0.0254					8	8	267	ENSMUSP00000069508
		Eif3h	15.31	0.0262					2	2	59	ENSMUSP00000022925
		Lamb1	2.48	0.0270					1	1	25	ENSMUSP00000002979
		Farsa	2.30	0.0288					5	5	243	ENSMUSP00000003906
		Usp5	46.03	0.0289					1	1	21	ENSMUSP00000041299
		Rpl37a	2.25	0.0311					1	1	25	ENSMUSP00000058919
		Atp5b	2.12	0.0317					9	9	542	ENSMUSP00000026459
		Smu1	42.38	0.0323					2	2	159	ENSMUSP00000030117
		Hsp90ab1	5.43	0.0332					12	6	616	ENSMUSP00000024739
		Cct7	28.78	0.0351					9	9	576	ENSMUSP00000032078
		Eif3m	579.30	0.0352					1	1	68	ENSMUSP00000028592
		Ldhb	71.97	0.0373					1	1	33	ENSMUSP00000032373
		Eif2a	2.01	0.0382					14	14	623	ENSMUSP00000029387
		Rpl3	2.48	0.0387					12	11	641	ENSMUSP00000080354
		Rps23	5.69	0.0397					3	3	139	ENSMUSP00000054490
		Epb4.1	24.86	0.0409					1	1	17	ENSMUSP00000030739
		Spnb2	44.87	0.0420					2	2	56	ENSMUSP00000006629
		Cct8	6.84	0.0424					5	5	140	ENSMUSP00000026704
		2610101N10Rik	85.89	0.0430					1	1	25	ENSMUSP00000077482
		Arpc3	3.73	0.0435					1	1	57	ENSMUSP00000031421
		Ywhag	2.14	0.0446					6	4	241	ENSMUSP00000051223
		Prdx2	171.64	0.0460					1	1	52	ENSMUSP00000005292
		AC239834.2	4.29	0.0471					12	12	768	ENSMUSP00000100550
		Ddx39b	2.47	0.0485					11	10	436	ENSMUSP00000070682
(B)	Binding at the rs2881479 T allele (non-risk)	Asxl2	0.40	0.0227	1 out o 3 criteria			transcription cofactor	4	4	118	ENSMUSP00000106846
		Vdac1	0.45	0.0242	none				1	1	21	ENSMUSP00000020673
		Sik3	0.32	0.0401					1	1	14	ENSMUSP00000112749

9.3 Supplementary table S3: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted *cis*-regulatory variant rs4684847 (A, B), rs7647481 (C, D) and non *cis*-regulatory variant rs17036342 (E, F) and rs2881479 (G, H)

For all supplementary tables S3: [a]*P*-values were derived from Fisher´s exact test, [b]Genes observed refers to the number of genes within the input list associated canonical signal transduction pathways (GePS-tool, Genomatix, based on NCI-nature Pathway Interaction Database (http://pid.nci.nih.gov) and The Cancer Cell Map (www.pathwaycommons.org)), [c]Genes expected refers to the number of genes which would expected randomly, [d]Genes total refers to the total number of genes within canonical signal transduction pathways (GePS-tool, Genomatix, based on NCI-nature Pathway Interaction Database (http://pid.nci.nih.gov) and The Cancer Cell Map (http://www.pathwaycommons.org)).

9.3.1 Supplementary table S3A: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted *cis*-regulatory variant rs4684847

In total 41 proteins binding at the rs4684847 surrounding genomic region (200 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 8 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
E2F transcription factor network	NCI-nature:e2f_pathway	5.84E-04	3	0.1820	78	ATM, YY1, MCM3	472, 7528, 4172	11920, 22632, 17215
hypoxia and p53 in the cardiovascular system	BioCarta:p53hypoxiapathway	1.51E-03	2	0.0607	26	ATM, HSP90AA1	472, 3320	11920, 15519
tumor suppressor arf inhibits ribosomal biogenesis	BioCarta:arfpathway	1.75E-03	2	0.0653	28	ATM, HSP90AA1	472, 3320	11920, 15519
Regulation of glucocorticoid receptor	NCI-nature:reg_gr_pathway	2.88E-03	2	0.0840	36	HSP90AA1, PCK2	3320, 5106	15519, 74551
p53 pathway	NCI-nature:p53regulationpathway	7.63E-03	2	0.1376	59	ATM, YY1	472, 7528	11920, 22632
ahr signal transduction pathway	BioCarta:ahrpathway	9.30E-03	1	0.0093	4	HSP90AA1	3320	15519
Regulation of Telomerase	NCI-nature:telomerasepathway	1.09E-02	2	0.1656	71	ATM, HSP90AA1	472, 3320	11920, 15519
multi-drug resistance factors	BioCarta:mrppathway	1.39E-02	1	0.0140	6	GSTP1	2950	14870
cdc25 and chk1 regulatory pathway in response to dna damage	BioCarta:cdc25pathway	1.62E-02	1	0.0163	7	ATM	472	11920
PLK3 signaling events	NCI-nature:plk3_pathway	1.62E-02	1	0.0163	7	ATM	472	11920
rb tumor suppressor/checkpoint signaling in response to dna damage	BioCarta:rbpathway	3.00E-02	1	0.0303	13	ATM	472	11920
apoptotic signaling in response to dna damage	BioCarta:chemicalpathway	3.22E-02	1	0.0327	14	ATM	472	11920
the prc2 complex sets long-term gene silencing through modification of histone tails	BioCarta:prc2pathway	3.22E-02	1	0.0327	14	YY1	7528	22632
ErbB receptor signaling network	NCI-nature:erbb_network_pathway	3.45E-02	1	0.0350	15	HSP90AA1	3320	15519
hypoxia-inducible factor in the cardivascular system	BioCarta:hifpathway	3.68E-02	1	0.0373	16	HSP90AA1	3320	15519
cdk regulation of dna replication	BioCarta:mcmpathway	4.13E-02	1	0.0420	18	MCM3	4172	17215
regulation of cell cycle progression by plk3	BioCarta:plk3pathway	4.13E-02	1	0.0420	18	ATM	472	11920
atm signaling pathway	BioCarta:atmpathway	4.35E-02	1	0.0443	19	ATM	472	11920
akt signaling pathway	BioCarta:aktpathway	4.35E-02	1	0.0443	19	HSP90AA1	3320	15519
Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	NCI-nature:hif1apathway	4.58E-02	1	0.0467	20	HSP90AA1	3320	15519

9.3.2 Supplementary table S3B: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted *cis*-regulatory variant rs4684847

In total 165 proteins binding at the rs4684847 surrounding genomic region (300 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 38 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
mechanisms of transcriptional repression by dna methylation	BioCarta:mbdpathway	1.61E-05	4	0.1662	15	CHD4, HDAC2, RBBP7, RBBP4	1108, 3066, 5931, 5928	107932, 15182, 245688, 19646
Regulation of Telomerase	NCI-nature:telomerasepathway	1.04E-04	6	0.7868	71	HDAC2, HSP90AA1, TGFB1, RBBP7, ATM, RBBP4	3066, 3320, 7040, 5931, 472, 5928	15182, 15519, 21803, 245688, 11920, 19646
the prc2 complex sets long-term gene silencing through modification of histone tails	BioCarta:prc2pathway	4.20E-04	3	0.1551	14	HDAC2, RBBP7, RBBP4	3066, 5931, 5928	15182, 245688, 19646
antisense pathway	BioCarta:antisensepathway	7.08E-04	2	0.0443	4	SFPQ, NONO	6421, 4841	71514, 53610
Hedgehog signaling events mediated by Gli proteins	NCI-nature:hedgehog_glipathway	1.76E-03	4	0.5319	48	HDAC2, CSNK1A1, RBBP7, RBBP4	3066, 1452, 5931, 5928	15182, 93687, 245688, 19646
hypoxia and p53 in the cardiovascular system	BioCarta:p53hypoxiapathway	2.74E-03	3	0.2881	26	CSNK1A1, HSP90AA1, ATM	1452, 3320, 472	93687, 15519, 11920
tumor suppressor arf inhibits ribosomal biogenesis	BioCarta:arfpathway	3.40E-03	3	0.3103	28	CSNK1A1, HSP90AA1, ATM	1452, 3320, 472	93687, 15519, 11920
Signaling events mediated by HDAC Class I	NCI-nature:hdac_classi_pathway	8.88E-03	4	0.8311	75	CHD4, HDAC2, RBBP7, RBBP4	1108, 3066, 5931, 5928	107932, 15182, 245688, 19646
Urokinase-type plasminogen activator (uPA) and uPAR-mediated signaling	NCI-nature:upa_upar_pathway	1.29E-02	3	0.4987	45	FN1, TGFB1, MMP12	2335, 7040, 4321	14268, 21803, 17381
Integrin-linked kinase signaling	NCI-nature:ilk_pathway	1.37E-02	3	0.5098	46	PARP1, RUVBL1, HSP90AA1	142, 8607, 3320	11545, 56505, 15519
Alpha6Beta4Integrin	CellMap:Alpha6Beta4Integrin	1.81E-02	3	0.5652	51	VIM, DSP, PLEC	7431, 1832, 5339	22352, 109620, 18810
opposing roles of aif in apoptosis and cell survival	BioCarta:aifpathway	3.29E-02	1	0.0332	3	PARP1	142	11545
cell cycle: g1/s check point	BioCarta:g1pathway	3.53E-02	2	0.2992	27	TGFB1, ATM	7040, 472	21803, 11920
ahr signal transduction pathway	BioCarta:ahrpathway	4.36E-02	1	0.0443	4	HSP90AA1	3320	15519
mTOR signaling pathway	NCI-nature:mtor_4pathway	4.55E-02	2	0.3435	31	EEF2, EIF4B	1938, 1975	13629, 75705
BARD1 signaling events	NCI-nature:bard1pathway	4.81E-02	2	0.3546	32	EWSR1, ATM	2130, 472	14030, 11920

9.3.3 Supplementary table S3C: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted *cis*-regulatory variant rs7647481

In total 108 proteins binding at the rs7647481 surrounding genomic region (200 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 29 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
antisense pathway	BioCarta:antisensepathway	4.10E-04	2	0.0338	4	SFPQ, NONO	6421, 4841	71514, 53610
overview of telomerase rna component gene hterc transcriptional regulation	BioCarta:tercpathway	1.41E-03	2	0.0592	7	SP3, SP1	6670, 6667	20687, 20683
E2F transcription factor network	NCI-nature:e2f_pathway	3.82E-03	4	0.6597	78	SP1, XRCC1, RBBP4, YY1	6667, 7515, 5928, 7528	20683, 22594, 19646, 22632
ATR signaling pathway	NCI-nature:atr_pathway	4.05E-03	3	0.3298	39	RFC3, RFC2, MCM7	5983, 5982, 4176	69263, 19718, 17220
effects of calcineurin in keratinocyte differentiation	BioCarta:calcineurinpathway	5.09E-03	2	0.1099	13	SP3, SP1	6670, 6667	20687, 20683
the prc2 complex sets long-term gene silencing through modification of histone tails	BioCarta:prc2pathway	5.90E-03	2	0.1184	14	RBBP4, YY1	5928, 7528	19646, 22632
Regulation of retinoblastoma protein	NCI-nature:rb_1pathway	1.81E-02	3	0.5666	67	DNMT1, UBTF, RBBP4	1786, 7343, 5928	13433, 21429, 19646
Regulation of Telomerase	NCI-nature:telomerasepathway	2.11E-02	3	0.6005	71	SP3, SP1, RBBP4	6670, 6667, 5928	20687, 20683, 19646
Signaling events mediated by HDAC Class I	NCI-nature:hdac_classi_pathway	2.44E-02	3	0.6343	75	NUP210, RBBP4, YY1	23225, 5928, 7528	54563, 19646, 22632
Regulation of nuclear SMAD2/3 signaling	NCI-nature:smad2_3nuclearpathway	3.18E-02	3	0.7020	83	SP3, SP1, RBBP4	6670, 6667, 5928	20687, 20683, 19646
FOXM1 transcription factor network	NCI-nature:foxm1pathway	4.83E-02	2	0.3552	42	SP1, XRCC1	6667, 7515	20683, 22594

9.3.4 Supplementary table S3D: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted *cis*-regulatory variant rs7647481

In total 142 proteins binding at the rs7647481 surrounding genomic region (300 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 32 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
ATR signaling pathway	NCI-nature:atr_pathway	4.09E-04	4	0.3640	39	SMARCAL1, RFC3, RFC2, MCM7	50485, 5983, 5982, 4176	54380, 69263, 19718, 17220
Validated targets of C-MYC transcriptional activation	NCI-nature:myc_activpathway	8.08E-03	4	0.8119	87	UBTF, RUVBL1, HSP90AA1, RUVBL2	7343, 8607, 3320, 10856	21429, 56505, 15519, 20174
Integrin-linked kinase signaling	NCI-nature:ilk_pathway	8.54E-03	3	0.4293	46	RUVBL1, HSP90AA1, RUVBL2	8607, 3320, 10856	56505, 15519, 20174
hypoxia-inducible factor in the cardivascular system	BioCarta:hifpathway	9.33E-03	2	0.1493	16	ASPH, HSP90AA1	444, 3320	65973, 15519
cdk regulation of dna replication	BioCarta:mcmpathway	1.18E-02	2	0.1680	18	MCM6, MCM7	4175, 4176	17219, 17220
Regulation of C-MYC	NCI-nature:myc_pathway	2.05E-02	2	0.2240	24	RUVBL1, RUVBL2	8607, 10856	56505, 20174
Lissencephaly gene (LIS1) in neuronal migration and development	NCI-nature:lis1pathway	3.32E-02	2	0.2893	31	MAP1B, DYNC1H1	4131, 1778	17755, 13424
ahr signal transduction pathway	BioCarta:ahrpathway	3.68E-02	1	0.0373	4	HSP90AA1	3320	15519
ATM pathway	NCI-nature:atm_pathway	4.15E-02	2	0.3266	35	TP53BP1, SMC3	7158, 9126	27223, 13006
Regulation of glucocorticoid receptor	NCI-nature:reg_gr_pathway	4.36E-02	2	0.3360	36	SMARCA4, HSP90AA1	6597, 3320	20586, 15519

9.3.5 Supplementary table S3E: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted non *cis*-regulatory variant rs17036342

In total 29 proteins binding at the rs17036342 surrounding genomic region (200 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 7 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
Signaling events mediated by HDAC Class II	NCI-nature:hdac_classii_pathway	5.37E-05	3	0.0837	41	GNB1, NUP210, HSP90AA1	2782, 23225, 3320	14688, 54563, 15519
cdk regulation of dna replication	BioCarta:mcmpathway	5.38E-04	2	0.0367	18	MCM6, MCM7	4175, 4176	17219, 17220
corticosteroids and cardioprotection	BioCarta:gcrpathway	1.13E-03	2	0.0531	26	GNB1, HSP90AA1	2782, 3320	14688, 15519
actions of nitric oxide in the heart	BioCarta:no1pathway	2.96E-03	2	0.0857	42	GNB1, HSP90AA1	2782, 3320	14688, 15519
ion channels and their functional role in vascular endothelium	BioCarta:raccpathway	3.10E-03	2	0.0878	43	GNB1, HSP90AA1	2782, 3320	14688, 15519
TNF alpha/NF-kB	CellMap:TNF_alpha_NF_kB	4.67E-03	3	0.3797	186	MCM7, AKAP8, HSP90AA1	4176, 10270, 3320	17220, 56399, 15519
ahr signal transduction pathway	BioCarta:ahrpathway	8.14E-03	1	0.0082	4	HSP90AA1	3320	15519
regulation of spermatogenesis by crem	BioCarta:crempathway	2.03E-02	1	0.0204	10	GNB1	2782	14688
g-protein signaling through tubby proteins	BioCarta:tubbypathway	2.23E-02	1	0.0225	11	GNB1	2782	14688
cxcr4 signaling pathway	BioCarta:cxcr4pathway	2.23E-02	1	0.0225	11	GNB1	2782	14688
mechanism of protein import into the nucleus	BioCarta:npcpathway	2.43E-02	1	0.0245	12	NUP210	23225	54563
visual signal transduction	BioCarta:rhodopsinpathway	2.63E-02	1	0.0265	13	GNB1	2782	14688
akap95 role in mitosis and chromosome dynamics	BioCarta:akap95pathway	2.63E-02	1	0.0265	13	AKAP8	10270	56399
cycling of ran in nucleocytoplasmic transport	BioCarta:ranpathway	2.83E-02	1	0.0286	14	NUP210	23225	54563
attenuation of gpcr signaling	BioCarta:agpcrpathway	2.83E-02	1	0.0286	14	GNB1	2782	14688
ErbB receptor signaling network	NCI-nature:erbb_network_pathway	3.02E-02	1	0.0306	15	HSP90AA1	3320	15519
Sumoylation by RanBP2 regulates transcriptional repression	NCI-nature:ranbp2pathway	3.02E-02	1	0.0306	15	NUP210	23225	54563
hypoxia-inducible factor in the cardivascular system	BioCarta:hifpathway	3.22E-02	1	0.0327	16	HSP90AA1	3320	15519
aspirin blocks signaling pathway involved in platelet activation	BioCarta:sppapathway	3.42E-02	1	0.0347	17	GNB1	2782	14688
sumoylation by ranbp2 regulates transcriptional repression	BioCarta:ranbp2pathway	3.82E-02	1	0.0388	19	NUP210	23225	54563
akt signaling pathway	BioCarta:aktpathway	3.82E-02	1	0.0388	19	HSP90AA1	3320	15519
Noncanonical Wnt signaling pathway	NCI-nature:wnt_calcium_pathway	4.01E-02	1	0.0408	20	GNB1	2782	14688
cystic fibrosis transmembrane conductance regulator (cftr) and beta 2 adrenergic receptor (b2ar) pathway	BioCarta:cftrpathway	4.01E-02	1	0.0408	20	GNB1	2782	14688
Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	NCI-nature:hif1apathway	4.01E-02	1	0.0408	20	HSP90AA1	3320	15519
pkc-catalyzed phosphorylation of inhibitory phosphoprotein of myosin phosphatase	BioCarta:myosinpathway	4.41E-02	1	0.0449	22	GNB1	2782	14688
ccr3 signaling in eosinophils	BioCarta:ccr3pathway	4.60E-02	1	0.0470	23	GNB1	2782	14688
how progesterone initiates the oocyte maturation	BioCarta:mprpathway	4.60E-02	1	0.0470	23	GNB1	2782	14688
Visual signal transduction: Rods	NCI-nature:rhodopsin_pathway	4.80E-02	1	0.0490	24	GNB1	2782	14688
regulation of ck1/cdk5 by type 1 glutamate receptors	BioCarta:ck1pathway	4.80E-02	1	0.0490	24	GNB1	2782	14688

9.3.6 Supplementary table S3F: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted non *cis*-regulatory variant rs17036342

In total 44 proteins binding at the rs17036342 surrounding genomic region (300 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 11 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
cdk regulation of dna replication	BioCarta:mcmpathway	1.95E-05	3	0.0577	18	MCM6, MCM2, MCM7	4175, 4171, 4176	17219, 17216, 17220
TNF alpha/NF-kB	CellMap:TNF_alpha_NF_kB	1.57E-04	5	0.5967	186	HSP90AB1, SMARCA4, FLNA, MCM7, HSP90AA1	3326, 6597, 2316, 4176, 3320	15516, 20586, 192176, 17220, 15519
Regulation of glucocorticoid receptor	NCI-nature:reg_gr_pathway	5.55E-03	2	0.1155	36	SMARCA4, HSP90AA1	6597, 3320	20586, 15519
ATR signaling pathway	NCI-nature:atr_pathway	6.50E-03	2	0.1251	39	MCM2, MCM7	4171, 4176	17216, 17220
Integrin-linked kinase signaling	NCI-nature:ilk_pathway	8.97E-03	2	0.1476	46	RUVBL1, HSP90AA1	8607, 3320	56505, 15519
ahr signal transduction pathway	BioCarta:ahrpathway	1.28E-02	1	0.0128	4	HSP90AA1	3320	15519
srebp control of lipid synthesis	BioCarta:s1ppathway	2.54E-02	1	0.0257	8	LIMA1	51474	65970
Validated targets of C-MYC transcriptional activation	NCI-nature:myc_activpathway	3.02E-02	2	0.2791	87	RUVBL1, HSP90AA1	8607, 3320	56505, 15519
Validated nuclear estrogen receptor beta network	NCI-nature:erb_genomic_pathway	4.71E-02	1	0.0481	15	SMARCA4	6597	20586
Validated nuclear estrogen receptor beta network	NCI-nature:erb_genomic_pathway-1	4.71E-02	1	0.0481	15	SMARCA4	6597	20586
ErbB receptor signaling network	NCI-nature:erbb_network_pathway	4.71E-02	1	0.0481	15	HSP90AA1	3320	15519

9.3.7 Supplementary table S3G: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted non *cis*-regulatory variant rs2881479

In total 25 proteins binding at the rs2881479 surrounding genomic region (200 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, P < 0.05 (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways (P < 0.05, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 6 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
LKB1 signaling events	NCI-nature:lkb1_pathway	2.23E-03	2	0.0752	43	SIK3, HSP90AA1	23387, 3320	70661, 15519
ahr signal transduction pathway	BioCarta:ahrpathway	6.98E-03	1	0.0070	4	HSP90AA1	3320	15519
E2F transcription factor network	NCI-nature:e2f_pathway	7.22E-03	2	0.1365	78	TRRAP, MCM3	8295, 4172	100683, 17215
Validated targets of C-MYC transcriptional activation	NCI-nature:myc_activpathway	8.93E-03	2	0.1522	87	TRRAP, HSP90AA1	8295, 3320	100683, 15519
ErbB receptor signaling network	NCI-nature:erbb_network_pathway	2.60E-02	1	0.0262	15	HSP90AA1	3320	15519
hypoxia-inducible factor in the cardivascular system	BioCarta:hifpathway	2.77E-02	1	0.0280	16	HSP90AA1	3320	15519
cdk regulation of dna replication	BioCarta:mcmpathway	3.11E-02	1	0.0315	18	MCM3	4172	17215
akt signaling pathway	BioCarta:aktpathway	3.28E-02	1	0.0332	19	HSP90AA1	3320	15519
Hypoxic and oxygen homeostasis regulation of HIF-1-alpha	NCI-nature:hif1apathway	3.45E-02	1	0.0350	20	HSP90AA1	3320	15519
TNF alpha/NF-kB	CellMap:TNF_alpha_NF_kB	3.80E-02	2	0.3255	186	KPNA3, HSP90AA1	3839, 3320	16648, 15519
how progesterone initiates the oocyte maturation	BioCarta:mprpathway	3.96E-02	1	0.0402	23	CAP1	10487	12331
Regulation of C-MYC	NCI-nature:myc_pathway	4.13E-02	1	0.0420	24	TRRAP	8295	100683
hypoxia and p53 in the cardiovascular system	BioCarta:p53hypoxiapathway	4.47E-02	1	0.0455	26	HSP90AA1	3320	15519
corticosteroids and cardioprotection	BioCarta:gcrpathway	4.47E-02	1	0.0455	26	HSP90AA1	3320	15519
tumor suppressor arf inhibits ribosomal biogenesis	BioCarta:arfpathway	4.80E-02	1	0.0490	28	HSP90AA1	3320	15519
VEGFR1 specific signals	NCI-nature:vegfr1_pathway	4.80E-02	1	0.0490	28	HSP90AA1	3320	15519

9.3.8 Supplementary table S3H: Canonical signaling pathways overrepresented in the set of significant allele-specific binding proteins at the predicted non *cis*-regulatory variant rs2881479

In total 82 proteins binding at the rs2881479 surrounding genomic region (300 mM NaCl eluate of affinity chromatography) with a significance allelic fold-change > 2.0 or < 0.5, $P < 0.05$ (normalized mean protein abundance from three independent experiments) were assessed to canonical signaling pathways using the GePS tool (Genomatix, Munich, Germany). Overrepresentation of canonical signaling pathways ($P < 0.05$, enrichment of identified proteins, Fisher´s exact test) are shown with the " Pathway ID", the respective *P*-value, the number of "Genes (observed)", the number of "Genes (expected)", the number of "Genes (total)", the "List of observed genes", the "Gene IDs (Human)" and "Original gene IDs (Mouse)". A total of 21 input genes were subjected to the analysis.

Canonical pathway	Pathway ID	*P-value*[a]	Genes (observed)[b]	Genes (expected)[c]	Genes (total)[d]	List of observed genes	Gene IDs (Human)	Original gene IDs (Mouse)
TNF alpha/NF-kB	CellMap:TNF_alpha_NF_kB	7.50E-05	7	1.1391	186	MCM7, HSP90AA1, HSP90AB1, FLNA, SMARCC2, YWHAG, DDX3X	4176, 3320, 3326, 2316, 6601, 7532, 1654	17220, 15519, 15516, 192176, 68094, 22628, 13205
cdk regulation of dna replication	BioCarta:mcmpathway	1.52E-04	3	0.1102	18	MCM6, MCM7, MCM3	4175, 4176, 4172	17219, 17220, 17215
LKB1 signaling events	NCI-nature:lkb1_pathway	2.09E-03	3	0.2633	43	HSP90AA1, SIK3, YWHAG	3320, 23387, 7532	15519, 70661, 22628
Regulation of glucocorticoid receptor	NCI-nature:reg_gr_pathway	1.99E-02	2	0.2205	36	HSP90AA1, SMARCC2	3320, 6601	15519, 68094
ahr signal transduction pathway	BioCarta:ahrpathway	2.43E-02	1	0.0245	4	HSP90AA1	3320	15519
Class I PI3K signaling events mediated by Akt	NCI-nature:pi3kciaktpathway	2.90E-02	2	0.2695	44	HSP90AA1, YWHAG	3320, 7532	15519, 22628
a6b1 and a6b4 Integrin signaling	NCI-nature:a6b1_a6b4_integrin_pathway	2.90E-02	2	0.2695	44	YWHAG, LAMB1	7532, 3912	22628, 16777
TGF-beta receptor signaling	NCI-nature:tgfbrpathway	3.94E-02	2	0.3185	52	SPTBN1, EIF2A	6711, 83939	20742, 229317

9.4 Supplementary table S4: Transcriptional cofactors identified at the predicted *cis*-regulatory variants rs4684847 and the rs7647481

Label-free quantitative proteomic analysis identified in total 824 -869 proteins binding at the predicted *cis*-regulatory variants. All identified proteins which were annotated as transcription cofactor (MatBase tool, Genomatix, Munich, Germany) are listed. Both sets of coregulators were used for calculation of enrichment of cofactor-identifications co-cited with the prioritized transcription factors PRRX1 (at the rs4684847 variant), YY1 and NFAFC4 (both at the rs7647481 variant) and for network interaction analysis (details see methods).

cofactors identified at the rs4684847 adjacent regions	cofactors identified at the 7647481 adjacent regions
Calr	Apex1
Apex1	Atn1
Brd4	Bcl9
Ddx17	Calr
Ddx21	Cbx3
Ddx5	Crebzf
Dek	Ddx17
Dhx9	Ddx21
Ewsr1	Ddx5
Fus	Dek
Gatad2b	Dhx9
Hcfc1	Dido1
Hmgb1	Ewsr1
Hmgb2	Fus
Hmgb3	Hcfc1
Hmgn1	Hmgb1
Hnrnpa2b1	Hmgb2
Khdrbs1	Hmgb3
Mbd3	Hmgn1
Mecp2	Hnrnpa2b1
Morf4l2	Hnrnpa2b1
Mta1	Khdrbs1
Mta2	Morf4l2
Mtdh	Mta2
Mybbp1a	Mtdh
Npm1	Mybbp1a
Pa2g4	Ncoa5
Pfn1	Npm1
Ptma	Pa2g4
Puf60	Pfn1
Rbbp4	Phb
Rbbp7	Ptma
Rbm39	Puf60
Rps3	Rbbp4
Rybp	Rbm14
Sfpq	Rbm39
Ssbp1	Rps3
Sub1	Rybp
Supt16h	Sfpq
Tardbp	Ssbp1
Trim28	Sub1
Uhrf1	Supt16h
Wdr5	Taf6l
Ywhah	Tardbp
Ywhaq	Trim28
	Uhrf1
	Wdr5
	Yaf2
	Ywhah
	Ywhaq

10. Literature

1. Wild, S. Roglic, G. Green, A. Sicree, R. & King, H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030, *Diabetes Care* **27,** 1047–1053 (2004).
2. International Diabetes Federation. IDF Diabetes Atlas, 6th edn. Brussels, Belgium: International Diabetes Federation.
3. WHO fact sheet (No. 311) on obesity and overweight. http://www.who.int/mediacentre/factsheets/fs311/en/index.html. (Accessed April 2014).
4. NHLBI Obesity Education Initiative Expert Panel on the Identification, Evaluation, and Treatment of Obesity in Adults (US). Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report. Bethesda (MD): National Heart, Lung, and Blood Institute; 1998 Sep. Report No.: 98-4083. Available from: http://www.ncbi.nlm.nih.gov/books/NBK2003/ ,
5. Guilherme, A. Virbasius, J. V. Puri, V. & Czech, M. P. Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes, *Nat Rev Mol Cell Biol* **9,** 367–377 (2008).
6. Halmos, T. New diagnostic and classification system in diabetic syndrome, *Orv Hetil* **143,** 2533–2541 (2002).
7. Barma, P. D. Ranabir, S. Prasad, L. & Singh, T. P. Clinical and biochemical profile of lean type 2 diabetes mellitus, *Indian J Endocrinol Metab* **15,** S40-3 (2011).
8. Unnikrishnan, A. G. Singh, S. K. & Sanjeevi, C. B. Prevalence of GAD65 antibodies in lean subjects with type 2 diabetes, *Ann N Y Acad Sci* **1037,** 118–121 (2004).
9. Polonsky, K. S. Dynamics of insulin secretion in obesity and diabetes, *Int J Obes Relat Metab Disord* **24 Suppl 2,** S29-31 (2000).
10. Karelis, A. D. To be obese--does it matter if you are metabolically healthy?, *Nat Rev Endocrinol* **7,** 699–700 (2011).
11. Velho, S. Paccaud, F. Waeber, G. Vollenweider, P. & Marques-Vidal, P. Metabolically healthy obesity: different prevalences using different criteria, *Eur J Clin Nutr* **64,** 1043–1051 (2010).
12. Appleton, S. L. *et al.* Diabetes and cardiovascular disease outcomes in the metabolically healthy obese phenotype: a cohort study, *Diabetes Care* **36,** 2388–2394 (2013).
13. Eckel, R. H. *et al.* Obesity and type 2 diabetes: what can be unified and what needs to be individualized?, *The Journal of clinical endocrinology and metabolism* **96,** 1654–1663 (2011).
14. Goran, M. I. Ball, Geoff D C & Cruz, M. L. Obesity and risk of type 2 diabetes and cardiovascular disease in children and adolescents, *J Clin Endocrinol Metab* **88,** 1417–1427 (2003).
15. Hall, M. A. *et al.* Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank, *Pac Symp Biocomput,* 200–211 (2014).
16. Franks, P. W. The complex interplay of genetic and lifestyle risk factors in type 2 diabetes: an overview, *Scientifica (Cairo)* **2012,** 482186 (2012).
17. Lyssenko, V. *et al.* Clinical risk factors, DNA variants, and the development of type 2 diabetes, *N Engl J Med* **359,** 2220–2232 (2008).
18. Hindorff LA, M. J. W. A. J. H. H. P. K. A. a. M. T. A Catalog of Published Genome-Wide Association Studies. Available from: www.genome.gov/gwastudies.
19. Diagnosis and Classification of Diabetes Mellitus, *Diabetes Care* **34,** S62 (2011).
20. Alberti, K. G. & Zimmet, P. Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation, *Diabet Med* **15,** 539–553 (1998).
21. Soltesz, G. Diabetes in the young: a paediatric and epidemiological perspective, *Diabetologia* **46,** 447–454 (2003).
22. Rother, K. I. Diabetes treatment--bridging the divide, *N Engl J Med* **356,** 1499–1501 (2007).

23. Barrett-Connor, E. Epidemiology, obesity, and non-insulin-dependent diabetes mellitus, *Epidemiol Rev* **11,** 172–181 (1989).

24. Leto, D. & Saltiel, A. R. Regulation of glucose transport by insulin: traffic control of GLUT4, *Nat Rev Mol Cell Biol* **13,** 383–396 (2012).

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

26. Morton, G. J. Hypothalamic leptin regulation of energy homeostasis and glucose metabolism, *J Physiol* **583,** 437–443 (2007).

27. Martin-Merino, E. Fortuny, J. Rivero-Ferrer, E. & Garcia-Rodriguez, L. A. Incidence of retinal complications in a cohort of newly diagnosed diabetic patients, *PLoS ONE* **9,** e100283 (2014).

28. Morrish, N. J. Wang, S. L. Stevens, L. K. Fuller, J. H. & Keen, H. Mortality and causes of death in the WHO Multinational Study of Vascular Disease in Diabetes, *Diabetologia* **44 Suppl 2,** S14-21 (2001).

29. Tesfaye, S. *et al.* Diabetic neuropathies: update on definitions, diagnostic criteria, estimation of severity, and treatments, *Diabetes Care* **33,** 2285–2293 (2010).

30. Al-Rubeaan, K. *et al.* Diabetic Nephropathy and Its Risk Factors in a Society with a Type 2 Diabetes Epidemic: A Saudi National Diabetes Registry-Based Study, *PLoS ONE* **9,** e88956 EP - (2014).

31. Schwenk, R. W. Vogel, H. & Schurmann, A. Genetic and epigenetic control of metabolic health, *Mol Metab* **2,** 337–347 (2013).

32. Aggarwal, B. B. Targeting Inflammation-Induced Obesity and Metabolic Diseases by Curcumin and Other Nutraceuticals, *Annu. Rev. Nutr.* **30,** 173–199 (2010).

33. Schwartz, M. W. Woods, S. C. Porte, D. JR, Seeley, R. J. & Baskin, D. G. Central nervous system control of food intake, *Nature* **404,** 661–671 (2000).

34. Meister, B. Control of food intake via leptin receptors in the hypothalamus, *Vitam Horm* **59,** 265–304 (2000).

35. Williams, L. M. Hypothalamic dysfunction in obesity, *Proc Nutr Soc* **71,** 521–533 (2012).

36. Boden, G. Obesity, insulin resistance and free fatty acids, *Curr Opin Endocrinol Diabetes Obes* **18,** 139–143 (2011).

37. Boden, G. Obesity and free fatty acids, *Endocrinol Metab Clin North Am* **37,** 635-46, viii-ix (2008).

38. Westley, R. L. & May, Felicity E B. A twenty-first century cancer epidemic caused by obesity: the involvement of insulin, diabetes, and insulin-like growth factors, *Int J Endocrinol* **2013,** 632461 (2013).

39. Unger, R. H. & Zhou, Y. T. Lipotoxicity of beta-cells in obesity and in other causes of fatty acid spillover, *Diabetes* **50 Suppl 1,** S118-21 (2001).

40. Chandrasekera, P. C. & Pippin, J. J. Of rodents and men: species-specific glucose regulation and type 2 diabetes research, *ALTEX* **31,** 157–176 (2014).

41. Noble, D. Mathur, R. Dent, T. Meads, C. & Greenhalgh, T. Risk models and scores for type 2 diabetes: systematic review, *BMJ* **343,** d7163 (2011).

42. Gunasekaran, U. & Gannon, M. Type 2 diabetes and the aging pancreatic beta cell, *Aging (Albany NY)* **3,** 565–575 (2011).

43. Szoke, E. *et al.* Effect of aging on glucose homeostasis: accelerated deterioration of beta-cell function in individuals with impaired glucose tolerance, *Diabetes Care* **31,** 539–543 (2008).

44. Age- and sex-specific prevalences of diabetes and impaired glucose regulation in 13 European cohorts, *Diabetes Care* **26,** 61–69 (2003).

45. Iozzo, P. *et al.* Independent influence of age on basal insulin secretion in nondiabetic humans. European Group for the Study of Insulin Resistance, *J Clin Endocrinol Metab* **84,** 863–868 (1999).

46. Dechenes, C. J. Verchere, C. B. Andrikopoulos, S. & Kahn, S. E. Human aging is associated with parallel reductions in insulin and amylin release, *Am J Physiol* **275,** E785-91 (1998).

47. Meisinger, C. *et al.* Sex differences in risk factors for incident type 2 diabetes mellitus: the MONICA Augsburg cohort study, *Arch Intern Med* **162,** 82–89 (2002).

48. Van Buren, Dorothy J & Tibbs, T. L. Lifestyle interventions to reduce diabetes and cardiovascular disease risk among children, *Curr Diab Rep* **14,** 557 (2014).

49. Medici, F. Hawa, M. Ianari, A. Pyke, D. A. & Leslie, R. D. Concordance rate for type II diabetes mellitus in monozygotic twins: actuarial analysis, *Diabetologia* **42,** 146–150 (1999).

50. Poulsen, P. *et al.* Heritability of insulin secretion, peripheral and hepatic insulin action, and intracellular glucose partitioning in young and old Danish twins, *Diabetes* **54,** 275–283 (2005).

51. Meigs, J. B. Cupples, L. A. & Wilson, P. W. Parental transmission of type 2 diabetes: the Framingham Offspring Study, *Diabetes* **49,** 2201–2207 (2000).

52. Vassy, J. L. & Meigs, J. B. Is genetic testing useful to predict type 2 diabetes?, *Best Pract Res Clin Endocrinol Metab* **26,** 189–201 (2012).

53. Bluher, S. & Schwarz, P. Metabolically healthy obesity from childhood to adulthood - Does weight status alone matter?, *Metabolism* **63,** 1084–1092 (2014).

54. Lyssenko, V. & Laakso, M. Genetic screening for the risk of type 2 diabetes: worthless or valuable?, *Diabetes Care* **36 Suppl 2,** S120-6 (2013).

55. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits, *PNAS* **106,** 9362–9367 (2009).

56. Lutsey, P. L. Pereira, M. A. Bertoni, A. G. Kandula, N. R. & Jacobs, David R Jr. Interactions between race/ethnicity and anthropometry in risk of incident diabetes: the multi-ethnic study of atherosclerosis, *Am J Epidemiol* **172,** 197–204 (2010).

57. Shai, I. *et al.* Ethnicity, obesity, and risk of type 2 diabetes in women: a 20-year follow-up study, *Diabetes Care* **29,** 1585–1590 (2006).

58. The National Center for Chronic Disease Prevention and Health Promotion: Agespecific prevalence of diagnosed diabetes, by race/ethnicity and sex, United States [article online], 2004. Available from:www.cdc.gov/diabetes/statistics/prev/national/fig2004.htm. Accessed 13 March 2005 (2004).

59. Abate, N. & Chandalia, M. The impact of ethnicity on type 2 diabetes, *J Diabetes Complications* **17,** 39–58 (2003).

60. Astrup, A. & Finer, N. Redefining type 2 diabetes: 'diabesity' or 'obesity dependent diabetes mellitus?, *Obes Rev* **1,** 57–59 (2000).

61. Sanderson, S. C. *et al.* Genetic and lifestyle causal beliefs about obesity and associated diseases among ethnically diverse patients: a structured interview study, *Public Health Genomics* **16,** 83–93 (2013).

62. Hu, F. B. Globalization of diabetes: the role of diet, lifestyle, and genes, *Diabetes Care* **34,** 1249–1257 (2011).

63. Marti, A. Moreno-Aliaga, M. J. Hebebrand, J. & Martinez, J. A. Genes, lifestyles and obesity, *Int J Obes Relat Metab Disord* **28 Suppl 3,** S29-36 (2004).

64. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology, *Nature* **518,** 197–206 (2015).

65. Herrera, B. M. Keildson, S. & Lindgren, C. M. Genetics and epigenetics of obesity, *Maturitas* **69,** 41–49 (2011).

66. Rahati, S. Shahraki, M. Arjomand, G. & Shahraki, T. Food pattern, lifestyle and diabetes mellitus, *Int J High Risk Behav Addict* **3,** e8725 (2014).

67. Misra, A. Ramchandran, A. Jayawardena, R. Shrivastava, U. & Snehalatha, C. Diabetes in South Asians, *Diabet Med* (2014).

68. Tuomilehto, J. *et al.* Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impaired glucose tolerance, *N Engl J Med* **344,** 1343–1350 (2001).

69. Hu, F. B. *et al.* Diet, lifestyle, and the risk of type 2 diabetes mellitus in women, *N Engl J Med* **345,** 790–797 (2001).

70. Ding, M. Bhupathiraju, S. N. Chen, M. van Dam, Rob M & Hu, F. B. Caffeinated and decaffeinated coffee consumption and risk of type 2 diabetes: a systematic review and a dose-response meta-analysis, *Diabetes Care* **37,** 569–586 (2014).

71. Willi, C. Bodenmann, P. Ghali, W. A. Faris, P. D. & Cornuz, J. Active smoking and the risk of type 2 diabetes: a systematic review and meta-analysis, *JAMA* **298,** 2654–2664 (2007).

72. Altshuler, D. Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease, *Science* **322,** 881–888 (2008).

73. Sturtevant, A. H. A THIRD GROUP OF LINKED GENES IN DROSOPHILA AMPELOPHILA, *Science* **37,** 990–992 (1913).

74. Ahlqvist, E. Ahluwalia, T. S. & Groop, L. Genetics of Type 2 Diabetes, *Clinical Chemistry* **57,** 241–254 (2011).

75. Rahim, N. G. Harismendy, O. Topol, E. J. & Frazer, K. A. Genetic determinants of phenotypic diversity in humans, *Genome Biol* **9,** 215 (2008).

76. Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes, *Nat Genet* **38,** 320–323 (2006).

77. Stranger, B. E. Stahl, E. A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics, *Genetics* **187,** 367–383 (2011).

78. Farooqi, I. S. *et al.* Beneficial effects of leptin on obesity, T cell hyporesponsiveness, and neuroendocrine/metabolic dysfunction of human congenital leptin deficiency, *J Clin Invest* **110,** 1093–1103 (2002).

79. Pearson, E. R. *et al.* Switching from insulin to oral sulfonylureas in patients with diabetes due to Kir6.2 mutations, *N Engl J Med* **355,** 467–477 (2006).

80. O'Rahilly, S. Human genetics illuminates the paths to metabolic disease, *Nature* **462,** 307–314 (2009).

81. Teumer, A. *et al.* Comparison of genotyping using pooled DNA samples (allelotyping) and individual genotyping using the affymetrix genome-wide human SNP array 6.0, *BMC Genomics* **14,** 506 (2013).

82. Sham, P. Bader, J. S. Craig, I. O'Donovan, M. & Owen, M. DNA Pooling: a tool for large-scale association studies, *Nat Rev Genet* **3,** 862–871 (2002).

83. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs, *Nature* **449,** 851–861 (2007).

84. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations, *Nature* **467,** 52–58 (2010).

85. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations, *Nucleic Acids Res* **42,** D1001-6 (2014).

86. Sun, X. Yu, W. & Hu, C. Genetics of Type 2 Diabetes: Insights into the Pathogenesis and Its Clinical Application, *Biomed Res Int* **2014,** 926713 (2014).

87. Yudkin, J. S. Kumari, M. Humphries, S. E. & Mohamed-Ali, V. Inflammation, obesity, stress and coronary heart disease: is interleukin-6 the link?, *Atherosclerosis* **148,** 209–214 (2000).

88. Florez, J. C. *et al.* A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets, *Diabetes* **56,** 3063–3074 (2007).

89. Hanson, R. L. *et al.* A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array, *Diabetes* **56,** 3045–3052 (2007).

90. Hayes, M. G. *et al.* Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies, *Diabetes* **56,** 3033–3044 (2007).

91. Rampersaud, E. *et al.* Identification of novel candidate genes for type 2 diabetes from a genome-wide association scan in the Old Order Amish: evidence for replication from diabetes-related quantitative traits and from independent populations, *Diabetes* **56,** 3053–3062 (2007).

92. Salonen, J. T. *et al.* Type 2 diabetes whole-genome association study in four populations: the DiaGen consortium, *Am J Hum Genet* **81,** 338–345 (2007).

93. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes, *Nature* **518,** 317–330 (2015).

94. Gerhardt, C. C. Romero, I. A. Cancello, R. Camoin, L. & Strosberg, A. D. Chemokines control fat accumulation and leptin secretion by cultured human adipocytes, *Molecular and Cellular Endocrinology* **175,** 81–92 (2001).

95. Ginsburg, D. Genetics and Genomics to the Clinic: A Long Road ahead, *Cell* **147,** 17–19 (2011).

96. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing, *Nat.Rev.Genet* **11,** 415–425 (2010).

97. Cipolletta, D. *et al.* PPAR-γ is a major driver of the accumulation and phenotype of adipose tissue Treg cells, *Nature* (2012).

98. Claussnitzer, M. *et al.* Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms, *Cell* **156,** 343–358 (2014).

99. Brunner, C. Laumen, H. Nielsen, P. J. Kraut, N. & Wirth, T. Expression of the Aldehyde Dehydrogenase 2-like Gene Is Controlled by BOB.1/OBF.1 in B Lymphocytes, *J.Biol.Chem* **278,** 45231–45239 (2003).

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

101. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci, *Nat Commun* **6,** 6326 (2015).

102. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3, *Nature* **507,** 371–375 (2014).

103. Stancakova, A. *et al.* Association of 18 confirmed susceptibility loci for type 2 diabetes with indices of insulin release, proinsulin conversion, and insulin sensitivity in 5,327 nondiabetic Finnish men, *Diabetes* **58,** 2129–2136 (2009).

104. Palmer, N. D. *et al.* Quantitative trait analysis of type 2 diabetes susceptibility loci identified from whole genome association studies in the Insulin Resistance Atherosclerosis Family Study, *Diabetes* **57,** 1093–1100 (2008).

105. Palmer, N. D. *et al.* Association of TCF7L2 gene polymorphisms with reduced acute insulin response in Hispanic Americans, *J Clin Endocrinol Metab* **93,** 304–309 (2008).

106. Zhong, H. *et al.* Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes, *PLoS Genet* **6,** e1000932 (2010).

107. Frayling, T. M. *et al.* A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity, *Science* **316,** 889–894 (2007).

108. Scott, L. J. *et al.* A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants, *Science* **316,** 1341–1345 (2007).

109. Dina, C. *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity, *Nat Genet* **39,** 724–726 (2007).

110. Gerken, T. *et al.* The Obesity-Associated FTO Gene Encodes a 2-Oxoglutarate-Dependent Nucleic Acid Demethylase, *Science* **318,** 1469–1472 (2007).

111. Cecil, J. E. Tavendale, R. Watt, P. Hetherington, M. M. & Palmer, Colin N A. An obesity-associated FTO gene variant and increased energy intake in children, *N Engl J Med* **359,** 2558–2566 (2008).

112. Xi, B. *et al.* Common polymorphism near the MC4R gene is associated with type 2 diabetes: data from a meta-analysis of 123,373 individuals, *Diabetologia* **55,** 2660–2666 (2012).

113. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index, *Nat Genet* **42,** 937–948 (2010).

114. Scherag, A. *et al.* Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups, *PLoS Genet* **6,** e1000916 (2010).

115. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity, *Nat Genet* **41,** 18–24 (2009).

116. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation, *Nat Genet* **41,** 25–34 (2009).

117. Loos, Ruth J F *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity, *Nat Genet* **40,** 768–775 (2008).

118. Issemann, I. & Green, S. Activation of a member of the steroid hormone receptor superfamily by peroxisome proliferators, *Nature* **347,** 645–650 (1990).

119. Dreyer, C. *et al.* Control of the peroxisomal beta-oxidation pathway by a novel family of nuclear hormone receptors, *Cell* **68,** 879–887 (1992).

120. Michalik, L. Desvergne, B. & Wahli, W. Peroxisome-proliferator-activated receptors and cancers: complex stories, *Nat Rev Cancer* **4,** 61–70 (2004).

121. Savkur, R. S. & Miller, A. R. Investigational PPAR-γ agonists for the treatment of Type 2 diabetes, *Expert Opin. Investig. Drugs* **15,** 763–778 (2006).

122. Nielsen, R. Grontved, L. Stunnenberg, H. G. & Mandrup, S. Peroxisome Proliferator-Activated Receptor Subtype- and Cell-Type-Specific Activation of Genomic Target Genes upon Adenoviral Transgene Delivery, *Mol.Cell.Biol* **26,** 5698–5714 (2006).

123. Yu, S. & Reddy, J. K. Transcription coactivators for peroxisome proliferator-activated receptors, *Biochim Biophys Acta* **1771,** 936–951 (2007).

124. Brun, R. P. *et al.* Differential activation of adipogenesis by multiple PPAR isoforms, *Genes Dev* **10,** 974–984 (1996).

125. Paterniti, I. *et al.* Molecular evidence for the involvement of PPAR-δ and PPAR-γ in anti-inflammatory and neuroprotective activities of palmitoylethanolamide after spinal cord trauma, *J Neuroinflammation* **10,** 20 (2013).

126. Moller, D. E. & Berger, J. P. Role of PPARs in the regulation of obesity-related insulin sensitivity and inflammation, *Int J Obes Relat Metab Disord* **27,** S17 (2003).

127. Greene, M. E. *et al.* Isolation of the human peroxisome proliferator activated receptor gamma cDNA: expression in hematopoietic cells and chromosomal mapping, *Gene Expr* **4,** 281–299 (1995).

128. Jiang, M. *et al.* Disruption of PPARgamma signaling results in mouse prostatic intraepithelial neoplasia involving active autophagy, *Cell Death Differ* **17,** 469–481 (2010).

129. Heikkinen, S. Auwerx, J. & Argmann, C. A. PPARgamma in human and mouse physiology, *Biochim Biophys Acta* **1771,** 999–1013 (2007).

130. Vidal-Puig, A. *et al.* Regulation of PPAR gamma gene expression by nutrition and obesity in rodents, *J Clin Invest* **97,** 2553–2561 (1996).

131. Berger, J. *et al.* Thiazolidinediones produce a conformational change in peroxisomal proliferator-activated receptor-gamma: binding and activation correlate with antidiabetic actions in db/db mice, *Endocrinology* **137,** 4189–4195 (1996).

132. Sun, K. Wang, Q. & Huang, X.-H. PPAR gamma inhibits growth of rat hepatic stellate cells and TGF beta-induced connective tissue growth factor expression, *Acta Pharmacol Sin* **27,** 715–723 (2006).

133. Law, R. E. *et al.* Expression and function of PPARgamma in rat and human vascular smooth muscle cells, *Circulation* **101,** 1311–1318 (2000).

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

135. Fajas, L. *et al.* The Organization, Promoter Analysis, and Expression of the Human PPARγ Gene, *J.Biol.Chem* **272,** 18779–18789 (1997).

136. Mukherjee, R. Jow, L. Croston, G. E. & Paterniti, J R Jr. Identification, characterization, and tissue distribution of human peroxisome proliferator-activated receptor (PPAR) isoforms PPARgamma2 versus PPARgamma1 and activation with retinoid X receptor agonists and antagonists, *J Biol Chem* **272,** 8071–8076 (1997).

137. Ricote, M. Li, A. C. Willson, T. M. Kelly, C. J. & Glass, C. K. The peroxisome proliferator-activated receptor-gamma is a negative regulator of macrophage activation, *Nature* **391,** 79–82 (1998).

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

139. Poulsen, Lars la Cour, Siersbaek, M. & Mandrup, S. PPARs: fatty acid sensors controlling metabolism, *Semin Cell Dev Biol* **23,** 631–639 (2012).

140. Lehmann, J. M. *et al.* An antidiabetic thiazolidinedione is a high affinity ligand for peroxisome proliferator-activated receptor gamma (PPAR gamma), *J Biol Chem* **270,** 12953–12956 (1995).

141. Tontonoz, P. Hu, E. Graves, R. A. Budavari, A. I. & Spiegelman, B. M. mPPAR gamma 2: tissue-specific regulator of an adipocyte enhancer, *Genes & Development* **8,** 1224–1234 (1994).

142. Kung, J. & Henry, R. R. Thiazolidinedione safety, *Expert Opin Drug Saf* **11,** 565–579 (2012).

143. Jones, A. B. Peroxisome proliferator-activated receptor (PPAR) modulators: diabetes and beyond, *Med Res Rev* **21,** 540–552 (2001).

144. Sugii, S. *et al.* PPARgamma activation in adipocytes is sufficient for systemic insulin sensitization, *Proc Natl Acad Sci U S A* **106,** 22504–22509 (2009).

145. Housley, W. J. *et al.* Peroxisome Proliferator-Activated Receptor Is Required for CD4+ T Cell-Mediated Lymphopenia-Associated Autoimmunity, *The Journal of Immunology* **187,** 4161–4169 (2011).

146. Clark, R. B. *et al.* The nuclear receptor PPAR gamma and immunoregulation: PPAR gamma mediates inhibition of helper T cell responses, *J Immunol* **164,** 1364–1371 (2000).

147. Garcia-Bates, T. M. *et al.* Peroxisome Proliferator-Activated Receptor Ligands Enhance Human B Cell Antibody Production and Differentiation, *The Journal of Immunology* **183,** 6903–6912 (2009).

148. Housley, W. J. *et al.* PPARgamma regulates retinoic acid-mediated DC induction of Tregs, *J Leukoc Biol* **86,** 293–301 (2009).

149. Faveeuw, C. *et al.* Peroxisome proliferator-activated receptor gamma activators inhibit interleukin-12 production in murine dendritic cells, *FEBS Lett* **486,** 261–266 (2000).

150. Pignatelli, M. Cocca, C. Santos, A. & Perez-Castillo, A. Enhancement of BRCA1 gene expression by the peroxisome proliferator-activated receptor gamma in the MCF-7 breast cancer cell line, *Oncogene* **22,** 5446–5450 (2003).

151. Memisoglu, A. Hankinson, S. E. Manson, J. E. Colditz, G. A. & Hunter, D. J. Lack of association of the codon 12 polymorphism of the peroxisome proliferator-activated receptor gamma gene with breast cancer and body mass, *Pharmacogenetics* **12,** 597–603 (2002).

152. Segawa, Y. *et al.* Expression of peroxisome proliferator-activated receptor (PPAR) in human prostate cancer, *Prostate* **51,** 108–116 (2002).

153. Sarraf, P. *et al.* Differentiation and reversal of malignant changes in colon cancer through PPARgamma, *Nat Med* **4,** 1046–1052 (1998).

154. Bishop-Bailey, D. Peroxisome proliferator-activated receptors in the cardiovascular system, *Br J Pharmacol* **129,** 823–834 (2000).

155. Jiang, Q. Heneka, M. & Landreth, G. E. The role of peroxisome proliferator-activated receptor-gamma (PPARgamma) in Alzheimer's disease: therapeutic implications, *CNS Drugs* **22,** 1–14 (2008).

156. Al-Shali, K. *et al.* A Single-Base Mutation in the Peroxisome Proliferator-Activated Receptor γ4 Promoter Associated with Altered in Vitro Expression and Partial Lipodystrophy, *The Journal of Clinical Endocrinology & Metabolism* **89,** 5655–5660 (2004).

157. Sabatino, L. Fucci, A. Pancione, M. & Colantuoni, V. PPARG Epigenetic Deregulation and Its Role in Colorectal Tumorigenesis, *PPAR Research* **2012,** 1–12 (2012).

158. Zhou, J. Wilson, K. M. & Medh, J. D. Genetic analysis of four novel peroxisome proliferator activated receptor-gamma splice variants in monkey macrophages, *Biochem Biophys Res Commun* **293,** 274–283 (2002).

159. Sundvold, H. & Lien, S. Identification of a novel peroxisome proliferator-activated receptor (PPAR) gamma promoter in man and transactivation by the nuclear receptor RORalpha1, *Biochem Biophys Res Commun* **287,** 383–390 (2001).

160. Strand, D. W. *et al.* PPARγ isoforms differentially regulate metabolic networks to mediate mouse prostatic epithelial differentiation, *Cell Death Dis* **3,** e361 (2012).

161. Saladin, R. *et al.* Differential regulation of peroxisome proliferator activated receptor gamma1 (PPARgamma1) and PPARgamma2 messenger RNA expression in the early stages of adipogenesis, *Cell Growth Differ* **10,** 43–48 (1999).

162. Chen, Y. Jimenez, A. R. & Medh, J. D. Identification and regulation of novel PPAR-gamma splice variants in human THP-1 macrophages, *Biochim Biophys Acta* **1759,** 32–43 (2006).

163. Deeb, S. S. *et al.* A Pro12Ala substitution in PPAR[gamma]2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity, *Nat Genet* **20,** 284–287 (1998).

164. Voight, B. F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis, *Nat Genet* **42,** 579–589 (2010).

165. Gouda, H. N. *et al.* The association between the peroxisome proliferator-activated receptor-gamma2 (PPARG2) Pro12Ala gene variant and type 2 diabetes mellitus: a HuGE review and meta-analysis, *Am J Epidemiol* **171,** 645–655 (2010).

166. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes, *Nat Genet* **40,** 638–645 (2008).

167. Gallicchio, L. *et al.* Genetic Polymorphisms of Peroxisome Proliferator-Activated Receptors and the Risk of Cardiovascular Morbidity and Mortality in a Community-Based Cohort in Washington County, Maryland, *PPAR Research* **2008,** 1–9 (2008).

168. Tonjes, A. & Stumvoll, M. The role of the Pro12Ala polymorphism in peroxisome proliferator-activated receptor gamma in diabetes risk, *Curr Opin Clin Nutr Metab Care* **10,** 410–414 (2007).

169. Adamo, K. B. *et al.* Influence of Pro12Ala peroxisome proliferator-activated receptor gamma2 polymorphism on glucose response to exercise training in type 2 diabetes, *Diabetologia* **48,** 1503–1509 (2005).

170. Altshuler, D. *et al.* The common PPAR[gamma] Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes, *Nat Genet* **26,** 76–80 (2000).

171. Yen, C. J. *et al.* Molecular scanning of the human peroxisome proliferator activated receptor gamma (hPPAR gamma) gene in diabetic Caucasians: identification of a Pro12Ala PPAR gamma 2 missense mutation, *Biochem Biophys Res Commun* **241,** 270–274 (1997).

172. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes, *Nature* **491,** 56–65 (2012).

173. Chistiakov, D. A. *et al.* The PPARgamma Pro12Ala variant is associated with insulin sensitivity in Russian normoglycaemic and type 2 diabetic subjects, *Diab Vasc Dis Res* **7,** 56–62 (2010).

174. Lyssenko, V. *et al.* Genetic prediction of future type 2 diabetes, *PLoS Med* **2,** e345 (2005).

175. Gouda, H. N. *et al.* The Association Between the Peroxisome Proliferator-Activated Receptor- 2 (PPARG2) Pro12Ala Gene Variant and Type 2 Diabetes Mellitus: A HuGE Review and Meta-Analysis, *American Journal of Epidemiology* **171,** 645–655 (2010).

176. Ludovico, O. *et al.* Heterogeneous effect of peroxisome proliferator-activated receptor gamma2 Ala12 variant on type 2 diabetes risk, *Obesity (Silver Spring)* **15,** 1076–1081 (2007).

177. Parikh, H. & Groop, L. Candidate genes for type 2 diabetes, *Rev Endocr Metab Disord* **5,** 151–176 (2004).

178. Lohmueller, K. E. Pearce, C. L. Pike, M. Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease, *Nat Genet* **33,** 177–182 (2003).

179. Ek, J. *et al.* Studies of the Pro12Ala polymorphism of the peroxisome proliferator-activated receptor-gamma2 (PPAR-gamma2) gene in relation to insulin sensitivity among glucose tolerant caucasians, *Diabetologia* **44,** 1170–1176 (2001).

180. Heikkinen, S. *et al.* The Pro12Ala PPARγ2 Variant Determines Metabolism at the Gene-Environment Interface, *Cell Metabolism* **9,** 88–98 (2009).

181. Ek, J. *et al.* Homozygosity of the Pro12Ala variant of the peroxisome proliferation-activated receptor-gamma2 (PPAR-gamma2): divergent modulating effects on body mass index in obese and lean Caucasian men, *Diabetologia* **42,** 892–895 (1999).

182. Kilpelainen, T. O. *et al.* SNPs in PPARG associate with type 2 diabetes and interact with physical activity, *Med Sci Sports Exerc* **40,** 25–33 (2008).

183. Adamo, K. B. *et al.* Influence of Pro12Ala peroxisome proliferator-activated receptor gamma2 polymorphism on glucose response to exercise training in type 2 diabetes, *Diabetologia* **48,** 1503–1509 (2005).

184. Weiss, E. P. *et al.* Endurance training-induced changes in the insulin response to oral glucose are associated with the peroxisome proliferator-activated receptor-gamma2 Pro12Ala genotype in men but not in women, *Metabolism* **54,** 97–102 (2005).

185. Kahara, T. *et al.* PPARgamma gene polymorphism is associated with exercise-mediated changes of insulin resistance in healthy men, *Metabolism* **52,** 209–212 (2003).

186. Peters, T. Ausmeier, K. Dildrop, R. & Ruther, U. The mouse Fused toes (Ft) mutation is the result of a 1.6-Mb deletion including the entire Iroquois B gene cluster, *Mamm Genome* **13,** 186–188 (2002).

187. Peters, T. Ausmeier, K. & Ruther, U. Cloning of Fatso (Fto), a novel gene deleted by the Fused toes (Ft) mouse mutation, *Mamm Genome* **10,** 983–986 (1999).

188. Stratigopoulos, G. *et al.* Regulation of Fto/Ftm gene expression in mice and humans, *AJP: Regulatory, Integrative and Comparative Physiology* **294,** R1185 (2008).

189. Tews, D. Fischer-Posovszky, P. & Wabitsch, M. Regulation of FTO and FTM Expression During Human Preadipocyte Differentiation, *Horm Metab Res* **43,** 17–21 (2011).

190. Lappalainen, T. *et al.* Gene Expression of *FTO* in Human Subcutaneous Adipose Tissue, Peripheral Blood Mononuclear Cells and Adipocyte Cell Line, *J Nutrigenet Nutrigenomics* **3,** 37–45 (2010).

191. Grunnet, L. G. *et al.* Regulation and Function of FTO mRNA Expression in Human Skeletal Muscle and Subcutaneous Adipose Tissue, *Diabetes* **58,** 2402–2408 (2009).

192. Fredriksson, R. *et al.* The obesity gene, FTO, is of ancient origin, up-regulated during food deprivation and expressed in neurons of feeding-related nuclei of the brain, *Endocrinology* **149,** 2062–2071 (2008).

193. Fischer, J. *et al.* Inactivation of the Fto gene protects from obesity, *Nature* **458,** 894–898 (2009).

194. Clifton, I. J. *et al.* Structural studies on 2-oxoglutarate oxygenases and related double-stranded beta-helix fold proteins, *J Inorg Biochem* **100,** 644–669 (2006).

195. Sanchez-Pulido, L. & Andrade-Navarro, M. A. The FTO (fat mass and obesity associated) gene codes for a novel member of the non-heme dioxygenase superfamily, *BMC Biochem* **8,** 23 (2007).

196. Jia, G. *et al.* N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO, *Nat Chem Biol* **7,** 885–887 (2011).

197. Tews, D. *et al.* FTO deficiency induces UCP-1 expression and mitochondrial uncoupling in adipocytes, *Endocrinology* **154,** 3141–3151 (2013).

198. McMurray, F. *et al.* Adult Onset Global Loss of the Fto Gene Alters Body Composition and Metabolism in the Mouse, *PLoS Genet* **9,** e1003166 (2013).

199. Albuquerque, D. Nóbrega, C. Manco, L. & López, M. Association of FTO Polymorphisms with Obesity and Obesity-Related Outcomes in Portuguese Children, *PLoS ONE* **8,** e54370 (2013).

200. Yang, J. *et al.* FTO genotype is associated with phenotypic variability of body mass index, *Nature* **490,** 267–272 (2012).

201. Peng, S. *et al.* FTO gene polymorphisms and obesity risk: a meta-analysis, *BMC Med* **9,** 71 (2011).

202. Sovio, U. *et al.* Association between Common Variation at the FTO Locus and Changes in Body Mass Index from Infancy to Late Childhood: The Complex Nature of Genetic Association through Growth and Development, *PLoS Genet* **7,** e1001307 (2011).

203. Al-Attar, S. A. *et al.* Association between the FTO rs9939609 polymorphism and the metabolic syndrome in a non-Caucasian multi-ethnic sample, *Cardiovasc Diabetol* **7,** 5 (2008).

204. Cha, S. W. *et al.* Replication of genetic effects of FTO polymorphisms on BMI in a Korean population, *Obesity (Silver Spring)* **16,** 2187–2189 (2008).

205. Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits, *PLoS Genet* **3,** e115 (2007).

206. Wing, M. R. *et al.* Analysis of FTO gene variants with obesity and glucose homeostasis measures in the multiethnic Insulin Resistance Atherosclerosis Study cohort, *Int J Obes (Lond)* **35,** 1173–1182 (2011).

207. Andreasen, C. H. *et al.* Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation, *Diabetes* **57,** 95–101 (2008).

208. Hertel, J. K. *et al.* FTO, type 2 diabetes, and weight gain throughout adult life: a meta-analysis of 41,504 subjects from the Scandinavian HUNT, MDC, and MPP studies, *Diabetes* **60,** 1637–1644 (2011).

209. Sonestedt, E. *et al.* Fat and carbohydrate intake modify the association between genetic variation in the FTO genotype and obesity, *Am J Clin Nutr* **90,** 1418–1425 (2009).

210. Chang, Y.-C. *et al.* Common variation in the fat mass and obesity-associated (FTO) gene confers risk of obesity and modulates BMI in the Chinese population, *Diabetes* **57,** 2245–2252 (2008).

211. Bressler, J. Kao, W H Linda, Pankow, J. S. & Boerwinkle, E. Risk of type 2 diabetes and obesity is differentially associated with variation in FTO in whites and African-Americans in the ARIC study, *PLoS ONE* **5,** e10521 (2010).

212. Wahlen, K. Sjolin, E. & Hoffstedt, J. The common rs9939609 gene variant of the fat mass- and obesity-associated gene FTO is related to fat cell lipolysis, *J Lipid Res* **49,** 607–611 (2008).

213. Kloting, N. *et al.* Inverse relationship between obesity and FTO gene expression in visceral adipose tissue in humans, *Diabetologia* **51,** 641–647 (2008).

214. Li, T. *et al.* Common variant rs9939609 in gene FTO confers risk to polycystic ovary syndrome, *PLoS ONE* **8,** e66250 (2013).

215. Barber, T. M. *et al.* Association of variants in the fat mass and obesity associated (FTO) gene with polycystic ovary syndrome, *Diabetologia* **51,** 1153–1158 (2008).

216. Keller, L. *et al.* The obesity related gene, FTO, interacts with APOE, and is associated with Alzheimer's disease risk: a prospective cohort study, *J Alzheimers Dis* **23,** 461–469 (2011).

217. Hubacek, J. A. *et al.* A FTO variant and risk of acute coronary syndrome, *Clin Chim Acta* **411,** 1069–1072 (2010).

218. Doney, Alex S F *et al.* The FTO gene is associated with an atherogenic lipid profile and myocardial infarction in patients with type 2 diabetes: a Genetics of Diabetes Audit and Research Study in Tayside Scotland (Go-DARTS) study, *Circ Cardiovasc Genet* **2,** 255–259 (2009).

219. Iles, M. M. *et al.* A variant in FTO shows association with melanoma risk not due to BMI, *Nat Genet* **45,** 428-32, 432e1 (2013).

220. Hubacek, J. A. *et al.* The FTO gene polymorphism is associated with end-stage renal disease: two large independent case-control studies in a general population, *Nephrol Dial Transplant* **27,** 1030–1035 (2012).

221. Olza, J. *et al.* Influence of FTO variants on obesity, inflammation and cardiovascular disease risk biomarkers in Spanish children: a case-control multicentre study, *BMC Med Genet* **14,** 123 (2013).

222. Freathy, R. M. *et al.* Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI, *Diabetes* **57,** 1419–1426 (2008).

223. Barber, T. M. *et al.* Association of variants in the fat mass and obesity associated (FTO) gene with polycystic ovary syndrome, *Diabetologia* **51,** 1153–1158 (2008).

224. Lyssenko, V. *et al.* Mechanisms by which common variants in the TCF7L2 gene increase risk of type 2 diabetes, *J Clin Invest* **117,** 2155–2163 (2007).

225. Jin, T. & Liu, L. The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus, *Mol Endocrinol* **22,** 2383–2392 (2008).

226. Yang, Y. Wnt signaling in development and disease, *Cell Biosci* **2,** 14 (2012).

227. MacDonald, B. T. Tamai, K. & He, X. Wnt/beta-catenin signaling: components, mechanisms, and diseases, *Dev Cell* **17,** 9–26 (2009).

228. Jin, T. The WNT signalling pathway and diabetes mellitus, *Diabetologia* **51,** 1771–1780 (2008).

229. Clevers, H. Wnt/beta-catenin signaling in development and disease, *Cell* **127,** 469–480 (2006).

230. Ip, W. Chiang, Y.-T. A. & Jin, T. The involvement of the wnt signaling pathway and TCF7L2 in diabetes mellitus: The current understanding, dispute, and perspective, *Cell Biosci* **2,** 28 (2012).

231. Yi, F. Brubaker, P. L. & Jin, T. TCF-4 mediates cell type-specific regulation of proglucagon gene expression by beta-catenin and glycogen synthase kinase-3beta, *J Biol Chem* **280,** 1457–1464 (2005).

232. Poy, F. Lepourcelet, M. Shivdasani, R. A. & Eck, M. J. Structure of a human Tcf4-beta-catenin complex, *Nat Struct Biol* **8,** 1053–1057 (2001).

233. Duval, A. *et al.* The human T-cell transcription factor-4 gene: structure, extensive characterization of alternative splicings, and mutational analysis in colorectal cancer cell lines, *Cancer Res* **60,** 3872–3879 (2000).

234. Osmark, P. *et al.* Unique splicing pattern of the TCF7L2 gene in human pancreatic islets, *Diabetologia* **52,** 850–854 (2009).

235. Cauchi, S. *et al.* Transcription factor TCF7L2 genetic study in the French population: expression in human beta-cells and adipose tissue and strong association with type 2 diabetes, *Diabetes* **55,** 2903–2908 (2006).

236. Lillioja, S. & Wilton, A. Agreement among type 2 diabetes linkage studies but a poor correlation with results from genome-wide association studies, *Diabetologia* **52,** 1061–1074 (2009).

237. Reynisdottir, I. *et al.* Localization of a susceptibility gene for type 2 diabetes to chromosome 5q34-q35.2, *Am J Hum Genet* **73,** 323–335 (2003).

238. Duggirala, R. *et al.* Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans, *Am J Hum Genet* **64,** 1127–1140 (1999).

239. Tong, Y. *et al.* Association between TCF7L2 gene polymorphisms and susceptibility to type 2 diabetes mellitus: a large Human Genome Epidemiology (HuGE) review and meta-analysis, *BMC Med Genet* **10,** 15 (2009).

240. Chandak, G. R. *et al.* Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population, *Diabetologia* **50,** 63–67 (2007).

241. Hayashi, T. Iwamoto, Y. Kaku, K. Hirose, H. & Maeda, S. Replication study for the association of TCF7L2 with susceptibility to type 2 diabetes in a Japanese population, *Diabetologia* **50,** 980–984 (2007).

242. Lehman, D. M. *et al.* Haplotypes of transcription factor 7-like 2 (TCF7L2) gene and its upstream region are associated with type 2 diabetes and age of onset in Mexican Americans, *Diabetes* **56,** 389–393 (2007).

243. Zeggini, E. & McCarthy, M. I. TCF7L2: the biggest story in diabetes genetics since HLA?, *Diabetologia* **50,** 1–4 (2007).

244. Humphries, S. E. *et al.* Common variants in the TCF7L2 gene and predisposition to type 2 diabetes in UK European Whites, Indian Asians and Afro-Caribbean men and women, *J Mol Med (Berl)* **84,** 1005–1014 (2006).

245. Florez, J. C. *et al.* TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program, *N Engl J Med* **355,** 241–250 (2006).

246. Groves, C. J. *et al.* Association analysis of 6,736 U.K. subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk, *Diabetes* **55,** 2640–2644 (2006).

247. Damcott, C. M. *et al.* Polymorphisms in the transcription factor 7-like 2 (TCF7L2) gene are associated with type 2 diabetes in the Amish: replication and evidence for a role in both insulin secretion and insulin resistance, *Diabetes* **55,** 2654–2659 (2006).

248. Scott, L. J. *et al.* Association of transcription factor 7-like 2 (TCF7L2) variants with type 2 diabetes in a Finnish sample, *Diabetes* **55,** 2649–2653 (2006).

249. Saxena, R. *et al.* Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals, *Diabetes* **55,** 2890–2895 (2006).

250. Zhang, C. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene and the risk of type 2 diabetes in large cohorts of U.S. women and men, *Diabetes* **55,** 2645–2648 (2006).

251. Pang, D. Smith, A. & Humphries, S. Functional analysis of TCF7L2 genetic variants associated with type 2 diabetes, *Nutrition, Metabolism and Cardiovascular Diseases* **23,** 550–556 (2013).

252. Wong, Newton Alexander Chiang Shuek & Pignatelli, M. Beta-catenin--a linchpin in colorectal carcinogenesis?, *Am J Pathol* **160,** 389–401 (2002).

253. Korinek, V. *et al.* Depletion of epithelial stem-cell compartments in the small intestine of mice lacking Tcf-4, *Nat Genet* **19,** 379–383 (1998).

254. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets, *Nat.Genet* **42,** 255–259 (2010).

255. Stitzel, M. L. *et al.* Global epigenomic analysis of primary human pancreatic islets provides insights into type 2 diabetes susceptibility loci, *Cell Metab* **12,** 443–455 (2010).

256. Savic, D. *et al.* Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism, *Genome Research* **21,** 1417–1425 (2011).

257. Zhang, K. *et al.* ICSNPathway: identify candidate causal SNPs and pathways from genome-wide association study by one analytical framework, *Nucleic Acids Research* **39,** W437 (2011).

258. Gerasimova, A. *et al.* Predicting cell types and genetic variations contributing to disease by combining GWAS and epigenetic data, *PLoS ONE* **8,** e54359 (2013).

259. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA, *Science* **337,** 1190–1195 (2012).

260. Nica, A. C. *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study, *PLoS Genet* **7,** e1002003 (2011).

261. Nicolae, D. L. *et al.* Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS, *PLoS Genet* **6,** e1000888 (2010).

262. Nica, A. C. *et al.* Candidate Causal Regulatory Effects by Integration of Expression QTLs with Complex Trait Genetic Associations, *PLoS Genet* **6,** e1000895 (2010).

263. Edwards, S. L. Beesley, J. French, J. D. & Dunning, A. M. Beyond GWASs: illuminating the dark road from association to function, *Am J Hum Genet* **93,** 779–797 (2013).

264. Kolovos, P. *et al.* Targeted Chromatin Capture (T2C): a novel high resolution high throughput method to detect genomic interactions and regulatory elements, *Epigenetics Chromatin* **7,** 10 (2014).

265. Mouchiroud, L. Eichner, L. J. Shaw, R. J. & Auwerx, J. Transcriptional Coregulators: Fine-Tuning Metabolism, *Cell Metab* **20,** 26–40 (2014).

266. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus, *Nature* **466,** 714–719 (2010).

267. Han, Y. J. Ma, S.-F. Wade, M. S. Flores, C. & Garcia, J. G. N. An intronic MYLK variant associated with inflammatory lung disease regulates promoter activity of the smooth muscle myosin light chain kinase isoform, *J Mol Med* **90,** 299–308 (2012).

268. Praetorius, C. *et al.* A polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway, *Cell* **155,** 1022–1033 (2013).

269. Spieler, D. *et al.* Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon, *Genome Res.* **24,** 592–603 (2014).

270. Maynard, N. D. Chen, J. Stuart, R. K. Fan, J.-B. & Ren, B. Genome-wide mapping of allele-specific protein-DNA interactions in human cells, *Nat Methods* **5,** 307–309 (2008).

271. Chu, X. *et al.* Dynamic conformational change regulates the protein-DNA recognition: an investigation on binding of a Y-family polymerase to its target DNA, *PLoS Comput Biol* **10,** e1003804 (2014).

272. Lin, W.-Z. Fang, J.-A. Xiao, X. & Chou, K.-C. iDNA-Prot: identification of DNA binding proteins using random forest with grey model, *PLoS ONE* **6,** e24756 (2011).

273. Langlois, R. E. & Lu, H. Boosting the prediction and understanding of DNA-binding domains from sequence, *Nucleic Acids Res* **38,** 3149–3158 (2010).

274. Qin, H. & Wang, Y. Exploring DNA-binding proteins with in vivo chemical cross-linking and mass spectrometry, *J Proteome Res* **8,** 1983–1991 (2009).

275. Helwa, R. & Hoheisel, J. D. Analysis of DNA-protein interactions: from nitrocellulose filter binding assays to microarray studies, *Anal Bioanal Chem* **398,** 2551–2561 (2010).

276. Furney, S. J. Higgins, D. G. Ouzounis, C. A. & Lopez-Bigas, N. Structural and functional properties of genes involved in human cancer, *BMC Genomics* **7,** 3 (2006).

277. Hellman, L. M. & Fried, M. G. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions, *Nat Protoc* **2,** 1849–1861 (2007).

278. Papoulas, O. Rapid separation of protein-bound DNA from free DNA using nitrocellulose filters, *Curr Protoc Mol Biol* **Chapter 12,** Unit 12.8 (2001).

279. Hall, K. B. & Kranz, J. K. Nitrocellulose filter binding for determination of dissociation constants, *Methods Mol Biol* **118,** 105–114 (1999).

280. Kuhl, A. J. Ross, S. M. & Gaido, K. W. Using a comparative in vivo DNase I footprinting technique to analyze changes in protein-DNA interactions following phthalate exposure, *J Biochem Mol Toxicol* **21,** 312–322 (2007).

281. Tullius, T. D. Dombroski, B. A. Churchill, M. E. & Kam, L. Hydroxyl radical footprinting: a high-resolution method for mapping protein-DNA contacts, *Methods Enzymol* **155,** 537–558 (1987).

282. Brenowitz, M. Senear, D. F. Shea, M. A. & Ackers, G. K. Quantitative DNase footprint titration: a method for studying protein-DNA interactions, *Methods Enzymol* **130,** 132–181 (1986).

283. Woodbury, C P Jr & von Hippel, P H. On the determination of deoxyribonucleic acid-protein interaction parameters using the nitrocellulose filter-binding assay, *Biochemistry* **22,** 4730–4737 (1983).

284. Whitson, P. A. & Matthews, K. S. Dissociation of the lactose repressor-operator DNA complex: effects of size and sequence context of operator-containing DNA, *Biochemistry* **25,** 3845–3852 (1986).

285. Sanger, F. Coulson, A. R. Hong, G. F. Hill, D. F. & Petersen, G. B. Nucleotide sequence of bacteriophage lambda DNA, *J Mol Biol* **162,** 729–773 (1982).

286. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity, *Nucleic Acids Res* **5,** 3157–3170 (1978).

287. Hampshire, A. J. Rusling, D. A. Broughton-Head, V. J. & Fox, K. R. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands, *Methods* **42,** 128–140 (2007).

288. Leblanc, B. & Moss, T. DNase I footprinting, *Methods Mol Biol* **148,** 31–38 (2001).

289. Geiselmann, J. & Boccard, F. Ultraviolet-laser footprinting, *Methods Mol Biol* **148,** 161–173 (2001).

290. Das, P. M. Ramachandran, K. vanWert, J. & Singal, R. Chromatin immunoprecipitation assay, *Biotechniques* **37,** 961–969 (2004).

291. Mundade, R. Ozer, H. G. Wei, H. Prabhu, L. & Lu, T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond, *Cell Cycle* **13,** 2847–2852 (2014).

292. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, *Genome Res* **22,** 1813–1831 (2012).

293. Furey, T. S. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions, *Nat Rev Genet* **13,** 840–852 (2012).

294. Reece-Hoyes, J. S. *et al.* Yeast one-hybrid assays for gene-centered human gene regulatory network mapping, *Nat Meth* **8,** 1050–1052 (2011).

295. Yanai, K. A modified yeast one-hybrid system for genome-wide identification of transcription factor binding sites, *Methods Mol Biol* **977,** 125–136 (2013).

296. Fuxman Bass, Juan I *et al.* Human gene-centered transcription factor networks for enhancers and disease variants, *Cell* **161,** 661–673 (2015).

297. Walhout, Albertha J M. What does biologically meaningful mean? A perspective on gene regulatory network validation, *Genome Biol* **12,** 109 (2011).

298. Reece-Hoyes, J. S. *et al.* Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping, *Nat Methods* **8,** 1059–1064 (2011).

299. Schulze, W. X. & Usadel, B. Quantitation in Mass-Spectrometry-Based Proteomics, *Annu. Rev. Plant Biol.* **61,** 491–516 (2010).

300. Mallick, P. & Kuster, B. Proteomics: a pragmatic perspective, *Nat Biotechnol* **28,** 695–709 (2010).

301. Cravatt, B. F. Simon, G. M. & Yates, John R 3rd. The biological impact of mass-spectrometry-based proteomics, *Nature* **450,** 991–1000 (2007).

302. Graves, P. R. & Haystead, Timothy A J. Molecular biologist's guide to proteomics, *Microbiol Mol Biol Rev* **66,** 39-63; table of contents (2002).

303. Nordhoff, E. & Lehrach, H. Identification and characterization of DNA-binding proteins by mass spectrometry, *Adv Biochem Eng Biotechnol* **104,** 111–195 (2007).

304. Soldi, M. Cuomo, A. Bremang, M. & Bonaldi, T. Mass spectrometry-based proteomics for the analysis of chromatin structure and dynamics, *Int J Mol Sci* **14,** 5402–5431 (2013).

305. Panesar, P. S. & Marwaha, S. S. *Biotechnology in Agriculture and Food Processing: Opportunities and Challenges* (Taylor & Francis, 2013).

306. Safarik, I. & Safarikova, M. Magnetic techniques for the isolation and purification of proteins and peptides, *Biomagn Res Technol* **2,** 7 (2004).

307. Gingras, A.-C. Gstaiger, M. Raught, B. & Aebersold, R. Analysis of protein complexes using mass spectrometry, *Nat Rev Mol Cell Biol* **8,** 645–654 (2007).

308. Kocher, T. & Superti-Furga, G. Mass spectrometry-based functional proteomics: from molecular machines to protein networks, *Nat Methods* **4,** 807–815 (2007).

309. Musso, G. A. Zhang, Z. & Emili, A. Experimental and computational procedures for the assessment of protein complexes on a genome-wide scale, *Chem Rev* **107,** 3585–3600 (2007).

310. Malovannaya, A. *et al.* Analysis of the human endogenous coregulator complexome, *Cell* **145,** 787–799 (2011).

311. Malovannaya, A. *et al.* Streamlined analysis schema for high-throughput identification of endogenous protein complexes, *Proc Natl Acad Sci U S A* **107,** 2431–2436 (2010).

312. Butter, F. *et al.* Proteome-Wide Analysis of Disease-Associated SNPs That Show Allele-Specific Transcription Factor Binding, *PLoS Genet* **8,** e1002982 (2012).

313. Jutras, B. L. Verma, A. & Stevenson, B. Identification of novel DNA-binding proteins using DNA-affinity chromatography/pull down, *Curr Protoc Microbiol* **Chapter 1,** Unit1F.1 (2012).

314. Himeda, C. L. *et al.* Quantitative Proteomic Identification of Six4 as the Trex-Binding Factor in the Muscle Creatine Kinase Enhancer, *Molecular and Cellular Biology* **24,** 2132–2143 (2004).

315. Nordhoff, E. *et al.* Rapid identification of DNA-binding proteins by mass spectrometry, *Nat Biotechnol* **17,** 884–888 (1999).

316. Gabrielsen, O. S. & Huet, J. Magnetic DNA affinity purification of yeast transcription factor, *Methods Enzymol* **218,** 508–525 (1993).

317. Kim, J. *et al.* Designed fabrication of multifunctional magnetic gold nanoshells and their application to magnetic resonance imaging and photothermal therapy, *Angew Chem Int Ed Engl* **45,** 7754–7758 (2006).

318. Kim, J.-S. *et al.* Cellular uptake of magnetic nanoparticle is mediated through energy-dependent endocytosis in A549 cells, *J Vet Sci* **7,** 321–326 (2006).

319. Choi, J.-s. *et al.* Biocompatible heterostructured nanoparticles for multimodal biological detection, *J Am Chem Soc* **128,** 15982–15983 (2006).

320. Jun, B.-H. *et al.* Protein separation and identification using magnetic beads encoded with surface-enhanced Raman spectroscopy, *Analytical Biochemistry* **391,** 24–30 (2009).

321. Levin, Y. Schwarz, E. Wang, L. Leweke, F. M. & Bahn, S. Label-free LC-MS/MS quantitative proteomics for large-scale biomarker discovery in complex samples, *J. Sep. Sci.* **30,** 2198–2203 (2007).

322. Wang, G. Wu, W. W. Zeng, W. Chou, C.-L. & Shen, R.-F. Label-Free Protein Quantification Using LC-Coupled Ion Trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application with Complex Proteomes, *J. Proteome Res.* **5,** 1214–1223 (2006).

323. Ewing, R. M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry, *Mol Syst Biol* **3,** 89 (2007).

324. Gavin, A.-C. *et al.* Proteome survey reveals modularity of the yeast cell machinery, *Nature* **440,** 631–636 (2006).

325. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature* **440,** 637–643 (2006).

326. Huang, P. H. & White, F. M. Phosphoproteomics: unraveling the signaling web, *Mol Cell* **31,** 777–781 (2008).

327. Kretschmer, A. *et al.* A common atopy-associated variant in the Th2 cytokine locus control region impacts transcriptional regulation and alters SMAD3 and SP1 binding, *Allergy* **69,** 632–642 (2014).

328. Fang, R. *et al.* Differential label-free quantitative proteomic analysis of Shewanella oneidensis cultured under aerobic and suboxic conditions by accurate mass and time tag approach, *Mol Cell Proteomics* **5,** 714–725 (2006).

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

330. Park, S.-S. *et al.* Effective correction of experimental errors in quantitative proteomics using stable isotope labeling by amino acids in cell culture (SILAC), *J Proteomics* **75,** 3720–3732 (2012).

331. Turck, C. W. *et al.* The Association of Biomolecular Resource Facilities Proteomics Research Group 2006 study: relative protein quantitation, *Mol Cell Proteomics* **6,** 1291–1298 (2007).

332. Ludwig, C. Claassen, M. Schmidt, A. & Aebersold, R. Estimation of absolute protein quantities of unlabeled samples by selected reaction monitoring mass spectrometry, *Mol Cell Proteomics* **11,** M111.013987 (2012).

333. Quadroni, M. Ducret, A. & Stocklin, R. Quantify this! Report on a round table discussion on quantitative mass spectrometry in proteomics, *Proteomics* **4,** 2211–2215 (2004).

334. Chelius, D. & Bondarenko, P. V. Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry, *J Proteome Res* **1,** 317–323 (2002).

335. Klaus, S. *et al.* Characterization of the novel brown adipocyte cell line HIB 1B. Adrenergic pathways involved in regulation of uncoupling protein gene expression, *J Cell Sci* **107 (Pt 1),** 313–319 (1994).

336. Rahman, A. *et al.* Proteomic analysis for inhibitory effect of chitosan oligosaccharides on 3T3-L1 adipocyte differentiation, *Proteomics* **8,** 569–581 (2008).

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

338. Skurk, T. & Hauner, H. Primary culture of human adipocyte precursor cells: expansion and differentiation, *Methods Mol Biol* **806,** 215–226 (2012).

339. Hauner, H. Skurk, T. & Wabitsch, M. Cultures of human adipose precursor cells, *Methods Mol.Biol* **155,** 239–247 (2001).

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

341. Hoffmann, C. *et al.* A Novel SP1/SP3 Dependent Intronic Enhancer Governing Transcription of the UCP3 Gene in Brown Adipocytes, *PLoS ONE* **8,** e83426 (2013).

342. Moxley, R. A. & Jarrett, H. W. Oligonucleotide trapping method for transcription factor purification systematic optimization using electrophoretic mobility shift assay (2005).

343. Wisniewski, J. R. Zougman, A. Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis, *Nat Meth* **6,** 359–362 (2009).

344. Hauck, S. M. *et al.* Deciphering membrane-associated molecular processes in target tissue of autoimmune uveitis by label-free quantitative mass spectrometry, *Molecular & Cellular Proteomics* **9,** 2292–2305 (2010).

345. Toerne, C. von *et al.* Apoe, Mbl2, and Psp Plasma Protein Levels Correlate with Diabetic Phenotype in NZO Mice—An Optimized Rapid Workflow for SRM-Based Quantification, *J. Proteome Res.* **12,** 1331–1343 (2013).

346. Arner, E. *et al.* Adipose Tissue MicroRNAs as Regulators of CCL2 Production in Human Obesity, *Diabetes* **61,** 1986–1993 (2012).

347. Bonora, E. *et al.* 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 (2000).

348. Holzapfel, C. *et al.* 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 (2008).

349. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes, *Nat Genet* **44,** 981–990 (2012).

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

351. Stranger, B. E. *et al.* Patterns of cis regulatory variation in diverse human populations, *PLoS Genet* **8,** e1002639 (2012).

352. Whitfield, T. W. *et al.* Functional analysis of transcription factor binding sites in human promoters, *Genome Biol* **13,** R50 (2012).

353. Mikkelsen, T. S. *et al.* Comparative Epigenomic Analysis of Murine and Human Adipogenesis, *Cell* **143,** 156–169 (2010).

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

355. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease, *Nat Biotechnol* **30,** 1095–1106 (2012).

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

357. Almlof, J. C. *et al.* Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression, *PLoS ONE* **7,** e52260 (2012).

358. Zeller, T. *et al.* Genetics and beyond--the transcriptome of human monocytes and disease susceptibility, *PLoS ONE* **5,** e10693 (2010).

359. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association, *Nature* **437,** 1365–1369 (2005).

360. Stranger, B. E. *et al.* Genome-wide associations of gene expression variation in humans, *PLoS Genet* **1,** e78 (2005).

361. Maurano, M. T. Wang, H. Kutyavin, T. & Stamatoyannopoulos, J. A. Widespread Site-Dependent Buffering of Human Regulatory Polymorphism, *PLoS Genet* **8,** e1002599 EP - (2012).

362. Cheung, V. G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association, *Nature* **437,** 1365–1369 (2005).

363. Gonzalez, M. W. & Kann, M. G. Chapter 4: Protein interactions and disease, *PLoS Comput Biol* **8,** e1002819 (2012).

364. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk, *Nat Genet* **42,** 105–116 (2010).

365. Zhang, J. Selective disruption of PPAR 2 impairs the development of adipose tissue and insulin sensitivity, *PNAS* **101,** 10703–10708 (2004).

366. Rosen, E. D. *et al.* PPARγ Is Required for the Differentiation of Adipose Tissue In Vivo and In Vitro, *Molecular Cell* **4,** 611–617 (1999).

367. Evangelisti, L. *et al.* PPARgamma promoter polymorphisms and acute coronary syndrome, *Atherosclerosis* **205,** 186–191 (2009).

368. Doney, Alex S F, Fischer, B. Leese, G. Morris, A. D. & Palmer, Colin N A. Cardiovascular risk in type 2 diabetes is associated with variation at the PPARG locus: a Go-DARTS study, *Arterioscler Thromb Vasc Biol* **24,** 2403–2407 (2004).

369. Haseeb, A. *et al.* Single-nucleotide polymorphisms in peroxisome proliferator-activated receptor gamma and their association with plasma levels of resistin and the metabolic syndrome in a South Indian population, *J Biosci* **34,** 405–414 (2009).

370. Dongxia, L. Qi, H. Lisong, L. & Jincheng, G. Association of peroxisome proliferator-activated receptorgamma gene Pro12Ala and C161T polymorphisms with metabolic syndrome, *Circ J* **72,** 551–557 (2008).

371. Mattevi, V. S. Zembrzuski, V. M. & Hutz, M. H. Effects of a PPARG gene variant on obesity characteristics in Brazil, *Braz J Med Biol Res* **40,** 927–932 (2007).

372. Hakanen, M. *et al.* FTO genotype is associated with body mass index after the age of seven years but not with energy intake or leisure-time physical activity, *J Clin Endocrinol Metab* **94,** 1281–1287 (2009).

373. Cornes, B. K. *et al.* Replication of the association of common rs9939609 variant of FTO with increased BMI in an Australian adult twin population but no evidence for gene by environment (G x E) interaction, *Int J Obes (Lond)* **33,** 75–79 (2009).

374. Lyssenko, V. The transcription factor 7-like 2 gene and increased risk of type 2 diabetes: an update, *Curr Opin Clin Nutr Metab Care* **11,** 385–392 (2008).

375. Hansson, O. Zhou, Y. Renstrom, E. & Osmark, P. Molecular function of TCF7L2: Consequences of TCF7L2 splicing for molecular function and risk for type 2 diabetes, *Curr Diab Rep* **10,** 444–451 (2010).

376. Peters, U. *et al.* A systematic mapping approach of 16q12.2/FTO and BMI in more than 20,000 African Americans narrows in on the underlying functional variation: results from the Population Architecture using Genomics and Epidemiology (PAGE) study, *PLoS Genet* **9,** e1003171 (2013).

377. Cauchi, S. *et al.* TCF7L2 is reproducibly associated with type 2 diabetes in various ethnic groups: a global meta-analysis, *J Mol Med (Berl)* **85,** 777–782 (2007).

378. Florez, J. C. *et al.* TCF7L2 polymorphisms and progression to diabetes in the Diabetes Prevention Program, *N Engl J Med* **355,** 241–250 (2006).

379. Helgason, A. *et al.* Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution, *Nat Genet* **39,** 218–225 (2007).

380. Savic, D. Park, S. Y. Bailey, K. A. Bell, G. I. & Nobrega, M. A. In vitro scan for enhancers at the TCF7L2 locus, *Diabetologia* **56,** 121–125 (2013).

381. Larouche, K. Bergeron, M. J. Leclerc, S. & Guerin, S. L. Optimization of competitor poly(dI-dC).poly(dI-dC) levels is advised in DNA-protein interaction studies involving enriched nuclear proteins, *Biotechniques* **20,** 439–444 (1996).

382. Neph, S. *et al.* Circuitry and dynamics of human transcription factor regulatory networks, *Cell* **150,** 1274–1286 (2012).

383. Figeys, D. McBroom, L. D. & Moran, M. F. Mass spectrometry for the study of protein-protein interactions, *Methods* **24,** 230–239 (2001).

384. Jaquinod, S. K. Chapel, A. Garin, J. & Journet, A. Affinity purification of soluble lysosomal proteins for mass spectrometric identification, *Methods Mol Biol* **432,** 243–258 (2008).

385. Gadgil, H. & Jarrett, H. W. Oligonucleotide trapping method for purification of transcription factors, *J Chromatogr A* **966,** 99–110 (2002).

386. Yang, V. W. Eukaryotic transcription factors: identification, characterization and functions, *J Nutr* **128,** 2045–2051 (1998).

387. Zhu, W. Smith, J. W. & Huang, C.-M. Mass spectrometry-based label-free quantitative proteomics, *J Biomed Biotechnol* **2010,** 840518 (2010).

388. Ożyhar, A. Gries, M. Kiltz, H.-H. & Pongs, O. Magnetic DNA affinity purification of ecdysteroid receptor, *The Journal of Steroid Biochemistry and Molecular Biology* **43,** 629–634 (1992).

389. Zhang, N. & Li, L. Effects of common surfactants on protein digestion and matrix-assisted laser desorption/ionization mass spectrometric analysis of the digested peptides using two-layer sample preparation, *Rapid Commun Mass Spectrom* **18,** 889–896 (2004).

390. Becktel, W. J. & Schellman, J. A. Protein stability curves, *Biopolymers* **26,** 1859–1877 (1987).

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

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

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

394. Li, F. Vijayasankaran, N. Shen, A. Y. Kiss, R. & Amanullah, A. Cell culture processes for monoclonal antibody production, *MAbs* **2,** 466–479 (2010).

395. Ohno, H. Shinoda, K. Spiegelman, B. M. & Kajimura, S. PPARgamma agonists induce a white-to-brown fat conversion through stabilization of PRDM16 protein, *Cell Metab* **15,** 395–404 (2012).

396. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index, *Nat Genet* **42,** 937–948 (2010).

397. Frayling, T. M. Genome-wide association studies provide new insights into type 2 diabetes aetiology, *Nat Rev Genet* **8,** 657–662 (2007).

398. Moriguchi, E. H. Tamachi, H. Homma, Y. & Goto, Y. An improved method for removal of antithrombin III from post-heparin plasma lipase fractions with a high recovery rate, *Tokai J Exp Clin Med* **16,** 33–41 (1991).

399. Farooqui, A. A. Purification of enzymes by heparin-sepharose affinity chromatography, *J Chromatogr* **184,** 335–345 (1980).

400. Srivastava, P. N. & Farooqui, A. A. Heparin-sepharose affinity chromatography for purification of bull seminal-plasma hyaluronidase, *Biochem J* **183,** 531–537 (1979).

401. Josic, D. Bal, F. & Schwinn, H. Isolation of plasma proteins from the clotting cascade by heparin affinity chromatography, *J Chromatogr* **632,** 1–10 (1993).

402. Wilson, K. & Walker, J. *Principles and Techniques of Biochemistry and Molecular Biology* (Cambridge University Press, 2010).

403. Jiang, D. Jia, Y. Zhou, Y. & Jarrett, H. W. Two-dimensional southwestern blotting and characterization of transcription factors on-blot, *J Proteome Res* **8,** 3693–3701 (2009).

404. Hodge, K. Have, S. T. Hutton, L. & Lamond, A. I. Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS, *J Proteomics* **88,** 92–103 (2013).

405. Wang, J. *et al.* Meta-analysis of associations between TCF7L2 polymorphisms and risk of type 2 diabetes mellitus in the Chinese population, *BMC Med Genet* **14,** 8 (2013).

406. McCarthy, M. I. Genomics, Type 2 Diabetes, and Obesity: New England Journal of Medicine, *N Engl J Med* **363,** 2339–2350 (2010).

407. Liu, P.-H. *et al.* Genetic variants of TCF7L2 are associated with insulin resistance and related metabolic phenotypes in Taiwanese adolescents and Caucasian young adults, *J Clin Endocrinol Metab* **94,** 3575–3582 (2009).

408. Prokopenko, I. McCarthy, M. I. & Lindgren, C. M. Type 2 diabetes: new genes, new understanding, *Trends Genet* **24,** 613–621 (2008).

409. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes, *Nature* **445,** 881–885 (2007).

410. Bailey, K. A. *et al.* Evidence of non-pancreatic beta cell-dependent roles of Tcf7l2 in the regulation of glucose metabolism in mice, *Hum Mol Genet* (2014).

411. Liu, B. *et al.* Common variants of transcription factor 7-like 2 (TCF7L2) are associated with reduced insulin secretion in women with polycystic ovary syndrome, *Gynecol Endocrinol* **28,** 594–597 (2012).

412. Skelin, M. Rupnik, M. & Cencic, A. Pancreatic beta cell lines and their applications in diabetes mellitus research, *ALTEX* **27,** 105–113 (2010).

413. Ravassard, P. *et al.* A genetically engineered human pancreatic beta cell line exhibiting glucose-inducible insulin secretion, *J Clin Invest* **121,** 3589–3597 (2011).

414. Ritchie, Graham R S, Dunham, I. Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants, *Nat Methods* **11,** 294–296 (2014).

415. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data, *Nucleic Acids Res* **41,** 827–841 (2013).

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

417. Meirhaeghe, A. & Amouyel, P. Impact of genetic variation of PPARγ in humans, *Molecular Genetics and Metabolism* **83,** 93–102 (2004).

418. Mittler, G. Butter, F. & Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements, *Genome Research* **19,** 284–293 (2008).

419. Hoglund, A. & Kohlbacher, O. From sequence to structure and back again: approaches for predicting protein-DNA binding, *Proteome Sci* **2,** 3 (2004).

420. García, E. Marcos-Gutiérrez, C. del Mar Lorente, M, Moreno, J. C. & Vidal, M. RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1, *EMBO J.* **18,** 3404–3418 (1999).

421. Fogarty, M. P. Cannon, M. E. Vadlamudi, S. Gaulton, K. J. & Mohlke, K. L. Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus, *PLoS Genet* **10,** e1004633 (2014).

422. Song, L. *et al.* Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity, *Genome Res* **21,** 1757–1767 (2011).

423. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies, *PLoS Genet* **10,** e1004722 (2014).

424. Chen, C.-Y. Chang, I.-S. Hsiung, C. A. & Wasserman, W. W. On the identification of potential regulatory variants within genome wide association candidate SNP sets, *BMC Med Genomics* **7,** 34 (2014).

425. Kowalczyk, M. S. *et al.* Intragenic enhancers act as alternative promoters, *Mol Cell* **45,** 447–458 (2012).

426. Shendure, J. Life after genetics, *Genome Medicine* **6,** 86 (2014).

427. Moyerbrailean, G. A. *et al.* Are all genetic variants in DNase I sensitivity regions functional?, *bioRxiv* (2014).

428. Frayling, T. M. & McCarthy, M. I. Genetic studies of diabetes following the advent of the genome-wide association study: where do we go from here?, *Diabetologia* **50,** 2229–2233 (2007).

429. Murea, M. Ma, L. & Freedman, B. I. Genetic and environmental factors associated with type 2 diabetes and diabetic vascular complications, *The Review of Diabetic Studies : RDS* **9,** 6–22 (2012).

430. Cookson, W. Liang, L. Abecasis, G. Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression, *Nat.Rev.Genet* **10,** 184–194 (2009).

431. Gibson, G. Powell, J. E. & Marigorta, U. M. Expression quantitative trait locus analysis for translational medicine, *Genome Med* **7,** 60 (2015).

432. McKenzie, M. Henders, A. K. Caracella, A. Wray, N. R. & Powell, J. E. Overlap of expression quantitative trait loci (eQTL) in human brain and blood, *BMC Med Genomics* **7,** 31 (2014).

433. Li, S. Lu, Q. & Cui, Y. A systems biology approach for identifying novel pathway regulators in eQTL mapping, *J Biopharm Stat* **20,** 373–400 (2010).

434. Li, G. *et al.* Identification of allele-specific alternative mRNA processing via transcriptome sequencing, *Nucleic Acids Res* **40,** e104 (2012).

435. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements, *Genome Research* **16,** 1299–1309 (2006).

436. Fullwood, M. J. *et al.* An oestrogen-receptor-alpha-bound human chromatin interactome, *Nature* **462,** 58–64 (2009).

437. Masugi, J. Tamori, Y. Mori, H. Koike, T. & Kasuga, M. Inhibitory effect of a proline-to-alanine substitution at codon 12 of peroxisome proliferator-activated receptor-gamma 2 on thiazolidinedione-induced adipogenesis, *Biochem Biophys Res Commun* **268,** 178–182 (2000).

438. Mori, H. *et al.* The Pro12 --Ala substitution in PPAR-gamma is associated with resistance to development of diabetes in the general population: possible involvement in impairment of insulin secretion in individuals with type 2 diabetes, *Diabetes* **50,** 891–894 (2001).

439. Masud, S. & Ye, S. Effect of the peroxisome proliferator activated receptor-gamma gene Pro12Ala variant on body mass index: a meta-analysis, *J Med Genet* **40,** 773–780 (2003).

440. Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits, *Genome Research* **24,** 1–13 (2014).

441. Gallicchio, L. *et al.* Single nucleotide polymorphisms in obesity-related genes and all-cause and cause-specific mortality: a prospective cohort study, *BMC Med Genet* **10,** 103 (2009).

442. Gallicchio, L. *et al.* Single Nucleotide Polymorphisms in Inflammation-related Genes and Mortality in a Community-based Cohort in Washington County, Maryland, *American Journal of Epidemiology* **167,** 807–813 (2008).

443. Tamori, Y. Masugi, J. Nishino, N. & Kasuga, M. Role of peroxisome proliferator-activated receptor-gamma in maintenance of the characteristics of mature 3T3-L1 adipocytes, *Diabetes* **51,** 2045–2055 (2002).

444. Auboeuf, D. *et al.* Tissue distribution and quantification of the expression of mRNAs of peroxisome proliferator-activated receptors and liver X receptor-alpha in humans: no alteration in adipose tissue of obese and NIDDM patients, *Diabetes* **46,** 1319–1327 (1997).

445. Kim, Y. C. Gomez, F. E. Fox, B. G. & Ntambi, J. M. Differential regulation of the stearoyl-CoA desaturase genes by thiazolidinediones in 3T3-L1 adipocytes, *J Lipid Res* **41,** 1310–1316 (2000).

446. Allott, E. H. *et al.* The SGBS cell strain as a model for the in vitro study of obesity and cancer, *Clin Transl Oncol* **14,** 774–782 (2012).

447. Rosen, E. D. & Spiegelman, B. M. Adipocytes as regulators of energy balance and glucose homeostasis, *Nature* **444,** 847–853 (2006).

448. Cannon, B. & Nedergaard, J. Brown adipose tissue: function and physiological significance, *Physiol Rev* **84,** 277–359 (2004).

449. Lin, J. *et al.* Defects in adaptive energy metabolism with CNS-linked hyperactivity in PGC-1alpha null mice, *Cell* **119,** 121–135 (2004).

450. Uldry, M. *et al.* Complementary action of the PGC-1 coactivators in mitochondrial biogenesis and brown fat differentiation, *Cell Metab* **3,** 333–341 (2006).

451. Kirkpatrick, C. L. *et al.* Type 2 diabetes susceptibility gene expression in normal or diabetic sorted human alpha and beta cells: correlations with age or BMI of islet donors, *PLoS ONE* **5,** e11053 (2010).

452. Qi, L. *et al.* Fat mass-and obesity-associated (FTO) gene variant is associated with obesity: longitudinal analyses in two cohort studies and functional test, *Diabetes* **57,** 3145–3151 (2008).

453. Pitman, R. T. Fong, J. T. Billman, P. & Puri, N. Knockdown of the fat mass and obesity gene disrupts cellular energy balance in a cell-type specific manner, *PLoS ONE* **7,** e38444 (2012).

454. Bravard, A. *et al.* FTO contributes to hepatic metabolism regulation through regulation of leptin action and STAT3 signalling in liver, *Cell Commun Signal* **12,** 4 (2014).

455. Russell, M. A. & Morgan, N. G. Conditional expression of the FTO gene product in rat INS-1 cells reveals its rapid turnover and a role in the profile of glucose-induced insulin secretion, *Clin Sci (Lond)* **120,** 403–413 (2011).

456. Chen, H. *et al.* An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape, *Sci Rep* **5,** 8465 (2015).

457. Jeziorska, D. M. Jordan, K. W. & Vance, K. W. A systems biology approach to understanding cis-regulatory module function, *Semin Cell Dev Biol* **20,** 856–862 (2009).

458. Arvey, A. Agius, P. Noble, W. S. & Leslie, C. Sequence and chromatin determinants of cell-type-specific transcription factor binding, *Genome Res* **22,** 1723–1734 (2012).

459. Shao, W. *et al.* The expression of dominant negative TCF7L2 in pancreatic beta cells during the embryonic stage causes impaired glucose homeostasis, *Mol Metab* **4,** 344–352 (2015).

460. Magdeldin, S. and Moser, A. (ed.). *Affinity Chromatography: Principles and Applications* (InTech, 2012).

461. Boulon, S. *et al.* Establishment of a Protein Frequency Library and Its Application in the Reliable Identification of Specific Protein Interaction Partners, *Molecular & Cellular Proteomics : MCP* **9,** 861–879 (2009).

462. Harper, S. & Speicher, D. W. Purification of proteins fused to glutathione S-tranferase, *Methods in molecular biology (Clifton, N.J.)* **681,** 259–280 (2011).

463. Korpela, T. & Kurkijarvi, K. A device for CNBr activation of agarose gels, *Anal Biochem* **104,** 150–152 (1980).

464. Stulik, J. Toman, R. Butaye, P. & Ulrich, R. G. *BSL3 and BSL4 Agents: Proteomics, Glycomics and Antigenicity* (Wiley-VCH Verlag, 2011).

465. Kim, H.-J. *et al.* Magnetic bead-based phage anti-immunocomplex assay (PHAIA) for the detection of the urinary biomarker 3-phenoxybenzoic acid to assess human exposure to pyrethroid insecticides, *Analytical Biochemistry* **386,** 45–52 (2008).

466. Wu, C. C. MacCoss, M. J. Howell, K. E. Matthews, D. E. & Yates, John R 3rd. Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis, *Anal Chem* **76,** 4951–4959 (2004).

467. Ishihama, Y. *et al.* Quantitative mouse brain proteomics using culture-derived isotope tags as internal standards, *Nat Biotechnol* **23,** 617–621 (2005).

468. Asara, J. M. Christofk, H. R. Freimark, L. M. & Cantley, L. C. A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen, *Proteomics* **8,** 994–999 (2008).

469. Hubner, N. C. & Mann, M. Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC), *Methods* **53,** 453–459 (2011).

470. Geiger, T. Cox, J. & Mann, M. Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation, *Mol Cell Proteomics* **9,** 2252–2261 (2010).

471. Nahnsen, S. Bielow, C. Reinert, K. & Kohlbacher, O. Tools for Label-free Peptide Quantification, *Molecular & Cellular Proteomics : MCP* **12,** 549–556 (2012).

472. Norris, R. A. The Identification of Prx1 Transcription Regulatory Domains Provides a Mechanism for Unequal Compensation by the Prx1 and Prx2 Loci, *Journal of Biological Chemistry* **276,** 26829–26837 (2001).

473. Hu, Y. S. *et al.* Expression of rat homeobox gene, rHOX, in developing and adult tissues in mice and regulation of its mRNA expression in osteoblasts by bone morphogenetic protein 2 and parathyroid hormone-related protein, *Mol Endocrinol* **12,** 1721–1732 (1998).

474. Nohno, T. *et al.* A chicken homeobox gene related to Drosophila paired is predominantly expressed in the developing limb, *Dev Biol* **158,** 254–264 (1993).

475. Cserjesi, P. *et al.* MHox: a mesodermally restricted homeodomain protein that binds an essential site in the muscle creatine kinase enhancer, *Development* **115,** 1087–1101 (1992).

476. Kern, M. J. Witte, D. P. Valerius, M. T. Aronow, B. J. & Potter, S. S. A novel murine homeobox gene isolated by a tissue specific PCR cloning strategy, *Nucleic Acids Res* **20,** 5189–5195 (1992).

477. Sugiyama, M. *et al.* Paired related homeobox 1 is associated with the invasive properties of glioblastoma cells, *Oncol Rep* **33,** 1123–1130 (2015).

478. Du, B. *et al.* The Transcription Factor Paired-Related Homeobox 1 (Prrx1) Inhibits Adipogenesis by Activating Transforming Growth Factor-β (TGFβ) Signaling, *J.Biol.Chem* **288,** 3036–3047 (2013).

479. Gulati, P. & Yeo, Giles S H. The biology of FTO: from nucleic acid demethylase to amino acid sensor, *Diabetologia* **56,** 2113–2121 (2013).

480. Liu, Z. & Habener, J. F. Glucagon-like peptide-1 activation of TCF7L2-dependent Wnt signaling enhances pancreatic beta cell proliferation, *J Biol Chem* **283,** 8723–8735 (2008).

481. Shu, L. *et al.* Transcription factor 7-like 2 regulates beta-cell survival and function in human pancreatic islets, *Diabetes* **57,** 645–653 (2008).

482. Elbein, S. C. *et al.* Transcription factor 7-like 2 polymorphisms and type 2 diabetes, glucose homeostasis traits and gene expression in US participants of European and African descent, *Diabetologia* **50,** 1621–1630 (2007).

483. Jolma, A. *et al.* DNA-binding specificities of human transcription factors, *Cell* **152,** 327–339 (2013).

484. Vaquerizas, J. M. Kummerfeld, S. K. Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution, *Nat Rev Genet* **10,** 252–263 (2009).

485. Auwerx, J. PPARgamma, the ultimate thrifty gene, *Diabetologia* **42,** 1033–1049 (1999).

486. Blättler, S. M. *et al.* Yin Yang 1 Deficiency in Skeletal Muscle Protects against Rapamycin-Induced Diabetic-like Symptoms through Activation of Insulin/IGF Signaling, *Cell Metabolism* **15,** 505–517 (2012).

487. Her, G. M. Pai, W.-Y. Lai, C.-Y. Hsieh, Y.-W. & Pang, H.-W. Ubiquitous transcription factor YY1 promotes zebrafish liver steatosis and lipotoxicity by inhibiting CHOP-10 expression, *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1831,** 1037–1051 (2013).

488. Lu, Y. *et al.* Yin Yang 1 Promotes Hepatic Gluconeogenesis Through Upregulation of Glucocorticoid Receptor, *Diabetes* **62,** 1064–1073 (2013).

489. He, C.-Q. Ding, N.-Z. & Fan, W. YY1 repressing peroxisome proliferator-activated receptor delta promoter, *Mol Cell Biochem* **308,** 247–252 (2008).

490. Yang, T. T. C. Xiong, Q. Enslen, H. Davis, R. J. & Chow, C.-W. Phosphorylation of NFATc4 by p38 Mitogen-Activated Protein Kinases, *Mol.Cell.Biol* **22,** 3892–3904 (2002).

491. Chen, L. *et al.* Significance of rs1271572 in the estrogen receptor beta gene promoter and its correlation with breast cancer in a southwestern Chinese population, *J Biomed Sci* **20,** 32 (2013).

492. Zhou, L. *et al.* A novel target of microRNA-29, Ring1 and YY1-binding protein (Rybp), negatively regulates skeletal myogenesis, *J. Biol. Chem.* **287,** 25255–25265 (2012).

493. Danen-van Oorschot, A A A M *et al.* Human death effector domain-associated factor interacts with the viral apoptosis agonist Apoptin and exerts tumor-preferential cell killing, *Cell Death Differ* **11,** 564–573 (2004).

494. Zheng, L. Schickling, O. Peter, M. E. & Lenardo, M. J. The death effector domain-associated factor plays distinct regulatory roles in the nucleus and cytoplasm, *J Biol Chem* **276,** 31945–31952 (2001).

495. Pirity, M. K. Locker, J. & Schreiber-Agus, N. Rybp/DEDAF is required for early postimplantation and for central nervous system development, *Mol Cell Biol* **25,** 7193–7202 (2005).

496. Chen, B. Fan, W. Liu, J. & Wu, F.-X. Identifying protein complexes and functional modules--from static PPI networks to dynamic PPI networks, *Brief Bioinform* **15,** 177–194 (2014).

497. Wang, X. *et al.* Modularity analysis based on predicted protein-protein interactions provides new insights into pathogenicity and cellular process of Escherichia coli O157:H7, *Theor Biol Med Model* **8,** 47 (2011).

498. Deng, Z. Chuaqui, C. & Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions, *J Med Chem* **47,** 337–344 (2004).

499. Spirin, V. & Mirny, L. A. Protein complexes and functional modules in molecular networks, *Proc Natl Acad Sci U S A* **100,** 12123–12128 (2003).

500. Hartwell, L. H. Hopfield, J. J. Leibler, S. & Murray, A. W. From molecular to modular cell biology, *Nature* **402,** C47-52 (1999).

501. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists, *Cell* **92,** 291–294 (1998).

502. Sharan, R. Ulitsky, I. & Shamir, R. Network-based prediction of protein function, *Mol Syst Biol* **3,** 88 (2007).

503. Hyde-DeRuyscher, R. P. Jennings, E. & Shenk, T. DNA binding sites for the transcriptional activator/repressor YY1, *Nucleic Acids Res* **23,** 4457–4465 (1995).

504. Medina-Gomez, G. *et al.* The Link Between Nutritional Status and Insulin Sensitivity Is Dependent on the Adipocyte-Specific Peroxisome Proliferator–Activated Receptor-γ2 Isoform, *Diabetes* **54,** 1706–1716 (2005).

505. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases, *Nat Genet* **44,** 1294–1301 (2012).

506. Yang, C. H. Axelrod, J. D. & Simon, M. A. Regulation of Frizzled by fat-like cadherins during planar polarity signaling in the Drosophila compound eye, *Cell* **108,** 675–688 (2002).

507. Sweetser, M. T. The Roles of Nuclear Factor of Activated T Cells and Ying-Yang 1 in Activation-induced Expression of the Interferon-gamma Promoter in T Cells, *J.Biol.Chem* **273,** 34775–34783 (1998).

508. Simicevic, J. *et al.* Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics, *Nat Meth* **10,** 570–576 (2013).

509. Bryois, J. *et al.* Cis and Trans Effects of Human Genomic Variants on Gene Expression, *PLoS Genetics* **10,** e1004461 (2014).

510. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins, *Nat Genet* **44,** 1084–1089 (2012).

511. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner, *Science* **325,** 1246–1250 (2009).

512. Stranger, B. E. *et al.* Population genomics of human gene expression, *Nat.Genet* **39,** 1217–1224 (2007).

513. Sanyal, A. The long-range interaction landscape of gene promoters.

514. Vadnais, C. *et al.* Long-range transcriptional regulation by the p110 CUX1 homeodomain protein on the ENCODE array, *BMC Genomics* **14,** 258 (2013).

515. Montgomery, S. B. & Dermitzakis, E. T. From expression QTLs to personalized transcriptomics, *Nat Rev Genet* **12,** 277–282 (2011).

516. Knight, J. C. Approaches for establishing the function of regulatory genetic variants involved in disease, *Genome Med* **6,** 92 (2014).

517. Small, K. S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes, *Nat Genet* **43,** 561–564 (2011).

518. Heinig, M. *et al.* A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk, *Nature* **467,** 460–464 (2010).

519. Cheung, V. G. *et al.* Polymorphic cis- and trans-regulation of human gene expression, *PLoS Biol* **8** (2010).

520. Sanghera, D. K. & Blackett, P. R. Type 2 Diabetes Genetics: Beyond GWAS, *J Diabetes Metab* **3** (2012).

11. Acknowledgements

Work on this thesis would not have been possible without encouragement and support from many people. First of I would like to extend my heartfelt gratitude to both Professor Hans Hauner and Professor Martin Hrabě de Angelis for giving me the opportunity to work. Without their support and encouragement throughout my entire PhD research, I would not have been able to finalize this thesis. In addition, this study was completed with support from the German Center for Diabetes Research and Clinical Cooperation Group Nutrigenomics and Type 2 Diabetes between Helmholtz Zentrum München and Technical University München.

Especially, I would like to express my gratitude to my supervisors Dr. Helmut Laumen. Thank you all for sharing with me your motivation, enthusiasm, passion for research, and for giving me the opportunity to be a part of many exciting projects and for creating an inspiring working environment. I received from you a perfect combination of guidance and support, as well as plenty of freedom and and responsibility during these four years of research. I also would like to especially thank Dr. Stephanie Hauck and Dr. Christine von Törne for their excellent cooperation in proteomics field. It was a wonderful experience to work with them, who provided a stimulating scientic environment, and without whom the research presented here would not have been possible.

Next, I would like to thank Christoph Hoffmann for providing HIB 1B cells andnhelping us creat the project design as well as for giving good advice during my PhD thesis. He showed me different ways to approach a research problem and the need to be persistent to accomplish any goal. Throughout the four years, I have been fortunate to have all extraordinary former and current member of the group, AG Hauner who helped me with my projects and with whom I shared not only an office, but also memorable times at conferences, dinner parties or coffee breaks. For their valuable friendship, encouragement I thank Dr, Simone Matthä, Tina Brand, Britta Fischer, Dr. Kerstin Ehlers, Vikoria Glunk, Kun Qian, Sebastian Dreyer. Special thanks to Dr, Melina Claussnitzer, Viktoria Glunk and Kun Qian who took the time to think about and discuss my work and gave me great input on experimental design.

I would also like to thank our amazing technical assistants, Elisabeth Hofmair, Manuela Hubersberger, Carola Herrmann not only for technical advicing in the lab, but also for their friendship and encouragement. Also thanks to Sylvia Heinrich for the help with paperwork and kindness.

I had multiple collaborations with groups at the Helmholtz Zentrum München, Germany and Karolinska Institutet, Sweden. Without Simone Wahl, Michaela Breier, Sophie Molnos Harald Grallert, Ingrid Dahlman, Prof. Peter Arner, an important part of my own paper would not exist.

Finally, let me also say ‘thank you’ to all my friends and my family from near and far who supported me throughout the whole time of the thesis, as well as everybody I unintentional forgot in this acknowledgment. Special thank to Anja Seffner for encouragement and helping me at any time during my study. Without her encouragement and constant helping, I could not have finished this work and study. Finally, I would like my family: my parents for giving me life in the first place, for educating me, for unconditional support and encouragement to pursue my interests. Thanks to my brother for believing in me and helping me whenever I need without complaining.

Thank you very much!

PERSONAL INFORMATION

Name	Heekyoung Lee (Dipl. Biotech.)
Address	Si-young Apt 20-103, seoung-san-2-dong, mapogu, 121-781 seoul, South korea
Telephone	(82)-10-9922-1830
Nationality	Republic of Korea

MY OBJECTIVES

"To pursue a postdoctoral position in the field of Molecular biology / Biochemistry where I can help advance the research interests of the group through my knowledge and skills and can be part of a team making a contribution to improving human health and welfare, and where I can learn and continue to develop my scientific skills."

EDUCATION

July 2010 – April 2014	Laboratory work of PhD thesis on the topic "Understanding the molecular mechanism underlying the effect of cis-regulatory variants on gene expression at type 2 diabetes (T2D) associated loci" with Prof. Dr. Hans Hauner under supervision of Dr. Helmut Laumen at the Technical University Munich (TUM), Else Kröner-Fresenius-Centre for Nutritional Medicine (EKFZ), Munich Germany in close cooperation with the Dr. Stefanie Hauck from the Protein Research Core Unit, Helmholtz Zentrum München, Germany
April 2008 – April 2009	Diploma thesis, entitled "CD95/CD95L non-apoptotic pathway through NF-κB" with Prof. Dr. Roland Eils under supervision of Dr. Eunice Hatada at the German Cancer Research Center (DKFZ), Heidelberg, Germany
April 2003 – May 2009	Diploma, Bielefeld University, Bielefeld, Germany- Molecular biotechnology
March 1999 – Dec. 2001	Konkuk University, Seoul, Southkorea- Animal science, not graduated

ADDITIONAL INFORMATION

- A PhD student with a broad range of skills in the analysis of signaling pathways, genetics, proteomics
- *Computer skills:* MS Word, Excel, Adobe illustrator, Power Point, Photoshop, ImageJ, Access, Hangeul, Statistical analysis using GraphPadPrism and SPSS
- *Technical skills:* Cell culture (also primary), Electrophoretic mobility shift assay (EMSA), Affinity chromatography, RT-PCR, Real Time PCR, Luciferase gene assay, Cloning, Transient transfections (siRNA, DNA plasmids), DNA, RNA isolation, Gel electrophoresis, Fluorescence microscopy, Differential staining, SDS-PAGE, Immunoblots, Preparation of cell lysates,

Bradford assay, Experimental design, Prepare tissue sections, Human primary adipocytes isolation, Retrovirus infection, Some experience with animal models, Some knowledge of FACS

- *Languages:* Korean (native), English, German (fluent)
- Experience as a private tutor (mathematics, science)
- Supervision of Master students (March 2012 ~ August 2012, Christina Bezold, April 2014 ~ present, Leili Jafari)
- Workshops/course participation
    - 2013 Useful Statistics and Publication Tips for Life Science PhDs, Technical University München, Germany
    - 2012 DZD Workshop, University of Tübingen, Germany
    - 2011 2nd International Workshop on Protein Analysis of Tissues, Helmholtz Zentrum München, Germany
    - 2010 DZD Workshop, University of Tübingen, Germany
- Former member of
    - Else Kroener-Fresenius-Centre for Nutritional Medicine, Chair of Nutritional Medicine, Technical University München, 85350 Freising-Weihenstephan, Germany
    - Research Centre for Nutrition and Food Sciences, Technical University München, 85350 Freising-Weihenstephan, Germany
    - Clinical Cooperation Group Nutrigenomics and Type 2 Diabetes, Helmholtz Zentrum München and Technical University München, 85350 Freising-Weihenstephan
    - Clinical German Center for Diabetes Research (DZD)
- Enjoys challenge in new fields and ideas; excellent ability to adapt to new situations and people; able to work well both independently as well as in a team; responsible and diligent.

BRIEF OUTLINE OF RESEARCH EXPERIENCE

July 2010 – April 2014: Prof. Dr. Hans Hauner's laboratory, Technical University Munich (TUM), Freising-Weihnstephan, Germany:

Genome-wide association studies (GWAS) have identified numerous risk loci (SNPs) associated with human diseases. However, most of the identified variants associated with diseases or trait are located in non-coding DNA regions, which hampered the further progress to assign the functional roles of those SNPs. Recently, advances in high throughput technologies enabled analysis of genome-wide expression quantitative trait loci in target tissues or cell types related to diseases or traits, suggesting that *cis*-regulatory SNPs might affect transcriptional regulation. However, an essential problem facing previously studies is a lack of validation data of the identification of *cis*-regulatoy SNPs. In only few studies it was shown that several transcription factors bind differentially at a SNP which alter gene expression. However, in most cases which transcription regulators and furthermore co-regulators binding to the *cis*-regulatory SNPs in which biological pathways still remain to be fully understood. Thus, combining bioinformatics and open-chromatin information with quantitative proteomics further supports the prediction of *cis*-regulatory variants and enables identification of allele-dependent binding of both, transcription factors (TFs) and co-regulators at the T2D associated loci. During my PhD thesis, I focused on the development of the label-free quantitative DNA protein interaction proteomics approach to identify allele-specific binding proteins, the identification of *cis*-regulatory SNPs and demonstration the biological role of those SNPs in target gene expression (*PPARG*, *TCF7L2* and *FTO*). Subsequently, I focused on the elucidation of the mechanisms mediating by *cis*-regulatory SNPs in the development of T2D through in-depth analysis of *PPARG* locus, and the efficient identification of both, transcription factors and co-regulators binding at the SNPs improved understanding of mechanisms underlying genetic associations.

April 2008 – April 2009: Prof. Dr. Roland Eils´s laboratory, German Cancer Research Center, Heidelberg, Germany:

CD95 is the best characterized death receptor inducing apoptosis; however, recent evidence indicates that CD95 is also involved in non-apoptotic processes. I have investigated the activation of NF-κB, one of the most important non-apoptotic signaling pathways mediated by CD95L, in two cancer cell lines, MCF-7 and HeLa. Using electrophoretic mobility shift assays (EMSAs) and immunoblots, I found that NF-κB activation by CD95L increased continuously over time, whereas TNF-mediated NF-κB activation showed a damped oscillation kinetic pattern. I also demonstrated by RT-PCR and qRT-PCR that the activation of NF-κB by CD95L led to the expression of the genes for A20, IκBα and IL-8 in both cell lines. Moreover, overexpression experiments suggested the involvement of RIP in the CD95L signaling pathway to NF-κB.

October 2007 - March 2008: Prof. Dr. Karsten Niehaus´s laboratory, Bielefeld University, Bielefeld, Germany:

Xanthomonas campestris is a bacterial species which causes a variety of plant diseases. It is used in the commercial production of a high molecular weight polysaccharide, xanthan gum, that has many important uses, especially in the food industry. I worked on the molecular cloning of three genes (*glg*A, *pgi and pfk*A) from *Xanthomonas campestri* encoding enzymes involved in the central metabolic pathway.

REFERENCES

- Prof. Dr. Hans Hauner, Chair for Nutritional Medicine at the Technical University of Munich and Director of the Else Kröner-Fresenius-Centre for Nutritional Medicine (EKFZ), Freising-Weihenstephan, Germany. E-mail: hauner@wzw.tum.de; Phone: (49)-8161-71-2001; Fax: (49)-8161-71-2097
- Dr. Helmut Laumen, Technische Universität München, Else Kröner-Fresenius-Centre for Nutritional Medicine (EKFZ), Freising-Weihenstephan, Germany. E-mail: helmut.laumen@tum.de; Phone: (49)-8161-71-2006; Fax: (49)-8161-71-2097
- Dr. Stefanie Hauck, Helmholtz Zentrum München, Research Unit Protein Science, Neuherberg, Germany. E-mail: hauck@helmholtz-muenchen.de; Phone: (49)-89-4140-3941; Fax: (49)-89-414-4426

LIST OF PUBLICATIONS AND CONFERENCE PRESENTATIONS

PUBLICATIONS

Claussnitzer M, Dankel SN, Klocke B, Grallert H, Glunk V, Berulava T, **Lee H**, Oskolkov N, Fadista J, Ehlers K, et al. 2014. Leveraging Cross-Species Transcription Factor Binding Site Patterns: From Diabetes Risk Loci to Disease Mechanisms. ***Cell*** 156: 343–358.

Kretschmer A, Möller G, **Lee H**, Laumen H, Toerne C, Schramm K, Prokisch H, Eyerich S, Wahl S, Baurecht H, et al. 2014. A common atopy-associated variant in the Th2 cytokine locus control region impacts transcriptional regulation and alters SMAD3 and SP1 binding. ***Allergy*** 69: 632–642.

Spieler D, Kaffe M, Knauf F, Bessa J, Tena JJ, Giesert F, Schormair B, Tilch E, **Lee H**, Horsch M, et al. 2014. Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. ***Genome Research*** 24: 592–603.

Lee H, von Toerne C, Claussnitzer M, Hoffmann C, Glunk V, Wahl S, Breier M, Molnos S, Grallert H, Dahlmann I, Arner P, Hauner H, Hauck SM, Laumen H. Unbiased allele-specific quantitative proteomics unravels molecular mechanisms modulated by *cis*-regulatory variation at the *PPARG* locus. **Submitted to PLOS Genetics (in review).**

CONFERENCE PRESENTATIONS

IR2013, Barcelona, Spain, 2013
XII International Symposium on Insulin Receptors and Insulin Action: Identification of allele-specific binding proteins at cis-regulatory variants using protein-DNA affinity-chromatography coupled with label-free quantitative proteomics analysis

IEG/GMC Symposium, Grassau, Germany, 2013
Identification of allele-specific binding proteins at *cis*-regulatory variants by MagBead-chromatography combined with label-free quantitative proteomics analysis