Integrated Computational and Experimental Analysis of the Neuroendocrine Transcriptome in Genetic Hypertension Identifies Novel Control Points for the Cardio-Metabolic Syndrome

Running title: Friese et al.; Analysis of the neuroendocrine transcriptome

Ryan S. Friese, PhD1,2; Chun Ye, PhD5; Caroline M. Nievergelt, PhD4; Andrew J. Schork, BS2; Nitish R. Mahapatra, PhD9; Fangwen Rao, MD2; Philip S. Napolitan, BS2; Jill Waalen, MD, MPH10; Geert W. Schmid-Schönbein, PhD1; Eleazar Eskin, PhD8; Daniel T. O’Connor, MD2,3,6,7 on behalf of the International Collaboration for Blood Pressure Genome Wide Association Studies (ICBP-GWAS)

1Depts of Bioengineering, 2Medicine, 3Pharmacology, 4Psychiatry, 5Bioinformatics Program, 6Institute for Genomic Medicine, University of California at San Diego; 7VA San Diego Healthcare System, San Diego; 8Depts of Computer Science & Human Genetics, University of California at Los Angeles, Los Angeles; 9Dept of Biotechnology, Indian Institute of Technology Madras, Chennai, India; 10Scripps Research Institute, La Jolla, CA

Address for correspondence:
Daniel T. O’Connor, MD
Department of Medicine
University of California at San Diego School of Medicine
VASDHS (0838)
Skaggs (SSPPS) Room 4256
9500 Gilman Drive
La Jolla, CA  92093-0838
Tel: (858)-5340661
Fax: (858)-5340626
E-mail: doconnor@ucsd.edu

Abstract:

Background - Essential hypertension, a common complex disease, displays substantial genetic influence. Contemporary methods to dissect the genetic basis of complex diseases, such as the genome-wide association study, are powerful, yet a large gap exists between the fraction of population trait variance explained by such associations and total disease heritability.

Methods and Results - We developed a novel, integrative method (combining animal models, transcriptomics, bioinformatics, molecular biology, and trait-extreme phenotypes) to identify candidate genes for essential hypertension and the metabolic syndrome. We first undertook transcriptome profiling on adrenal glands from blood pressure extreme mouse strains: the hypertensive BPH and hypotensive BPL. Microarray data clustering revealed a striking pattern of global underexpression of intermediary metabolism transcripts in BPH. The MITRA algorithm identified a conserved motif in the transcriptional regulatory regions of the underexpressed metabolic genes, and we then hypothesized that regulation through this motif contributed to the global underexpression. Luciferase reporter assays demonstrated transcriptional activity of the motif, via transcription factors HOXA3, SRY, and YY1. We finally hypothesized that genetic variation at HOXA3, SRY, and YY1 might predict blood pressure and other metabolic syndrome traits in humans. Tagging variants for each locus were associated with BP in a human population BP extreme sample, with the most extensive associations for YY1 tagging SNP rs11625658, on SBP, DBP, BMI, and fasting glucose. Meta-analysis extended the YY1 results into two additional large population samples, with significant effects preserved on DBP, BMI, and fasting glucose.

Conclusions - The results outline an innovative, systematic approach to the genetic pathogenesis of complex cardiovascular disease traits, and point to transcription factor YY1 as a potential candidate gene involved in essential hypertension and the cardio-metabolic syndrome.

Key words: BPH mouse strain; complex trait; essential (genetic) hypertension; human genetics; metabolic syndrome
Introduction

Essential hypertension, a common disease, displays substantial genetic influence, with heritability estimates for blood pressure up to ~50%\(^1\). The complex, multifactorial nature of this disorder, however, has made it difficult to identify underlying genetic contributors. Investigation of such complex traits with novel and integrative methods might be required to completely understand their genetic basis.

In the current report, we present a novel method that begins with genome-wide transcriptome profiling, and then sequentially integrates several disciplines and tools (animal models, microarray technology, bioinformatics, molecular biology, extreme phenotypes, and human genetics) to create, filter, and refine a list of candidate genes for essential hypertension and the metabolic syndrome (a cluster of risk factors for cardiovascular disease, including essential hypertension, insulin resistance, dyslipidemia, and elevated plasma cholesterol).

Our analysis is based on two inbred, blood pressure extreme mouse strains, the hypertensive BPH (Blood Pressure High) and the hypotensive BPL (Blood Pressure Low)\(^2\). The use of extreme phenotypes has been suggested for the study of complex traits as a strategy to increase the power of detecting underlying genetic determinants as well as rare variants\(^3\)-\(^6\). The BPH strain parallels human essential hypertension, with elevated blood pressure, increased heart rate, and early mortality\(^7\). Even though BPH was developed through selection solely on blood pressure, the strain exhibits several metabolic syndrome-like abnormalities, such as decreased plasma HDL cholesterol, enhanced rate of body weight gain on a high fat diet, and increased pulse rate\(^8\).

We chose the adrenal gland as the target of microarray analysis of BPH and BPL since its secretory products, both medullary and cortical, determine endocrine, cardiovascular, and
sympathetic functions important in hypertension and the metabolic syndrome. For example, medullary catecholamines act through G-protein-coupled adrenergic receptors to modulate force of contraction of the heart and constriction of blood vessels, and regulate carbohydrate and lipid metabolism through effects on glucagon and insulin secretion, glycogenolysis, glycolysis, and lipolysis. Cortical steroid hormones regulate reabsorption and secretion of sodium and potassium (mineralocorticoids), and modulate carbohydrate and lipid metabolism (e.g., through effects on peripheral glucose uptake, gluconeogenesis, and lipolysis), as well as inflammation (glucocorticoids).

Following microarray analysis, bioinformatic, molecular biology, and human genetic tools were applied sequentially to identify significant association (by meta-analysis) of YY1 rs11625658 SNP genotype with DBP, BMI and glucose across three large human cohorts, suggesting shared genetic determination for such metabolic syndrome traits. Uniquely, we focused on transcription factors as candidate gene “master switches” since functional changes in them are likely to be pleiotropic and, therefore, provide a unifying genetic mechanism for multiple traits of the metabolic syndrome.

Methods

Mouse strains

Juvenile (~5-week-old) and adult (~12-week-old), hypertensive (strain=BPH) and hypotensive (strain=BPL) inbred male mice were obtained from colonies at the Jackson Laboratory (Bar Harbor, ME). Systolic blood pressure of the BPH increases early in life. At 7 weeks of age, BPH systolic blood pressure (110 mmHg) is 40 mmHg higher than BPL (70 mmHg)\(^7\). The systolic blood pressure differential increases to 60 mmHg by 21 weeks of age (BPH=130 mmHg; BPL=70 mmHg)\(^7\). Mice were studied according to a protocol approved by the Animal Subjects
Committee of the University of California at San Diego, and research was conducted in accordance with institutional guidelines.

**Microarray experiments**

We previously presented genome-wide gene expression profiles of adrenal glands from BPH and BPL mice using standard Affymetrix (Santa Clara, CA) protocols and MG-U74Av2 GeneChips. In short, adrenal gene expression of BPH and BPL mice (n=3, juvenile BPH; n=3, adult BPH; n=3, juvenile BPL; and n=3, adult BPL) was determined using MG-U74Av2 GeneChips, and statistically significant changes in gene expression were determined with 2-factor analysis-of-variance (ANOVA) in order to assess the impact of strain (BPH vs. BPL), age (juvenile vs. adult), and strain-by-age interaction on gene expression.

The primary focus of the current report is the impact of strain on gene expression: BPH (3 juvenile BPH + 3 adult BPH) vs. BPL (3 juvenile BPL + 3 adult BPL). Juvenile BPH mice (5-weeks old) are “pre-hypertensive” — they do not yet have maximal elevation of blood pressure. Pre-hypertensive animals are useful in genetic studies since the effects of confounding factors (e.g., age, maximal blood pressure elevation) on gene expression are minimized. We also utilized 2-factor ANOVA to compare BPH with BPL gene expression across two age groups (juvenile and adult), in order to identify genes demonstrating a consistent pattern of differential expression (i.e., underexpressed in both juvenile and adult BPH, or overexpressed in both juvenile and adult BPH) in both preliminary and advanced stages of hypertensive disease. This consistent level of expression across two age groups is a key part of the analysis. Such analysis might enrich the set of differentially expressed genes for underlying genetic determinants of hypertension with effects independent from (or resistant to) confounding changes in age, blood pressure, and hypertensive disease processes. Genes with consistent patterns of expression across both age
groups might be stronger candidates for pathogenic drivers of disease than genes with changing
or inconsistent expression patterns. Microarray data are available at the NCBI Gene Expression
Omnibus database (http://www.ncbi.nlm.nih.gov/geo) under the following accession numbers:
GSE1674, GSE19342.

Our analysis was designed to gain a broad perspective on which biochemical pathways
and physiological systems exhibit global changes, rather than focus on specific changes in
individual candidate genes. Towards this end, we used GenMAPP v2.1
(http://www.genmapp.org)\textsuperscript{11} to perform functional clustering and pathway analysis. Fisher’s
exact tests were used to determine if significantly perturbed pathways (identified with
GenMAPP) displayed patterns of global under- or over-expression.

\textbf{Computational motif identification}

We defined a gene “regulatory sequence” to be the sequence of DNA from -1000 bp upstream to
+1000 bp downstream of the transcription start site, in order to capture the majority of the
proximal promoter transcriptional elements upstream of the TATA box, as well as any
transcriptional regulatory motifs within the 5’-untranslated region (5’-UTR) or first intron.

Regulatory sequences were extracted from an in-house database of the \textit{Mus musculus} genome,
originally downloaded from the University of California, Santa Cruz, genome database
(http://genome.ucsc.edu). We used MITRA\textsuperscript{12, 13} to discover conserved motifs within a set of
regulatory sequences. The MITRA algorithm identifies common motifs in unaligned DNA
sequences by maximizing a score that discriminates sequences between a positive set (i.e., our
set of regulatory sequences hypothesized to contain a common motif) and a negative set (i.e., the
set of regulatory sequences from all genes not represented in the positive set). The algorithm was
constrained to search for motifs 10 bp in length on both the forward and reverse DNA strands.
WebLogo\textsuperscript{14, 15} was used to create a consensus sequence representation of the motifs identified with MITRA.

\textbf{Computational prediction of transcription factor binding}

Predictions for transcription factors that bind the MITRA-identified motif were performed with CONSITE (which utilizes the JASPAR database; http://www.phylofoot.org/consite)\textsuperscript{16} and P-MATCH (which utilizes the TRANSFAC database; http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi)\textsuperscript{17}, two web-based tools for identifying \emph{cis}-regulatory elements in genomic sequences. Predictions for transcription factors that bind the SV40 promoter of the pGL3-Promoter control vector (Promega, Madison, WI) were performed with CONSITE.

\textbf{Construction of promoter or enhancer/luciferase reporter plasmids}

Cloning of the motif oligonucleotide sequences was accomplished at the unique XhoI (promoter) or SalI (enhancer) restriction enzyme sites in pGL3-Promoter. The two oligos [5’-(Phosphate) TCGAGACCATAGATAC-3’; 5’-(Phosphate) TCGAGTATCTATGGTC-3’] were annealed in the following reaction mixture: 162.5 picomole of each oligo plus water for a total of 75 \(\mu\)L, and 25 \(\mu\)L of annealing buffer (100 mM Tris pH 7.5, 100 mM MgCl\(_2\), 10 mM DTT). The annealing reaction mixture was heated at 90°C for 5 minutes and then allowed to cool to room temperature over a period of 45 minutes. To test the transcriptional effects of the motif when it is located in the proximal promoter region of a gene, the annealed oligos were inserted into the unique XhoI site just upstream of the SV40 promoter in the multiple cloning site of the firefly luciferase reporter vector, pGL3-Promoter (Promega, Madison, WI). To allow for testing of transcriptional “enhancer” (distant) effects of the motif, the annealed oligos were inserted in the SalI restriction enzyme site just downstream of the luciferase gene poly-A signal in the pGL3-Promoter vector (Promega, Madison, WI). The oligos inserted in the “enhancer” position were designed with the
following sequence: 5’-(Phosphate) CTAGCACCATACTAGATC -3’; 5’-(Phosphate) TCGAGTATCTATGCTGTG -3’. A total of four types of luciferase reporter plasmids were constructed with the motif being inserted in either the “promoter” or “enhancer” position, in either the forward or reverse orientation. Insertion of the motif in two directions enabled testing of orientation specific effects.

Adrenal cell promoter/reporter transfection and luciferase activity assays

Rat PC12 pheochromocytoma cells [grown in DMEM high glucose (Invitrogen) with 5% heat-inactivated fetal bovine serum (Gemini Bioproducts, Woodland, CA), 10% heat-inactivated horse serum (Gemini Bioproducts), penicillin (100 U/ml), streptomycin (100 μg/ml), and L-glutamine (0.292 mg/ml)] were transfected (at 50-60% confluence, 1 day after splitting 1:4) with motif-promoter (XhoI; forward or reverse orientation) or motif-enhancer (SalI; forward or reverse orientation) reporter vector DNA [1 μg supercoiled DNA per well; 12-well polystyrene plates (coated with poly-L-lysine; Sigma), 2.2-cm diameter wells, Corning Inc., Corning, NY] using the liposome method (Superfect; Qiagen). Co-transfection of mammalian expression (CMV-promoter) transcription factor plasmids was performed with 50 ng of pcDNA3.1(-) (empty vector control) (Invitrogen, Carlsbad, CA), HoxA3 (Mus musculus; MMM1013-98478526, Open Biosystems, Huntsville AL), MEF2A 18 (Homo sapiens), RunX1 (Mus musculus; MMM1013-9498895, Open Biosystems, Huntsville AL), SOX9 (Homo sapiens; MHS1010-9205725, Open Biosystems, Huntsville AL), Sox17 (Danio rerio; MDR1734-97029554, Open Biosystems, Huntsville AL), Sry (Rattus norvegicus; Sry1/pcDNA3.1) 19, or YY1 (Homo sapiens; SC118004, OriGene, Rockville, MD). Cells were lysed 20 hours after transfection with lysis buffer (300 μL per well) [0.1 M phosphate buffer (K2HPO4 + KH2PO4) (pH 7.8), 1 mM DTT, and 0.1% Triton-X 100].
The bioluminescent activity of luciferase in 80 µL of cell lysate was determined with the AutoLumat LB 953 luminometer (EG&G Berthold, Nashua, NH) by measuring light emission (incubation time = 0 seconds, measure time = 10 seconds, temperature = 25°C) after addition of assay buffer [100 µl per sample; 100 mM Tris-acetate (pH 7.8), 10 mM Mg-acetate, 1 mM EDTA (pH 8.0), 3 mM ATP, and 100 µM luciferin (Sigma-Aldrich)]. As a control for varying cell number between individual wells, the total protein content was measured in the cell lysate using the Bio-Rad Protein Assay (coomassie blue dye absorbance shift; based on the Bradford method) (Bio-Rad, Hercules, CA). Luciferase activity (n=4-5 wells/condition) is expressed as the normalized ratio of (luciferase activity)/(total protein) or (RLU/µg protein). Data were analyzed with analysis-of-variance (ANOVA), followed by pairwise t-tests corrected for multiple comparisons (Bonferroni).

**Human subjects.**

Population blood pressure extremes. We previously published a detailed description of the human subjects and subject selection criteria used in this investigation. Unrelated adults from the Kaiser-Permanente Medical Group (subscription-based Health Maintenance Organization, HMO) primary care population located in San Diego, CA, were selected from the lowest and highest (extreme) percentiles of diastolic blood pressure (DBP) distribution (<63 mm Hg or ≥90 mm Hg); subjects were ascertained on the DBP trait, because twin and family studies provide evidence that DBP is substantially heritable and SBP correlates highly with DBP. Two independent DBP extreme samples (cohort-1: n=996; cohort-2: n=1075) were obtained from different individuals within the source population. Blood pressure was measured in seated subjects using brachial sphygmomanometry. If DBP was elevated, repeat measurement was obtained for verification (only the initial value was reported; values were not averaged).
Approximately 48% of the hypertensive subjects from the upper DBP extreme reported being prescribed and taking ≥1 anti-hypertensive medication. Subjects in the lower extreme DBP group reported no history of hypertension or anti-hypertensive medication. Subjects did not have renal failure (serum creatinine concentration was ≤1.5 mg/dL in 98.6% of subjects). Self-identified ethnicity (including that of both parents and all 4 grandparents) for all subjects was specified as white (European ancestry). Medical information was obtained from annual health appraisal visits and questionnaire. Blood for preparation of genomic DNA was obtained with informed consent, and samples were de-identified.

*Extension of human genetic association: ICBP-GWAS (International Consortium of Blood Pressure – Genome Wide Association Studies).* Extension of associations in our two blood pressure extreme cohorts was sought in the International Consortium of Blood Pressure (ICBP GWAS)\(^\text{25}\). Complete details of ICBP methodology have previously been presented\(^\text{26}\). In short, ICBP data (http://www.igm.jhmi.edu/~gehret/icbp32413ahsfd134/icbp_088023401234-9812599.html) were analyzed in separate genome wide meta-analyses for SBP and DBP. Before meta-analysis, the association results for each cohort were filtered to exclude SNPs not in HapMap, SNPs with alleles different from HapMap, and SNPs with observed/expected ratio of coded genotype scores less than 0.3. A genomic control correction was applied. Where studies stratified their analyses by sex, genomic control was applied within each sex stratum. For each SNP, the per-coded-allele effects were combined across studies (and across sex strata, when used) using inverse variance weighted meta-analysis.

*Genotyping and statistical analysis.*

The HapMap (http://hapmap.ncbi.nlm.nih.gov) was used to select common (minor allele frequency >5%), validated tagging SNPs (single nucleotide polymorphisms) that capture the
linkage disequilibrium (LD) and haplotype block structure within candidate genes, using CEU (European ancestry) subject data. By inspection of CEU LD heat plots, each target locus (HOXA3, 7.8 kbp; YY1, 39.7 kbp; and SRY at 896 bp) was spanned by a single LD block in CEU subjects. A total of 7 SNPs were selected to tag these 3 loci (at 2-3 SNPs per locus): two at HOXA3 (rs10085570, rs6948297), two at SRY (rs2058276, rs1865680), and three at YY1 (rs8021803, rs11625658, rs4905941). The HOXA3 and YY1 tagging SNPs were located within introns. SRY tagging SNPs were located adjacent to the SRY gene in the 5’-upstream region.

Single nucleotide polymorphism (SNP) genotyping was performed on subjects’ genomic DNA with the matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF) system developed by Sequenom, as previously described20.

Statistical association between individual SNPs and individual metabolic syndrome traits was tested with univariate analysis-of-variance (ANOVA) [independent variable = diploid genotype (additive model); dependent variable = one metabolic syndrome trait; covariates = age, sex]. A False Discovery Rate (FDR <0.05) was used to control for testing of multiple genotypes and phenotypes, as described27 and applied28, 29 in the literature. A total of 6 phenotypes were tested for association: systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), (fasting) plasma glucose, total cholesterol, and HDL cholesterol. Plasma triglyceride and LDL cholesterol data were not available in these subjects. A second control for testing of multiple phenotypes was also achieved using multivariate analysis-of-variance (MANOVA), wherein statistical association between individual SNPs and multiple metabolic syndrome traits was evaluated simultaneously using one test (MANOVA: independent variable = diploid genotype (additive model); dependent variable = SBP, DBP, BMI, plasma glucose, plasma cholesterol, and/or plasma HDL cholesterol; covariates = age, sex). We included age and
sex as covariates in order to control for their effects on blood pressure and metabolic phenotypes. No adjustment was made for the use of anti-hypertensive medication. Blood pressure data can be adjusted for anti-hypertensive therapy (with, for example, the method described by Cui30), however, without adjustment our analysis was likely more conservative, i.e., biased toward the null (insignificance), since adjustment would tend to disproportionately increase the mean BP of diploid genotype groups consisting of a higher proportion of hypertensive subjects. The effect size for significant SNPs is presented as partial eta-squared ($\eta^2$): the proportion of total trait variance attributable to a specific SNP. Statistical analyses were performed using SPSS software (SPSS Inc., Chicago, IL). Data are presented as mean ± standard error of the mean (SEM).

To evaluate the cumulative effect of genetic variation at YY1 on metabolic syndrome traits in the 3 groups (San Diego cohorts 1&2; ICBP-GWAS) we turned to meta-analysis, using the outcomes of regression analysis (additive model) in each group, focusing on the effect size beta (slope per allele) and the SE of beta, using the command METAN within the program STATA (Stata Corporation, College Station, TX <http://www.stata.com>, reporting results of fixed-effect (i.e., genotype as independent variable) models.

Results

Microarray adrenal mRNA expression analysis

Previous statistical analysis of the BPH and BPL adrenal microarray data identified 2004 significantly differentially expressed genes—approximately 16% of all probe sets9,10 (Supplementary Table 1). The distribution of differential expression was split approximately evenly between overexpression (924/2004 genes, or 46%) and underexpression (1080/2004 genes, or 54%) in BPH versus BPL.
We used GenMAPP to functionally cluster the microarray data into statistically significant biochemical systems and pathways. GenMAPP identified 7 biochemical pathways within the intermediary metabolism domain (the set of canonical biochemical pathways responsible for intra-cellular energy production, i.e., ATP formation resulting from carbohydrate or lipid metabolism) of the BPH that were significantly perturbed: the electron transport chain (p<0.0001), fatty acid degradation (p<0.01), fatty acid synthesis (p<0.001), glycolysis/gluconeogenesis (p<0.001), mitochondrial long chain fatty acid beta-oxidation (p<0.001), pentose phosphate (p<0.004), and TCA (tricarboxylic acid) cycle (p<0.001) pathways (Supplementary Table 2). In addition, all of the significantly perturbed intermediary metabolism pathways were also globally under-expressed in BPH: the electron transport chain (p<0.0001), fatty acid degradation (p=0.011), fatty acid synthesis (p=0.031), glycolysis/gluconeogenesis (p=0.006), mitochondrial long chain fatty acid beta-oxidation (p=0.021), pentose phosphate (p=0.031), and TCA (tricarboxylic acid) cycle pathways (p=0.0005) (Fisher’s exact test; Supplementary Table 2). The TCA cycle pathway exhibited a pattern of global underexpression representative of that observed in all of the intermediary metabolism pathways (Supplementary Figure 1). In total, 82 genes of the intermediary metabolism pathways were differentially expressed, with the directional pattern of expression significantly and globally shifted towards underexpression: 79 of 82 genes were underexpressed, while 3 of 82 genes were overexpressed (p<0.0001; Fisher’s exact test).

Computational promoter motif identification

The MITRA algorithm was used to identify statistically over-represented transcription factor binding motifs in regulatory sequences from the set of 82 differentially expressed genes (79 underexpressed, 3 overexpressed) in intermediary metabolism of the BPH adrenal gland. MITRA
identified one significantly overrepresented (p<0.05) 10-bp motif that was present a total of 55 times in 38 of the 82 (38/82=46%) differentially expressed metabolic genes: 5’-ACCATAGNTN-3’ (Figure 1, Supplementary Table 3). The motif was not localized to a limited set of sub-pathways of intermediary metabolism (e.g., only in the TCA cycle and pentose phosphate sub-pathways), but instead, was present in an approximately equal proportion of differentially expressed genes (~50%; p=0.97, Chi-square test) in each of the sub-pathways of intermediary metabolism.

**Promoter/luciferase reporter assays: Determination of motif transcriptional activity**

Promoter/luciferase reporter plasmid constructs were used to determine if the motif conferred transcriptional activity. Since positions 8 and 10 in the motif consensus sequence (5’-ACCATAGNTN-3’) lacked nucleotide preference (“N”), adenine (“A”) was randomly and independently chosen to fill each “N” position in the motif (5’-ACCATAAGATA-3’). The motif was inserted in two regions of the pGL3-Promoter luciferase reporter plasmid (driven by the SV-40 promoter): 1) in the “promoter” site just upstream of the SV-40 promoter; and 2) in the “enhancer” site downstream of the luciferase gene. The motif was inserted in either the forward (sense strand: 5’-ACCATAAGATA-3’) or reverse (sense strand: 5’-TATCTATGGT-3’) orientation. Luciferase assays were performed in PC12 cells—a cell line derived from a rat adrenal medullary chromaffin cell tumor (pheochromocytoma)\textsuperscript{31}.

The motif significantly decreased transcriptional activity when inserted into the “promoter” position in both forward (0.50 ± 0.04 RLU/µg protein; p=0.0003) and reverse orientations (0.66 ± 0.06 RLU/µg protein, p=0.0055) compared to the control vector without motif insert (1.00 ± 0.07 RLU/µg protein) (Figure 2). There was no significant difference between luciferase activity of the promoter-forward and promoter-reverse orientation constructs.
The motif did not significantly affect transcripational activity when inserted into the “enhancer” position, in either the forward (1.02 ± 0.04 RLU/µg protein) or reverse orientations (0.94 ± 0.07 RLU/µg protein) compared to the control vector without motif insert (1.00 ± 0.07 RLU/µg protein) (Figure 2).

*Computational prediction of motif/transcription factor binding*

The CONSITE and P-MATCH algorithms were used to predict which transcription factors bind to the motif identified with MITRA (Table 1). CONSITE identified 6 putative transcription factors: MEF2A – myocyte enhancer factor 2A, RUNX1 – runt-related transcription factor 1, SOX9 – SRY-box 9, SOX17 – SRY-box 17, SRY – sex determining region Y, and YY1 – yin yang 1. P-MATCH identified 4 putative transcription factors: HOXA3 – homeobox A3, RUNX1, SRY, and YY1. The MEF2A, RUNX1, SOX9, and SOX17 transcription factors were also predicted to bind the SV-40 promoter, the promoter that drives expression of the pGL3-Promoter vector (the control vector in the luciferase reporter experiments).

*Luciferase assays: Trans-activation by transcription factor co-transfection*

Promoter/luciferase reporter co-transfection experiments were performed with the promoter-forward and promoter-reverse orientation constructs, as well as cDNA expression plasmids for the transcription factors computationally predicted to bind the motif. One set of transcription factors (HOXA3, SRY, YY1) had no effect on the control vector, while significantly increasing the expression of the promoter-forward or promoter-reverse constructs (Figure 3, Table 2). More specifically: HOXA3 increased expression of the promoter-reverse construct by 1.24-fold (p<0.05); SRY increased expression of the promoter-forward construct by 1.26-fold (p<0.05); and YY1 significantly increased expression of the promoter-forward construct by 1.49-fold (p<0.05) and the promoter-reverse construct by 1.63-fold (p<0.05).
Another set of transcription factors (RUNX1, SOX9, SOX17) significantly decreased expression of the insertless control vector (SV-40 promoter, pGL3-Promoter) while also reducing expression of the promoter-forward and/or promoter-reverse constructs (Figure 4, Table 2). Specific changes included: RUNX1 reduced expression of the promoter-forward construct by 0.69-fold (p<0.05); SOX9 reduced expression of the promoter-forward construct by 0.64-fold (p<0.05); SOX17 reduced expression of the promoter-forward construct by 0.73-fold (p<0.05) and the promoter-reverse construct by 0.75-fold (p<0.05). MEF2A reduced expression of the control vector by 0.77-fold (p<0.05) but had no effect on the promoter-forward or promoter-reverse constructs.

**Human blood pressure extremes: Statistical genetic associations with the metabolic syndrome.**

**Initial study.** The HapMap was used to select common (minor allele frequency >5%), validated tagging SNPs (single nucleotide polymorphisms) that capture the linkage disequilibrium (LD) and haplotype block structure within the human HOXA3, SRY, and YY1 loci. These tagging SNPs were genotyped in a sample from extremes of blood pressure in the San Diego population (all SNPs were genotyped in cohort-1): White (European ancestry) males and females from the lowest and highest percentiles of DBP distribution (≤63 mm Hg or ≥90 mm Hg). Statistical association was tested between tagging SNPs and several traits of the human metabolic syndrome: systolic blood pressure (SBP), diastolic blood pressure (DBP), body mass index (BMI), plasma glucose, plasma total cholesterol, and plasma HDL cholesterol (Table 3).

In the first cohort, **HOXA3 SNP rs10085570** showed significant association with DBP (p=0.017). The **SRY-tagging SNP rs2058276** in males significantly associated with SBP (p=0.027), DBP (p=0.018) and BMI (p=0.004). **YY1 SNP rs8021803** showed a significant association with both SBP (p=0.004) and DBP (p=0.002). **YY1 SNP rs11625658** significantly
associated with SBP (p=0.0002), DBP, (p=0.0006), BMI (p=0.006), and plasma glucose (p=0.024). Multivariate analysis-of-variance (MANOVA) for YY1 rs11625658 revealed significant association between YY1 genotype and the following set of metabolic syndrome traits: SBP, DBP, BMI, and plasma glucose (p=0.0005; effect size = 0.012 or 1.2% for the joint set of traits; specific individual trait effect size and directionality are shown in Tables 3 and 4). No significant associations were found for plasma total cholesterol or plasma HDL cholesterol for any of the HOXA3, SRY, or YY1 SNPs.

SNPs found to be significantly associated in cohort-1 were then examined within San Diego BP extreme cohort-2. Only YY1 SNP rs1625658 showed significant association in this second cohort; HOXA3 SNP rs10085570, SRY SNP rs2058276, and YY1 SNP rs8021803 did not show significant association in cohort-2 (data not shown).

Potential interdependence of the statistical associations of SBP, DBP, BMI, and plasma glucose (4 correlated traits) with YY1 SNP rs11625658 was examined using a modified statistical model wherein the four traits were systematically used as covariates for each other (using the FDR <0.05 multiple testing correction; age and sex remained as covariates in all analyses). In cohort-1, associations with SBP and DBP remained significant when BMI and/or plasma glucose were added as covariates; and BMI retained significance when plasma glucose was added as a covariate (data not shown).

**Meta-analysis.** Since rs11625658 showed the highest degree of association in cohort-1 and preliminary association in cohort-2, we sought to extend the results of its significant SBP, DBP, BMI, and glucose associations into two additional population samples: San Diego BP extreme cohort-2 (n=1075) and the ICBP-GWAS (n=66,741), for a total of n=68,812 subjects (Table 4). By meta-analysis, considering the effect size (beta, or slope per allele) and its SE for each
subgroup, we found nominally significant effects across the 3 groups for YY1 tagging variant rs11625658 on DBP (p=0.033), BMI (p=0.004), and fasting glucose (p=0.038). Inspection of beta (slope) values revealed directionally consistent effects in each subgroup, though inverse allelic effects (i.e., positive versus negative beta-slope values) were observed for DBP and BMI. SBP and cholesterol (total or HDL) were not significantly affected in the meta-analysis (each p>0.05). The previously reported, isolated effects of rs11625658 on SBP (p=0.502) and DBP (p=0.047) by the ICBP-GWAS, parallel the results of our meta-analysis. The Genetic Investigation of ANthropometric Traits (GIANT) consortium GWAS of BMI in n=249,796 individuals reported p=0.0603 for SNP rs11625658.

Discussion

Despite the substantial heritability of blood pressure, the genetic underpinnings of hypertension remain incompletely understood. Investigation of such complex traits may benefit from a comprehensive set of tools that includes not only the GWAS, but also integrative and novel methods that might ultimately be required to completely reveal complex trait genetic determinants. To this end, we developed a novel approach, integrating tools from several disciplines, including bioinformatics, molecular biology, and human genetics, to identify novel candidate genes for essential hypertension and the cardio-metabolic syndrome (Supplementary Figure 2).

Microarray expression analysis

We began with genome-wide gene expression analysis of adrenal glands from the BPH and BPL blood pressure trait extreme inbred mouse strains. Although the widespread expression changes (up to ~16%) were initially surprising (Supplementary Table 1), they likely reflected the
polygenic nature of essential hypertension, the comparison of extreme phenotypes, and the diverse adaptations or responses to disease processes. Functional clustering uncovered a striking pattern of global underexpression of intermediary metabolism pathways in the BPH (Supplementary Table 2), and we hypothesized that such consistent underexpression throughout interacting and functionally related pathways resulted from a common transcriptional mechanism.

**Computational motif identification**

The MITRA algorithm identified a conserved motif in the set of transcriptional regulatory sequences from the 82 differentially expressed genes of the intermediary metabolism sub-pathways (Figure 1). The motif was present in approximately half (38 of 82) of the differentially expressed genes (Supplementary Table 3) in an equal proportion in each of the metabolic sub-pathways, suggesting that if the motif altered transcription, it was likely to be important in regulation of not only each individual sub-pathway but also the function of intermediary metabolism as a whole.

**Determination of motif transcriptional activity**

Promoter/luciferase reporter assays demonstrated that the motif was functional in the promoter position, in both the forward and reverse orientations, but lacked activity in the enhancer position (in either orientation) (Figure 2). Classical *cis*-acting proximal promoter elements are typically dependent on both strand orientation and distance from the transcriptional start site, while classical *trans*-acting enhancer elements are independent of orientation and distance. If we assume the binding of only one transcription factor to the motif, the distance-dependent yet orientation-independent activity of the motif could represent an alternative mode of action for proximal promoter elements. It is also conceivable that the motif consensus sequence represents
a composite of two or more distinct motifs, with each one directing binding of different transcription factors in either the forward or reverse orientation.

**Transcription factor identification**

We then queried TRANSFAC and JASPAR to identify transcription factors that might bind the motif and, therefore, regulate its cis-acting transcriptional activity (Table 1). Expression plasmids for these transcription factors (HOXA3, MEF2A, RUNX1, SOX9, SOX17, SRY, and YY1) were co-transfected with the luciferase reporter constructs (with the motif inserted only in the promoter position) to determine which transcription factors could modulate transcription through the motif.

One important caveat of the luciferase experiments was that both the control vector (without motif insert) and the promoter-forward and -reverse constructs (with the motif insert) contained the SV-40 promoter (in pGL3-Promoter) to direct basal eukaryotic transcription. Transcription factor effects on the motif were therefore difficult to distinguish from effects on the SV-40 promoter if transcriptional activity was altered in both the control vector (pGL3-Promoter) and in the motif promoter-forward or -reverse constructs; such was the case for RUNX1, SOX9, and SOX17, which significantly decreased expression of both the control vector and the promoter-forward and/or promoter-reverse constructs (Figure 4, Table 2). Computational analysis of the SV-40 promoter revealed RUNX1, SOX9, and SOX17 binding sites, so their effects on the control vector were not unexpected.

Thus HOXA3, SRY, and YY1 emerged as the strongest candidates for transcriptional regulation of the motif, since they altered luciferase expression of the motif-containing constructs but lacked effects on the control vector (without a motif insert) (Figure 3, Table 2). These three candidate transcription factors had distinct patterns of regulation: HOXA3 had no
effect on the promoter-forward construct but increased expression of the promoter-reverse construct; SRY increased expression of the promoter-forward but had no effect on the promoter-reverse construct; YY1 increased expression of both the promoter-forward and the promoter-reverse constructs. It is conceivable that the motif consensus sequence might form a composite of HOXA3, SRY, and YY1 motifs, wherein binding specificity to target genes in vivo is dictated by strand orientation, deviation of the actual motif sequence from the consensus (shown in Supplementary Table 3), and the relative abundance and/or activity of endogenous HOXA3, SRY, YY1 and their co-factors in the nucleus.

There is precedent that inbred rodent models of genetic hypertension, such as the BPH, exhibit metabolic abnormalities. The SHR (Spontaneously Hypertensive Rat), the most widely studied inbred model of genetic hypertension, was developed in a selection paradigm similar to that of the BPH (i.e., selection only on the basis of elevated blood pressure), yet the SHR also exhibits dyslipidemia and insulin resistance, and is widely studied as model of the metabolic syndrome.

Though the BPH has not been as thoroughly investigated as SHR as a model of the metabolic syndrome, the BPH strain exhibits several metabolic abnormalities, including decreased plasma HDL cholesterol, enhanced rate of body weight gain on a high fat diet, and increased pulse rate. Perhaps pleiotropic genetic variants that affect both blood pressure and metabolism were fixed during the blood pressure selection program for the BPH. If such “master switches” exist, transcription factors (such as HOXA3, SRY, and YY1) would be logical and indeed compelling candidates. Since we identified HOXA3, SRY, and YY1 as candidate transcriptional regulators of perturbed intermediary metabolism in an inbred rodent model of human genetic hypertension, we hypothesized that blood pressure and other metabolic syndrome traits in humans might stem from genetic variation in the HOXA3, SRY, and YY1 loci.
Human SNP genotyping and statistical genetic association.

We then genotyped common HapMap tagging SNPs that captured the linkage disequilibrium (LD) structure within the human HOXA3, SRY, and YY1 loci. These tagging SNPs were located in non-coding regions (i.e., upstream of the gene or within an intron) of our candidate genes. Statistical association of the SNPs with several traits of the metabolic syndrome was tested in a sample of population blood pressure extremes (white males and females from the lowest and highest percentiles of DBP distribution). Although our blood pressure extreme groups were ascertained on a DBP criterion, recent evidence indicates that SBP is at least as important a risk factor for target organ damage; we plan future studies to explore the potential effect of polymorphism in isolated systolic hypertension.

Transcription factor YY1 emerged as the most compelling candidate gene for the human metabolic syndrome. In the BPH, YY1 was differentially expressed across both age groups (in juvenile pre-hypertensives and adult hypertensives), as well as in only in the juvenile pre-hypertensives (data not shown). In cohort-1, YY1 SNP rs11625658 significantly associated with not only blood pressure (SBP: p=0.0002; DBP: p=0.0006), but also BMI (p=0.006), and carbohydrate metabolism (plasma glucose: p=0.024), though not with lipid metabolism (neither plasma total cholesterol nor plasma HDL cholesterol) (Table 3). Furthermore, multivariate analysis (MANOVA) demonstrated that rs11625658 genotype significantly associated with SBP, DBP, BMI, and plasma glucose as a joint set of traits (p=0.0005). Finally, meta-analysis (Table 4) indicated that YY1 variant rs11625658 continued to predict DBP, BMI, and glucose in a very large number of subjects (n = up to 68,812), with unexpectedly inverse (or opposite) allelic effects on DBP and BMI. Studies of rs11625658 in additional human populations would bolster YY1 as a candidate gene for hypertension and the metabolic syndrome.
YY1 is a ubiquitous, multifunctional zinc-finger transcription factor with fundamental roles in biological processes such as embryogenesis, differentiation, cellular proliferation, and cell cycle progression\textsuperscript{35}. Indeed, homozygous knockout of the \textit{YY1} gene is lethal\textsuperscript{36}. The actions of YY1 are complex, since the transcription factor can directly or indirectly (via cofactors) activate or repress transcription and can also disrupt binding sites by changing DNA conformation\textsuperscript{35}. A large number of genes are regulated by YY1 including physiological systems important to the metabolic syndrome, such as the nervous system\textsuperscript{37} and inflammatory system\textsuperscript{35}.\textsuperscript{37, 38} For example, YY1 regulates expression of p53\textsuperscript{35, 39}, a transcription factor and tumor suppressor that controls cell cycle progression and the cellular stress response. Adipose expression of p53 plays a role in inflammation and the development of insulin resistance\textsuperscript{40}. YY1 regulates the expression of other transcription factors (e.g., CREB, c-MYC, and SP1)\textsuperscript{35} which, by themselves, can regulate extensive physiological pathways. YY1 could modulate multiple metabolic syndrome traits (e.g., SBP, DBP, BMI, and plasma glucose) through a network of direct or indirect gene-by-gene or protein-protein (transcription factor) interactions.

Genetic variation at the \textit{YY1} locus in rats is associated with development of Type 1 diabetes, potentially through its actions on cytokine-related genes\textsuperscript{41}. In mice, \textit{YY1} gene expression significantly correlates with cardiovascular function, specifically the P-R wave interval of an electrocardiogram\textsuperscript{42}. In addition, two quantitative trait loci (QTLs) for body weight and urinary albumin excretion in the rat have been mapped to the \textit{YY1}-containing region of the human genome using stringently filtered cross-species alignments\textsuperscript{43}. Knockout mice expressing graded amounts of YY1 (75\%, 50\%, 25\%) displayed dose-dependent changes in sensitivity to apoptosis, a key component of target organ damage in hypertension\textsuperscript{44}. RNA interference strategies have been used to generate transgenic mouse lines that express reduced levels of YY1.
protein and exhibit decreased birth weight and size.

Functional genetic variation at the YY1 locus could thereby provide a unifying genetic mechanism for a portion of metabolic syndrome trait variation in the population.

**Conclusions and Perspectives**

We developed a novel sequential strategy for identification of candidate genes in the cardio-metabolic syndrome. The method integrated several disciplines and tools (animal models, microarray technology, bioinformatics, molecular biology, extreme phenotypes, human genetics) and suggested a previously unexplored gene, YY1, as a candidate and potential mechanistic link for several traits of the human metabolic syndrome. Thus, development of novel methods to identify candidate genes might prove advantageous in the quest to understand the genetic basis of complex traits.

**Funding Sources:** Funding was provided by the National Center for Research Resources, M01RR000827, the Comprehensive Research Center of Excellence in Minority Health and Health Disparities (CRCHD), MD00020, the NIH/NIDDK DK007671 Nephrology Training Grant, Department of Veterans Affairs, and National Institutes of Health (DK 60702 and P01 HL58120 to DTO'C).

**Conflict of Interest Disclosures:** None

**References:**


Table 1. Transcription factors predicted to bind the MITRA-identified motif. The CONSITE and P-MATCH algorithms identified a total of 7 transcription factors predicted to bind the MITRA-identified motif. Three transcription factors (RUNX1, SRY, and YY1) were identified by both algorithms. The JASPAR accession number for CONSITE predictions and the TRANSFAC site accession for P-MATCH predictions are listed. CONSITE and P-MATCH binding scores and the DNA strand on which the transcription factor was predicted to bind are also presented.

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>Accession number</th>
<th>CONSITE score</th>
<th>P-MATCH score</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOX9</td>
<td>MA0077</td>
<td>6.194</td>
<td>--</td>
<td>+</td>
</tr>
<tr>
<td>SRY</td>
<td>MA0084; R07263</td>
<td>2.695</td>
<td>0.755</td>
<td>+</td>
</tr>
<tr>
<td>YY1</td>
<td>MA0095; R05992</td>
<td>7.027</td>
<td>0.964</td>
<td>+</td>
</tr>
<tr>
<td>HOXA3</td>
<td>R07263</td>
<td>--</td>
<td>0.938</td>
<td>-</td>
</tr>
<tr>
<td>MEF2A</td>
<td>MA0052</td>
<td>5.567</td>
<td>--</td>
<td>-</td>
</tr>
<tr>
<td>RUNX1</td>
<td>MA0002; R07891</td>
<td>5.491</td>
<td>--</td>
<td>-</td>
</tr>
<tr>
<td>SOX17</td>
<td>MA0078</td>
<td>4.273</td>
<td>--</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2. Effect of candidate transcription factors on motif activity. Luciferase reporter assays were performed with cotransfection of cDNA expression plasmids for transcription factors computationally predicted to bind the motif. The effects of transcription factors on the motif in the promoter position in either the forward or reverse orientation were tested. Data are presented as normalized intensity (RLU/µg protein). The pGL3-Promoter vector, which contains the SV-40 promoter driving expression of the luciferase gene, served as the control luciferase expression vector. The control for the transcription factor expression plasmids was an empty vector containing only the CMV promoter. Data were analyzed with ANOVA followed by pairwise t-tests corrected for multiple comparisons. Statistical significance is indicated with bold text.

<table>
<thead>
<tr>
<th>Transcription factor</th>
<th>Control vector (RLU/µg protein)</th>
<th>Promoter-forward (RLU/µg protein)</th>
<th>Promoter-reverse (RLU/µg protein)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control: empty vector</td>
<td>1.00 ± 0.06</td>
<td>0.74 ± 0.04</td>
<td>0.68 ± 0.04</td>
</tr>
<tr>
<td>No effect on control vector</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HOXA3</td>
<td>0.92 ± 0.08</td>
<td>0.66 ± 0.02</td>
<td>0.84 ± 0.03#</td>
</tr>
<tr>
<td>SRY</td>
<td>0.94 ± 0.05</td>
<td>0.93 ± 0.04#</td>
<td>0.79 ± 0.04</td>
</tr>
<tr>
<td>YY1</td>
<td>0.87 ± 0.07</td>
<td>1.10 ± 0.06#</td>
<td>1.11 ± 0.08#</td>
</tr>
<tr>
<td>Effect on control vector</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEF2A</td>
<td>0.77 ± 0.06*</td>
<td>0.73 ± 0.04</td>
<td>0.62 ± 0.07</td>
</tr>
<tr>
<td>RUNX1</td>
<td>0.62 ± 0.03*</td>
<td>0.51 ± 0.09*</td>
<td>0.65 ± 0.06</td>
</tr>
<tr>
<td>SOX9</td>
<td>0.69 ± 0.09*</td>
<td>0.47 ± 0.02*</td>
<td>0.54 ± 0.05</td>
</tr>
<tr>
<td>SOX17</td>
<td>0.62 ± 0.02*</td>
<td>0.54 ± 0.02*</td>
<td>0.51 ± 0.03#</td>
</tr>
</tbody>
</table>

*p<0.05 vs. control vector/empty vector.
+p<0.05 vs. promoter-forward construct/empty vector.
#p<0.05 vs. promoter-reverse construct/empty vector.
Table 3. Initial (cohort-1; n=996) associations of candidate genes with hypertension and metabolic syndrome traits.
Statistical association between candidate gene tagging SNPs and the following metabolic syndrome traits was tested: systolic blood pressure (BP), diastolic BP, body mass index (BMI), plasma glucose, plasma total cholesterol, and plasma HDL cholesterol. Data were analyzed with univariate ANOVA using an additive model (independent variable = diploid genotype; dependent variable = metabolic syndrome trait; covariates = age, sex). Nominal p-values that satisfy a False Discovery Rate (FDR=0.05) correction for testing of multiple genotypes and phenotypes are indicated with **bold** text. The effect size [partial eta-squared ($\eta^2$): the proportion of total trait variance attributable to the specific SNP, expressed as % of trait variance] for significant SNPs is presented in parenthesis following the p-values.

<table>
<thead>
<tr>
<th>Gene</th>
<th>RefSNP ID</th>
<th>Systolic BP</th>
<th>Diastolic BP</th>
<th>Body mass index (BMI)</th>
<th>Plasma glucose</th>
<th>HDL cholesterol</th>
<th>Total cholesterol</th>
</tr>
</thead>
<tbody>
<tr>
<td>YY1</td>
<td>rs8021803</td>
<td>p=0.004 (1.3%)</td>
<td>p=0.002 (1.4%)</td>
<td>p=0.098</td>
<td>p=0.359</td>
<td>p=0.176</td>
<td>p=0.065</td>
</tr>
<tr>
<td>YY1</td>
<td>rs11625658</td>
<td>p=0.0002 (1.9%)</td>
<td>p=0.0006 (1.6%)</td>
<td>p=0.006 (1.2%)</td>
<td>p=0.024 (0.9%)</td>
<td>p=0.137</td>
<td>p=0.281</td>
</tr>
<tr>
<td>YY1</td>
<td>rs4905941</td>
<td>p=0.103</td>
<td>p=0.166</td>
<td>p=0.037</td>
<td>p=0.636</td>
<td>p=0.154</td>
<td>p=0.901</td>
</tr>
<tr>
<td>SRY</td>
<td>rs2058276</td>
<td>p=0.027 (1.5%)</td>
<td>p=0.018 (1.7%)</td>
<td>p=0.004 (2.3%)</td>
<td>p=0.079</td>
<td>p=0.191</td>
<td>p=0.149</td>
</tr>
<tr>
<td>SRY</td>
<td>rs1865680</td>
<td>p=0.299</td>
<td>p=0.371</td>
<td>p=0.113</td>
<td>p=0.651</td>
<td>p=0.272</td>
<td>p=0.733</td>
</tr>
<tr>
<td>HOXA3</td>
<td>rs10085570</td>
<td>p=0.050</td>
<td>p=0.017 (1.0%)</td>
<td>p=0.614</td>
<td>p=0.879</td>
<td>p=0.411</td>
<td>p=0.665</td>
</tr>
<tr>
<td>HOXA3</td>
<td>rs6948297</td>
<td>p=0.054</td>
<td>p=0.116</td>
<td>p=0.081</td>
<td>p=0.973</td>
<td>p=0.652</td>
<td>p=0.867</td>
</tr>
</tbody>
</table>
Table 4. Meta-analysis: Extension of initial genetic associations into additional population samples. Effect of a human YY1 tagging variant on cardio-metabolic syndrome traits. Meta-analysis was conducted in STATA, using fixed effect models. Individual regression models were additive. Effect sizes (from regression) are given as beta (or slope per allele), ± SE of beta. **Bold**: significant (p<0.05) effects.

<table>
<thead>
<tr>
<th>YY1 RefSNP</th>
<th>Group</th>
<th>Allele_1/Allele_2</th>
<th>Freq_A1</th>
<th>n</th>
<th>Trait</th>
<th>Model</th>
<th>Beta (slope, per allele)</th>
<th>SE of beta</th>
<th>P-value from meta-analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs11625658</td>
<td>Cohort-1</td>
<td>A/C</td>
<td>25.3%</td>
<td>996</td>
<td>DBP</td>
<td>Additive</td>
<td>2.180</td>
<td>1.106</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-2</td>
<td>A/C</td>
<td>26.2%</td>
<td>1075</td>
<td>DBP</td>
<td>Additive</td>
<td>0.140</td>
<td>0.070</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>ICBP</td>
<td>A/C</td>
<td>23.3%</td>
<td>6674</td>
<td>DBP</td>
<td>Additive</td>
<td>0.148</td>
<td>0.069</td>
<td>0.033</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Meta-analysis</td>
<td>A/C</td>
<td>-</td>
<td>68812</td>
<td>DBP</td>
<td>Additive</td>
<td>0.148</td>
<td>0.069</td>
<td>0.033</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-1</td>
<td>A/C</td>
<td>25.3%</td>
<td>996</td>
<td>SBP</td>
<td>Additive</td>
<td>2.560</td>
<td>1.373</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-2</td>
<td>A/C</td>
<td>26.2%</td>
<td>1075</td>
<td>SBP</td>
<td>Additive</td>
<td>0.140</td>
<td>0.070</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>ICBP</td>
<td>A/C</td>
<td>23.3%</td>
<td>6674</td>
<td>SBP</td>
<td>Additive</td>
<td>0.140</td>
<td>0.070</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Meta-analysis</td>
<td>A/C</td>
<td>-</td>
<td>2071</td>
<td>BMI</td>
<td>Additive</td>
<td>-0.703</td>
<td>0.242</td>
<td>0.004</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-1</td>
<td>A/C</td>
<td>25.3%</td>
<td>996</td>
<td>BMI</td>
<td>Additive</td>
<td>-1.217</td>
<td>0.405</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-2</td>
<td>A/C</td>
<td>26.2%</td>
<td>1075</td>
<td>BMI</td>
<td>Additive</td>
<td>-0.415</td>
<td>0.303</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Meta-analysis</td>
<td>A/C</td>
<td>-</td>
<td>2071</td>
<td>Glucose</td>
<td>Additive</td>
<td>2.776</td>
<td>1.069</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-1</td>
<td>A/C</td>
<td>25.3%</td>
<td>996</td>
<td>Glucose</td>
<td>Additive</td>
<td>0.311</td>
<td>1.129</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-2</td>
<td>A/C</td>
<td>26.2%</td>
<td>1075</td>
<td>Glucose</td>
<td>Additive</td>
<td>0.311</td>
<td>1.129</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Meta-analysis</td>
<td>A/C</td>
<td>-</td>
<td>2071</td>
<td>Glucose</td>
<td>Additive</td>
<td>1.661</td>
<td>0.802</td>
<td>0.038</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-1</td>
<td>A/C</td>
<td>25.3%</td>
<td>996</td>
<td>HDL-chol</td>
<td>Additive</td>
<td>-0.55</td>
<td>0.944</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-2</td>
<td>A/C</td>
<td>26.2%</td>
<td>1075</td>
<td>HDL-chol</td>
<td>Additive</td>
<td>0.589</td>
<td>0.821</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Meta-analysis</td>
<td>A/C</td>
<td>-</td>
<td>2071</td>
<td>HDL-chol</td>
<td>Additive</td>
<td>0.589</td>
<td>0.821</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-1</td>
<td>A/C</td>
<td>25.3%</td>
<td>996</td>
<td>Total-chol</td>
<td>Additive</td>
<td>1.28</td>
<td>2.015</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Cohort-2</td>
<td>A/C</td>
<td>26.2%</td>
<td>1075</td>
<td>Total-chol</td>
<td>Additive</td>
<td>-1.559</td>
<td>1.968</td>
<td>-</td>
</tr>
<tr>
<td>rs11625658</td>
<td>Meta-analysis</td>
<td>A/C</td>
<td>-</td>
<td>2071</td>
<td>Total-chol</td>
<td>Additive</td>
<td>-1.73</td>
<td>1.407</td>
<td>0.902</td>
</tr>
</tbody>
</table>

Figure Legends:

Figure 1. Computationally identified motif in the regulatory sequences of differentially expressed genes in the intermediary metabolism pathways of the BPH. MITRA identified a significantly overrepresented (p<0.05) consensus 10-bp motif in the set of regulatory sequences from differentially expressed genes in intermediary metabolism pathways of the BPH. The WebLogo consensus sequence consists of stacks of nucleotide symbols, one stack for each position in the sequence. The overall height of each stack indicates the sequence conservation at that position (measured in bits), whereas the height of symbols within the stack reflects the relative frequency of the corresponding nucleic acid at that position. Positions 8 and 10, which lack symbols, showed no nucleotide preference and can be represented as “N” in the consensus sequence: 5’-ACCATAGNTN-3’.

Figure 2. The motif confers transcriptional activity in the promoter position but not in the enhancer position. Orientation- and distance-specific transcriptional effects of the motif were determined using promoter/luciferase reporter assays. The motif was inserted into two regions of the pGL3-Promoter luciferase reporter plasmid: 1) in the “promoter” (Xhol) site just upstream of the SV-40 promoter; and 2) in the “enhancer” (Sall) site downstream of the luciferase gene. The motif was inserted in either the forward (5’-ACCATAGATA-3’) or reverse (5’-TATCTATGGT-3’) orientation. The control plasmid was the pGL3-Promoter vector (with eukaryotic transcription driven by the SV-40 early promoter) without a motif insert. The motif significantly altered transcriptional activity in the “promoter” position in both orientations, while showing no effect in the “enhancer” position in either orientation. Data were analyzed with ANOVA followed by pairwise t-tests corrected for multiple comparisons (Bonferroni). *p=0.0003 vs. Control. +p=0.0055 vs. Control.

Figure 3. Transcriptional activity of the motif was enhanced by the transcription factors HOXA3, SRY, and YY1. Promoter/luciferase reporter cotransfection experiments were performed with cDNA expression plasmids (pCMV) for the HOXA3, SRY, and YY1 transcription factors. The effects of these transcription factors on the motif in the promoter position in the forward or reverse orientations were tested. HOXA3, SRY, and YY1 had no
effect on the pGL3-Promoter control vector (SV-40 promoter; no motif insert). The control for the transcription factor expression plasmids was an empty vector containing only the CMV promoter (pcDNA-3.1). HOXA3 increased expression of the promoter-reverse construct by 1.24-fold (24%). SRY significantly increased expression of the promoter-forward construct by 1.26-fold (26%). YY1 significantly increased expression of the promoter-forward construct by 1.49-fold (49%) and the promoter-reverse construct by 1.63-fold (63%). Data are presented as normalized intensity (RLU/µg protein) and were analyzed with ANOVA followed by pairwise t-tests corrected for multiple comparisons. +p<0.05 vs. promoter-forward/empty vector. #p<0.05 vs. promoter-reverse/empty vector.

**Figure 4.** Transcriptional activity of the motif and the control vector were reduced by the transcription factors MEF2A, RUNX1, SOX9, and SOX17. Promoter/luciferase reporter cotransfection experiments were performed with cDNA expression plasmids (pCMV) for the MEF2A, RUNX1, SOX9, and SOX17. The effects of these transcription factors on the motif in the promoter position in the forward or reverse orientation were tested. MEF2A, RUNX1, SOX9, and SOX17 significantly reduced expression of the pGL3-Promoter control vector (SV-40 promoter; no motif insert). The control for the transcription factor expression plasmids was an empty vector containing only the CMV promoter (pcDNA-3.1). RUNX1 significantly reduced expression of the promoter-forward construct by 1.45-fold (31%). SOX9 significantly reduced expression of the promoter-forward construct by 1.57-fold (36%). SOX17 significantly reduced expression of the promoter-forward construct by 1.37-fold (27%) and the promoter-reverse construct by 1.33-fold (25%) (p<0.05). Data are presented as normalized intensity (RLU/µg protein) and were analyzed with ANOVA followed by pair-wise t-tests corrected for multiple comparisons. *p<0.05 vs. control vector/empty vector. +p<0.05 vs. promoter-forward/empty vector. #p<0.05 vs. promoter-reverse/empty vector.
Normalized intensity (RLU/ug protein)

Transcription factor

Control (empty vector) MEF2A RUNX1 SOX9 SOX17

- Control vector - no motif insert
- Motif - promoter position - forward orientation
- Motif - promoter position - reverse orientation

* p<0.05 vs. control vector/empty vector
+ p<0.05 vs. promoter-forward/empty vector
# p<0.05 vs. promoter-reverse/empty vector
Integrated Computational and Experimental Analysis of the Neuroendocrine Transcriptome in Genetic Hypertension Identifies Novel Control Points for the Cardio-Metabolic Syndrome

Ryan S. Friese, Chun Ye, Caroline M. Nievergelt, Andrew J. Schork, Nitish R. Mahapatra, Fangwen Rao, Philip S. Napolitan, Jill Waalen, Geert W. Schmid-Schönbein, Eleazar Eskin and Daniel T. O'Connor

*Circ Cardiovasc Genet.* published online June 5, 2012;
*Circulation: Cardiovascular Genetics* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2012 American Heart Association, Inc. All rights reserved.
Print ISSN: 1942-325X. Online ISSN: 1942-3268

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circgenetics.ahajournals.org/content/early/2012/06/05/CIRCGENETICS.111.962415

Data Supplement (unedited) at:
http://circgenetics.ahajournals.org/content/suppl/2012/06/05/CIRCGENETICS.111.962415.DC1

**Permissions:** Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Circulation: Cardiovascular Genetics* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

**Reprints:** Information about reprints can be found online at:
http://www.lww.com/reprints

**Subscriptions:** Information about subscribing to *Circulation: Cardiovascular Genetics* is online at:
http://circgenetics.ahajournals.org//subscriptions/
The Implications of Inheritance for Clinical Management: Supplemental Material

Srijita Sen-Chowdhry, MBBS, MD, FESC, Daniel Jacoby, MD, William J. McKenna, MD, DSc, FESC

**Estimating Heritability**

Quantifying heritability in a population therefore depends on the partitioning of observed phenotypic variation ($\sigma_p$) into unobserved genetic ($\sigma_g$) and environmental ($\sigma_i$) factors. The genetic variance can be further subdivided into the variance of additive genetic effects ($\sigma_A$), of dominance genetic effects (interactions between alleles at the same locus, $\sigma_D$), and of epistatic genetic effects (interactions between alleles at different loci, $\sigma_I$): $\sigma_G = \sigma_A + \sigma_D + \sigma_I$ (Figure 3). Broad sense heritability $H^2$ is the ratio of the total genetic variance to the total phenotypic variance ($\sigma_g / \sigma_p$), whereas narrow sense heritability $h^2$ is the ratio of the additive genetic variance to the total phenotypic variance ($\sigma_A / \sigma_p$).

![Illustrative pedigree from a large family with hypertension, a complex trait associated with multiple susceptibility loci.](image)

In this example, we hypothesise that four genes, A-D, control variation in systolic blood pressure in the family. The major allele of each gene is represented by the capital letter (A-D), while the lower case letter (a-d) denotes the minor allele, with lower frequency in the general population. We arbitrarily quantify the contribution of each allele to the variation in systolic blood pressure as follows: A, 1; a, 2; B, 1; b, 3; C, 2; c, 3; D, 1; d, 4.
The genotypes of the relatives in the box in Figure 3 are shown below (subscript m, p – maternal and paternal alleles respectively):

Father: Aa bb cc Dd
Mother: Aa BB Cc DD
Son (IV.9): a_m A_p b_m b_p C_m c_p D_m d_p
Son (IV.10): A_m a_p b_m b_p c_m c_p D_m d_p
Son (IV.11): A_m a_m b_m b_p c_m c_p D_m D_p
Daughter (IV.12): A_m A_p b_m b_p C_m c_p D_m D_p

Additive model
In an additive model, the combined effect of genetic alleles at two or more gene loci is equal to the sum of their individual effects. In the father, this amounts to 20 (i.e., 1+2+3+3+3+1+4). The combined genetic effect for the other relatives in Box-1 will be as follows: Mother, 12; Son (IV.9), 18; Son (IV.10), 18; Son (IV.11), 14; Daughter (IV.12), 14. We can further hypothesise that a total genetic effect exceeding 15 will be associated with clinical hypertension; consequently, the father and two older sons are hypertensive, while the remaining family members are normotensive.

Dominance model
Alleles that exhibit a dominant action will mask the contribution of recessive partners at the same locus. For simplicity, let us assume that the major allele at each locus is dominant over its minor partner. In this model, the combined genetic effect of Aa is 2 (i.e., 1+1), equivalent to the homozygous dominant genotype AA. The combined genetic effect for the relatives in Box-1 will be: Father, 16; Mother, 10; Son (IV.9), 12; Son (IV.10), 12; Son (IV.11), 10; Daughter, 10. If we further hypothesise that a total genetic effect exceeding 11 will be associated with clinical hypertension, then the father and two older sons will be affected while other family members are not, consistent with the status depicted on the pedigree.

Of note, dominance genetic effects have no influence on phenotypic resemblance between generations. Both the father and the two older sons have hypertension, for example, because of the shared a, b, c, and d alleles. Similarly, both the mother and the youngest son are normotensive because of the shared B, C, and D alleles. Since the children can inherit only one allele at each locus from each parent, interallelic interactions do not factor into intergenerational similarities. Dominance effects may, however, contribute to phenotypic resemblance between siblings. Both the youngest son and daughter are normotensive and share the BbCc genotype, with the B and C alleles dominant over their minor partners.

Epistasis model
Phenotypic expression of a gene may be modified, suppressed, or enhanced by genes at other loci in a phenomenon known as epistasis. As an example, we hypothesise the DD genotype has an inhibitory effect (-2) on the B/b locus. Assuming otherwise additive inheritance, the combined genetic effect for the relatives in Box-1 will be: Father, 20; Mother, 10; Son (IV.9), 18; Son (IV.10), 18; Son (IV.11), 12; Daughter, 12.
Empirical studies of genetic variance components show that additive variance typically accounts for over half, and frequently almost all, of the total genetic variance, even if non-additive effects are operating at the level of gene action \(^i\). For dominant genetic effects to contribute to intrafamilial phenotypic resemblance, the relatives must share both alleles at any particular locus, which is true only of full siblings; since parents transmit only a single copy of each gene to their children, interactions of this type cannot come into play between generations, where selection pressures are at work. Narrow sense heritability is therefore the more frequently cited and reproducible parameter \(^i\). Conventional heritability studies include simple functions of the regression of offspring on parental phenotypes, or correlation of offspring/ parental phenotypes, full/ half siblings, and monozygotic/ dizygotic twin pairs. A linear mixed model is generally optimal for unbalanced designs that incorporate continuous phenotypic measures from individuals with a mixture of relationships, within and across multiple generations \(^i\). One approach for binary traits, such as disease status, is to parameterise familial resemblance on a continuous liability scale so that the heritability is independent of disease prevalence \(^i\).

**“Missing Heritability” of Complex Traits**

Variants of low minor allele frequency (MAF), defined here as \(<0.5\)-5\%, or rare variants (MAF < 0.5\%) are not captured by current GWA genotyping arrays. Neither are their effect sizes large enough to allow consistent detection by classical linkage analysis in family studies \(^iv\). Direct sequencing may therefore be necessary to assess the contribution of infrequent genetic variants to complex traits. Mendelian disorders at the extremes of the phenotypic spectrum may provide candidate genes. Bartter’s and Gitelman’s syndromes, for example, are recessive Mendelian diseases associated with renal salt wasting and hypotension, with a respective prevalence of 1/ million and 1/ 40,000 in the general population. The heterozygous disease alleles have an estimated prevalence of \(<1\%. Screening of the Framingham Heart Study participants for variation in three of the genes implicated in these disorders (SLC12A3 (NCCT), SLC12A1 (NKCC2) and KCNJ1 (ROMK)) revealed a number of rare, heterozygous mutations, which were associated with clinically significant blood pressure reduction and protection from the development of hypertension \(^v\).

Rare sequence variants also contribute to variation in plasma HDL and triglyceride levels in the general population \(^vi\ \, vii\). It has been argued that genes subject to purifying selection may show a preponderance of rare independent mutations rather than common functional variants \(^v\). Mathematical modelling, however, suggests that if the susceptibility genotypes are rare (e.g., frequencies of \(<1\) in 5,000), then very large numbers of genes (n=183-556) are needed to explain 50\% of a common disease in the population, even with high individual risk ratios (RR= 10-20). In contrast, only 10-50 genes are required to explain 50\% of the burden of a common disease in the population if the predisposing genotypes are common (frequencies of 10-50\%), even with a risk ratio as low as 1.2 \(^vii\). Large scale whole genome sequencing studies are awaited to further define the role of infrequent variants, and establish whether they account for the residual heritability of complex traits.

In the interim, an algorithm for estimating the proportion of all SNPs in GWA studies for a quantitative trait has shed light on the source of the “missing heritability”. The combination of linear model analysis and simulations for validation was first applied to adult height. Although highly significant and well-replicated SNPs accounted for only \(<5\% of the variance in height, common SNPs in total explained a further 40\% \(^x\). Application of the same algorithm, modified for binary traits, to case control studies demonstrated that a substantial proportion of variation in susceptibility to Crohn’s disease and type-1 diabetes was tagged by common SNPs \(^x\). Much of the “missing heritability” was therefore not missing. Returning to height as an example, more than 80\% of the variation due to SNPs had gone undetected in published studies because the effects of the SNPs were too modest to reach the stringent level for genome-wide significance \(^ix\).
The results further suggested that the discrepancy between the 80% heritability of height and the 45% explained by SNPs was a corollary of incomplete linkage disequilibrium between contributory variants and SNPs. This in turn might be due to the residual contributory variants having a lower MAF, on average, than the SNPs typed on the array, although it was not necessary to invoke very low frequency variants (MAF < 0.001) to account for the outstanding heritability \(^{vii}\). More recently, a set of tools has been developed to estimate the total number of susceptibility loci and the distribution of their effect sizes for a trait based on discoveries from previous GWA studies. By integrating power over the predicted unidentified susceptibility loci, accounting for relative risk and allele frequency, the algorithms also estimate the expected number of discoveries for future GWA studies and the requisite sample sizes. The initial test cases were height, Crohn’s disease, and breast, prostate, and colorectal cancers; the results suggested that each trait was likely to harbour additional loci within the spectrum of low-penetrance common variants \(^{xi}\).