Biologie

Comparative genomics and genetic analysis of hypertensive end organ damage

(Vergleichende Genomik und genetische Analysen hypertensiver Endorganschäden)

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Biologie
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von

Anika Sietmann

aus Dülmen

2009

_____

Dekan: Prof. Dr. Christian Klämbt

Erster Gutachter: Prof. Dr. Thorsten Reusch

Zweite Gutachterin: Prof. Dr. Monika Stoll

Tag der mündlichen Prüfung: 02.03.2010

Tag der Promotion: 23.04.2010

Summary - Zusammenfassung

Die Erforschung der genetischen Grundlage von Volkskrankheiten wie Bluthochdruck und damit verbundene Endorganschäden ist eine der Herausforderungen unserer Zeit. Neben Umweltfaktoren ist die Entstehung der linksventrikulären (LV) Hypertrophie (LVH) als ein typischer hypertensiver Endorganschaden bedingt durch eine genetische Prädisposition. Die Entdeckung solcher Suszeptibilitätsfaktoren ist schwierig auf Grund des multifaktoriellen und komplexen Charakters der Erkrankung und erfordert eine geeignete Herangehensweise zur Analyse des genetisch determinierten Risikos. Um diesen Anforderungen gerecht zu werden wurde eine genomweite Assoziationsstudie als Fall-Kontroll-Studie in einem repräsentativen Patientenkollektiv mit Bluthochdruck und LVH durchgeführt. In einem zweiten Schritt wurde der Einfluss der resultierenden Nukleotidvarianten auf die LV Masse getestet. Die nun vorliegenden Kandidaten-Polymorphismen konnten nicht in einer Erweiterung des Ausgangs-kollektivs oder in zur Verfügung stehenden unabhängigen Patientenproben repliziert werden. Auf Grund dessen wurden in einem weiteren Schritt zur Verfügung stehende genetische Daten eines LVH Modells aus der Ratte mit Hilfe vergleichender Genomik in die Analyse integriert. Es konnten zwei Polymorphismen identifiziert werden die eine Assoziation mit bluthochdruckbedingter LVH im Menschen zeigen und im Rattengenom innerhalb einer Kopplungsregion für eine erhöhte linksventrikuläre Masse liegen. Die beiden Varianten ließen sich robust replizieren, sowohl in den erweiterten Proben der Ausgangskohorte als auch in unabhängigen Patientenproben. Auch wurde ein weiterer Polymorphismus in dem betroffenen Gen (PACS1) gefunden der sich im Kopplungsungleichgewicht zu den beiden oben beschriebenen Varianten befindet. PACS1, ein Membranprotein das in der Lokalisation des *trans*-Golgi-Netzwerkes eine Rolle spielt ist somit ein neues Kandidatengen für die Entstehung einer LVH als hypertensiver Endorganschaden und steht damit weiteren funktionellen Analysen zur Verfügung.

Table of contents

1. Introduction 1

1.1 Genetic dissection of complex traits 1

1.1.1 The variability and structure of the human genome 2

1.1.2 Family-based Linkage Analyses 3

1.1.3 Population-based Association Analyses 4

1.1.4 The Use of Comparative Genomics 6

1.2 Arterial Hypertension and hypertensive end organ damage 7

1.2.1 Left ventricular hypertrophy 7

1.3 Aim of the study 9

2. Material and Methods 10

2.1 Material 10

2.1.1 Consumables and Kits 10

2.1.2 Instrumentation 10

2.1.3 Software and internet resources 10

2.2 Methods 11

2.2.1 Genome-wide association study design 11

2.2.1.1 Sample recruitment and characterization 12

2.2.1.2 DNA preparation 14

2.2.1.3 Genome-wide SNP genotyping 15

2.2.1.4 TaqMan® SNP genotyping 17

2.2.1.5 PLINK data formats 19

2.2.2 Examination of linkage disequilibrium and haplotype blocks 20

2.2.2.1 Linkage Disequilibrium 20

2.2.2.2 Haplotype Estimation and tagging SNP selection 21

2.2.3 Linkage Analysis and QTL mapping 23

2.2.3.1 Parental strains and F_2-intercrossing 23

2.2.3.2 Rat Phenotyping 23

2.2.3.3 Genotyping of microsatellite markers 24

2.2.3.4 Linkage Analysis 24

2.2.3.5 Comparative Mapping 25

2.2.4 Statistics 26

2.2.4.1 Hardy-Weinberg Equilibrium 26

2.2.4.2 Basic allelic association testing 27

2.2.4.3 Logistic and linear Regression 29

2.2.4.4 Student´s *t*-test of a slope 31

2.2.4.5 Multiple-testing correction 31

2.2.4.6 Calculation of combined *p*-values 32

3. Results 33

3.1 A genome-wide association study of left ventricular hypertrophy as a hypertensive end organ damage 33

3.1.1 Quality control and SNP marker recovery 33

3.1.2 Genetic associations in the screening cohort 34

3.1.2.1 Basic allelic association 36

3.1.2.2 Age and sex adjustment of the screening cohort 40

3.1.2.3 Left ventricular mass as a quantitative trait 45

3.1.2.4 Replication of genome-wide candidate SNPs 50

3.2 Comparative mapping approach 53

3.2.1 QTL mapping 53

3.2.2 Comparative mapping of the QTL region 56

3.2.3 Replication of SNPs identified by comparative mapping 59

3.2.4 Fine mapping of PACS1 region 63

4. Discussion **71**

4.1 The genome-wide association study design 71

4.2 How to find associations in screening samples 74

4.3 Replication - a demanding task 76

4.4 To learn one´s lesson from comparative genomics 78

4.5 Outlook 81

5. Literature **83**

6. Appendix

A: Acknowledgements

B: List abbreviations

C: Publications

D: Curriculum Vitae

1. Introduction

1.1. Genetic dissection of complex traits

Most of the observed phenotypes in all organisms, including behavior or the susceptibility to common diseases are genetically complex traits, which are controlled by multiple genes. Monogenic or Mendelian traits are complementary to complex traits and, as far as the influence of the genetic background on the phenotype is concerned, most Mendelian traits presumably possess a multifactorial component. In addition, both complex and single-gene traits are affected by environmental factors also contributing to the examined phenotypes (Glazier AM 2002).

Common diseases such as arterial hypertension, diabetes, age-related macular degeneration or heart disease are classical complex traits of increasing public and scientific importance due to socioeconomic factors such as rising costs in the public health sector. These factors will become even more important since most common complex diseases are typical attributes of an aging society. However, the dissection of complex traits is a demanding challenge due to incomplete penetrance as well as phenotypic and genetic heterogeneity. Several well-established approaches have been applied to dissect genetic complex traits including family-based linkage analysis and population-based association studies (Lander E and Schork NJ 1994). Furthermore, the availability of a set of human genomes has rapidly increased the knowledge of human genetic variation, providing a powerful tool to dissect complex traits (Frazer KA 2009).

1.1.1. The variability and structure of the human genome

Considering the architecture of the human genome is an important step in understanding human diversity, which in turn may result in different disease susceptibilities. The availability of complete genome sequences rapidly increased the knowledge of several forms of human genetic variation, their evolutionary history and the correlation between them. Common human genetic variations or synonymous polymorphisms with a minor allele frequency of one percent in a distinct population can be divided into two different classes: single variations like single nucleotide polymorphisms (SNPs) or structural variations like copy number variations (CNVs). The human genome contains at least nine to ten million common SNPs, hence being the most prevalent variant (Frazer KA 2009).

Fortunately, it is not necessary to genotype all known SNPs to determine the disease influencing polymorphisms due to the block structure of the human genome. Genetic variants that are located close to each other tend to be inherited together. The association between alleles at linked loci is called linkage disequilibrium (LD). LD is structured in haplotype blocks, a particular combination of alleles along a chromosome. Haplotypes in the human genome are a result of recombination or mutation and the haplotype distribution is influenced by population specific factors such as genetic drift, natural selection or the number of population founding individuals (Collins A 2009). Therefore, the strength of allelic association and the occurrence of a particular haplotype differ between distinct populations. Utilizing this phenomenon, the International HapMap Consortium discovered haplotype blocks in four representative populations by genotyping about 3.1 million SNPs in phase I and II resulting in a SNP density of approximately one SNP per kilo base pairs. Overall, 30 trios of northern and western European ancestry living in Utah from the Centre d´Etude du Polymorphisme Humain (CEPH) collection were assayed called CEU samples (The International HapMap Consortium 2005; The International HapMap Consortium 2007). The resulting haplotype

information can be used to select representative tagging SNPs in order to identify genetic variation within haplotype blocks without the necessity to genotype every single SNP within this block. The approach provides an important basis for genome-wide genetic analysis.

1.1.2. Family-based Linkage Analyses

As mentioned above, the prevalence of common complex disease results from interactions between numerous environmental factors and genetic variation of many genes. The identification of the alleles affecting disease risk will help to better understand disease etiology. Genome-wide linkage analysis using polymorphic markers like microsatellite markers or SNPs spread across the genome is one of the most traditional method mapping disease genes (Hirschhorn JN and Daly MJ 2005).

The basic for linkage analysis was established early on fruit flies. It is the simplest form of genetic mapping measuring correlated segregation of Mendelian inherited markers for a trait (disease) of interest in families. Given meiotic recombination, cosegregating or linked markers are supposed to be located in close proximity to a disease influencing gene (Altshuler D 2008). In this way, disease associated genomic regions or quantitative trait loci (QTL) for a continuous phenotype are identified that are more likely to harbor a causal genetic variant contributing to the trait. Due to relatively broad (because of less marker resolution and limited number of meiotic breaks) regions the linkage approach requires further investigation such as positional cloning or candidate gene approaches.

While linkage analysis has been proven to be successful in mapping genes underlying mono- or oligogenic diseases like Huntington disease (Gusella JF 1983) the approach is limited in detecting genes underlying multifactorial complex traits.

Despite the identification of linkage regions for disorders like type I diabetes (Nisticò L 1996) a lot of human linkage studies in common disease were unsuccessful due to several factors including low total heritability, insufficient phenotyped families and insufficient power to detect common genetic variants with modest effects on disease. Another limitation of linkage mapping in humans is the recruitment of a sufficient number of affected families (Almasy L 2009). The use of an experimental population of animal models e.g. inbred mouse or rat strains counters these limitations due to a large number of family samples and larger pedigrees. Therefore linkage mapping in the animal model provides a more promising approach in dissecting the nature of common complex diseases.

1.1.3. Population-based Association Analyses

Another widely used method identifying genes underlying complex diseases is to compare allele or genotypes frequencies of a given variant, mainly SNPs between two groups. If a particular genetic variant is observed more often than expected by chance among these two groups the variant is called associated with the disease or trait of interest. In this case the variant serves as a marker and not as a mandatory causal relationship with the disease or trait. Most prevalent, the correlation between genetic variants and trait differences is assessed in affected case and unaffected comparison control samples on a population scale. These case control studies are relatively straightforward due to rapid and sufficient assembly of samples. Until now, only candidate gene association studies were performed in order to dissect common variants within genes of previously identified linkage regions or putative diseased pathways. However, association studies based on one or few candidate genes examine only a small fraction of the universe of sequence variation in each patient, disregarding the multifactorial nature of complex traits.

Moreover, the hypothesis driven approach leads to potential bias in selecting candidate genes (Hirschhorn JN and Daly MJ 2005).

In the recent past, the completion of the human genome project and the progress in the International HapMap Project as well as the development of cost-efficient high-throughput SNP genotyping platforms assaying hundreds of thousands of SNPs simultaneously have set the stage for genome-wide association studies (GWAS). GWAS are an important step beyond candidate gene studies because they allow for SNP marker queries of the entire genome at levels of high resolution unaffected by prior hypotheses. The expanding knowledge of the correlation among SNPs generated by the International HapMap Project provides the basis of performing genome-wide studies (Pearson TA 2008). GWAS are based on the common disease / common variant hypotheses suggesting that a distinct number of variants with allele frequencies of more than 1 % to 5 % in the observed population contribute to the more common forms of human diseases. These common disease variants targeted by GWAS have modest and variable phenotypic effects influenced by complex genetics and environmental factors (Collins A 2009; Manolio AT 2009). The number of GWAS is exponentially growing addressing several disease categories including metabolic, autoimmune, neurodegenerative disease and cancer. Numerous, approximately 300 novel genetic loci underlying disease susceptibility have been discovered for over 80 phenotypes e.g. HLA-DQAI or HBB in Crohn´s disease or IL23R in Psoriasis (Frazer KA 2009; Johnsons AD and O´Donell CJ 2009).

Despite the recent success of GWAS which use a case control approach, there are disadvantages regarding the recruitment of representative affected and unaffected samples and possible population stratification due to differences in background population. Because of these reasons and due to insufficient sample size or random errors associated SNP markers found in GWAS often exhibit a lack of reproducibility (Pearson TA 2008).

1.1.4. The Use of Comparative Genomics

As mentioned before, several strategies exist to dissect a complex trait. All of these approaches exhibit assets and drawbacks regarding their efficiency, effectiveness and practicability. Hence, comparative genomics is often a valuable tool to bridge between data available from human and animal model studies.

A considerable part of the human genome consists of conserved sequence regions and homologous DNA sections exhibiting sequence similarities to other species (orthologs) or within the same organism (paralogs). It is believed that common features of different organisms are often encoded by homologous DNA sequences. Comparing the genome sequences of different species such as human and rodents using sequence alignments algorithms is the major principle of comparative genomics. Especially for rodents, extensive collections of physiological and genetic data are available surveyed in numerous mouse and rat strains. In particular, the rat serves as an excellent model organism for common diseases like hypertension, providing a rich source for comparative genomic approaches. Furthermore, the overlapping examination of rodent orthologs and genes associated with human disease were enabled due to the availability of not only the human but also the mouse and rat genome sequences and accessible whole-genome aligned genomes (Hardison RC 2003).

Several promising comparative genomic studies were published in the recent past addressing the problem of multifactorial human disorders using available animal model data. In 2000 Stoll *et. al.* published a comparative genomic map targeting human hypertension loci based on QTL data from several experimental rat crosses in order to select new regions for genetic and functional studies (Stoll M 2000). Using a similar comparative genetics approach among mouse and human, known human atherosclerosis QTLs could be mapped to homologous QTLs in mice, identifying novel candidate genes for atherosclerosis (Wang X 2005).

1.2. Arterial Hypertension and hypertensive end organ damage

According to the current guidelines of the world health organization, arterial hypertension is defined as blood pressure values exceeding 140 mmHg for systolic blood pressure and 90 mmHg for diastolic blood pressure. Approximately 15 % to 20 % of the western population suffers from this serious disease (Classen M 2004). The arterial hypertension prevalence is projected at one billion cases worldwide (Kearney PM 2005). Above 90 % to 95 % of adult hypertension cases are essential or primary hypertension which are defined as a hypertension without apparent reasons like kidney damage (Carretero OA 2000). As a consequence of cardiovascular circulation system stress under continuous high blood pressure conditions, end organ damage develops as a hypertensive complication. These complications can affect various organs including the brain (e.g. cerebrovascular accident), the eye (e.g. hypertensive retinopathy), the kidney (e.g. hypertensive nephropathy) or the heart. In the following section, the role of left ventricular hypertrophy as typical cardiac end organ damage is discussed.

1.2.1. Left ventricular hypertrophy

The pathological enlargement of the myocardial wall of the left ventricle (LV) with or without LV chamber enlargement is called Left Ventricular Hypertrophy (LVH). Epidemiological studies identified LVH as an important and independent risk factor for several diseases like stroke, myocardial infarction, heart failure and cardiovascular death in high-risk patients and the general population (Baessler A 2006; SMART Study Group 2007). LVH is a common complex disease occurring in 16 % of white men and 21 % of white women and 33 % to 34 % of African Americans. The prevalence in hypertensives varies between 22 % and 60 % (Arnett DK 2004; Sharma P 2006).

In the past, the LV enlargement was merely understood as an adaptive hypertrophy due to increased biomechanical stress like high blood pressure. Besides hypertension, several additional risk factors for developing LVH like age, diabetes or obesity are known (Arnett DK 2004). Recent studies have shown, that the LV mass increase is indeed a compensatory process but the LVH progress and magnitude varies despite equal blood pressure levels across individuals and well-comparable environmental factors (Baessler A 2006). These findings suggest a heritable component in the development of LVH. Furthermore, a normal distribution of LV mass in the population was observed, providing evidence that the phenotype is a complex quantitative trait influenced by multiple genes (Arnett DK 2009).

Up to date, mainly candidate gene approaches were applied to dissect the genetic background of LVH in human. Several polymorphisms in a number of candidate genes in distinct pathways have been identified but not consistently replicated e.g. the angiotensin-converting enzyme, the guanine nucleotide-binding protein gene, the Ghrelin receptor gene region or the transforming growth factor β1 (Semplicini A 2001; Baessler A 2006; Xu HY 2009). In the recent past a first pilot case control genome-wide association study for LV mass in Caucasians, the HyperGen (Hypertension Genetic Epidemiology Network) Study was performed. 11 valid SNPs were identified and a intragenic SNP within KCNBI (rs756529) was suggested as a valid candidate gene for LVH development (Arnett DK 2009). In addition, a genome-wide linkage scan for LV mass and function in Caucasian and African-American participants of the HyperGen Study population identified a region on chromosome 7 linked to LV wall thickness in Whites and several LOD peaks in both ethnic groups giving evidence for putative genes involved in LVH (Arnett DK 2009; Tang W 2009).

1.3. Aim of the study

Finding appropriate preventive and therapeutic methods for common diseases including arterial hypertension or the susceptibility for subsequent hypertension end organ damages remains a difficult task. It requires the dissection of the underlying genetic mechanism and the associated molecular pathways. In this study, the relatively unexplored genetic basis for left ventricular hypertrophy (LVH) as a prevalent hypertensive end organ damage was studied in order to characterize genetic susceptibility loci that determine LV mass and the resulting and LV dysfunction.

LVH as a typical common complex disease requires appropriate methods to investigate the underlying genetic basis. Previously reported approaches unfortunately do not accommodate for the complex nature of LVH pathogenesis and progression. For this reason a genome-wide association study approach was chosen to shed light on the disease influencing variants in a representative, well phenotyped German population with arterial hypertension and heart disease. Subsequently, integration of available linkage data from a rat model combined with a comparative genomics approach was used to:

1) dissect the global mechanism of developing LVH under high blood pressure conditions and

2) to provide a proof-of-concept for the analysis of complex traits.

2. Materials and Methods

2.1. Materials

2.1.1. Consumables and Kits

The following consumables and Kits were used:

100x TE-solution and sodium hypochlorite solution (Sigma-Aldrich Chemie GmbH, Steinheim, Ger); TaqMan® SNP Genotyping Assay, TaqMan® Genotyping Master Mix, 384-Well reaction plate, MicroAmp® optical adhesive film (Applied Biosystems Inc., Foster City, USA); Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen GmbH, Karlsruhe, Ger)

2.1.2. Instrumentation

Tecan Genesis RSP 100 and GENios plate reader (Tecan Deutschland GmbH, Crailsheim, Ger); pipettes (Eppendorf AG, Hamburg, Ger); 7900HT Fast Real-Time PCR System (Applied Biosystems Inc., Foster City, USA)

2.1.3. Software and internet resources

BioMart Project (http://www.biomart.org/)

Ensembl Genome Browser (http://www.ensembl.org/)

HaploView (http://www.broadinstitute.org/haploview)

International HapMap Project (http://www.hapmap.org/)

Microsoft® Office 2007 (Access and Excel)

NCBI National Center for Biotechnology Information
(http://www.ncbi.nlm.nih.gov/)

NetAffyx™ Analysis Center (http://www.affymetrix.com/analysis/index.affx)

PLINK toolset v1.06 (http://pngu.mgh.harvard.edu/purcell/PLINK/)

Rat Genome Database (http://rgd.mcw.edu/)

SISA (Simple Interactive Statistical Analysis,
http://www.quantitativeskills.com/sisa/)

STATA 9.0 (StataCorp LP, Texas, USA)

SDS 2.3 (Applied Biosystems Inc, Foster City, USA)

2.2. Methods

2.2.1. Genome-wide association study design

The described genome-wide association study (GWAS) resulted from the National Genome Research Network (NGFN), a cooperating network of several clinical centers and researchers. The entire sample recruitment and collection took place within this network. After sample processing for genome-wide genotyping, the resulting data was made available for the statistical analysis performed in this thesis. The applied statistical methods (chapter 2.2.2) were implemented in two software packages: STATA and the open-source PLINK toolset (Purcell S 2007). Replication of candidate single nucleotide polymorphisms (SNPs) was performed using TaqMan® SNP genotyping and the LIFA facility for high through-put genotyping.

2.2.1.1. Sample recruitment and characterization

For the GWAS, two patients groups were recruited within the NGFN network: a group of screening samples consisting of affected patients (cases) and healthy controls for genome-wide SNP genotyping using array technology (see 2.2.1.3) and a group of replications samples, consisting of cases and controls for candidate SNP replication.

Screening cases and replication cases I (see Table 2.1) were selected within a survey of patients with arterial hypertension in cardiological rehabilitation hospitals, the ESTher (***E****ndorgan****s****chäden, **Th****erapie und V****er****lauf*) *Register* (for further information see http://www.esther-register.de/). The ESTher cohort was recruited in about 30 German rehabilitation hospitals, a total of 1,400 patients were collected over a period of six month. Inclusion criteria for the ESTher cohort were known or new diagnosed arterial hypertension and a rehabilitation program due to cardiovascular disease.

Arterial hypertension was defined as follows:

- Blood pressure after a 5 min rest period greater than 140 / 90 mmHg (systolic / diastolic)
- Mean blood pressure (over 24 h) greater than 130 / 80 mmHg
- Exceeding one of the following blood pressure values during exercise electrocardiogram (ECG):

Age	Blood pressure [mmHg] (75 watt)	Blood pressure [mmHg] (100 watt)
20 -50	185 / 100	200 / 100
51 – 60	195 / 105	210 / 105
> 60	205 / 110	220 / 110

Left ventricular mass (LVM) was determined using

$$LVM\,[g] = 0.8 * [1.04 * (IVSd + PWd + LVEDD)^3 - LVEDD^3] + 0.6\,g$$

(IVSd = end-diastolic interventricular septum thickness, PWd = left ventricular posterior wall diameter in diastole, LVEDD = left ventricular enddiastolic diameter, all parameters were determined using echocardiography). The normalization of LVM for height to the allometric power of 2.7 was used to define left ventricular hypertrophy (LVH). Partition values for LVH were 50 (male) or 47 (female) $g/m^{2.7}$ (de Simone G 2005). Replication samples (cases II) were recruited from the Department of Cardiology, Angiology and Pulmology at the Heidelberg University Hospital by the group of Dr. med. Norbert Frey. All cases II exhibit an arterial hypertension (definition see above) and a striking increase in size of the left ventricle during echocardiography. LVH status was determined as described.

Healthy controls for screening and replication were available within the NGFN network: PopGen controls were recruited in the region of Kiel and were included in the PopGen database (Krawczak M 2006). Replication of candidate SNPs was performed using KORA-gen F3 controls based on the KORA platform (Wichmann HE 2005). For detailed information see http://epi.gsf.de/kora-gen/seiten/kora500k_e.php. All participating patients and control samples were Caucasian. In Table 2.1, detailed information of population composition, size and available phenotype data is given.

Table 2.1: Overall, 3938 DNA samples were included in the replication study. Due to the case-control design of the study, screening and replication samples were divided into case and control samples. For each group sex ratio, age (with standard deviation, SD), left ventricular mass (LVM), left ventricular hypertrophy (LVH), arterial hypertension (a.h.) status, 24 h mean systolic blood pressure (s. BP) in mmHg, cardiac heart disease (CHD), body mass index (BMI), waist circumference (waist circ.) and diabetes status (Diabetes mellitus type I or II) are given (missing data is indicated by "/").

Cases and healthy controls of the whole GWAS (n = 3938)					
Characteristics	Screening samples (n = 969) Cases (n = 492)	Controls (n = 476)	Replication samples (n = 2969) Cases I (n = 855)	Cases II (n = 470)	Controls (n = 1644)
sex (female/male)	92 / 400	223 / 253	297 / 501	222 / 248	831 / 813
female [%]	18.69	46.85	37.22	47.47	50.55
age (SD) [years]	57.65 ± 10.12	39.59 ± 11.17	54.60 ± 19.18	68.03 ± 9.62	62.52 ± 10.09
LVM (SD) [g]	211.48 ± 74.23	/	257.51 ± 112.64	236.62 ± 45.22	/
LVH %	41.38	/	54.89	87.20	/
arterial hypertension	all	no	all	all	770 a.h. 867 no a.h.
24h mean s. BP [mm Hg]	124.41 ± 14.48	/	124.88 ± 14.64	/	/
CHD [%]	84.79	/	77.82	no	/
BMI (SD)	28.87 ± 4.80	/	26.56 ± 10.35	29.37 ± 6.35	27.95 ± 5.16
waist circ. (SD) [cm]	101.17 ± 11.70	/	102.21 ± 12.01	/	/
Diabetes (type I+II) [%]	26.42	/	28.07	/	/

2.2.1.2. DNA preparation

Blood lymphocytes DNA from cases and controls was isolated in the recruiting clinical center using standard techniques (QIAGEN). Quant-iT™ PicoGreen® dsDNA Assay Kit was used to quantify the isolated DNA. The Quant-iT™ PicoGreen® dsDNA reagent is a Hoechst 33258-based assay for sensitive fluorescent nucleic acid staining. The samples were suited at 480 nm and the fluorescence emission

intensity was measured at 520 nm using a spectrofluorometer. The required dilution series, fluorescence measurement and adjustment to the aspired DNA concentration using 0.1 % TE-solution was carried out by means of Tecan Genesis RSP 100 and GENios plate reader.

2.2.1.3. Genome-wide SNP genotyping

Genome-wide SNP genotyping of screening samples was performed by Affymetrix® Research Services Laboratory (Affymetrix Inc., San Francisco, USA) using the Affymetrix® Genome-Wide Human SNP Array 5.0. The SNP chip represents 500 568 SNPs, 440 794 SNPs are available using the current Affymetrix® Genotyping Console (BAT 2.0). The participating SNPs content was chosen randomly, exhibiting 65 % coverage of a standardized central European population (CEU) of the ENCODE regions of the International HapMap Project (Bickeböller H 2007).

In brief, 500 ng total genomic DNA was digested with Nsp I and Sty I restriction enzymes and all resulting fragments were ligated to adaptor sequences due to cohesive 4 bp overhangs. The fragments were used as PCR templates, after amplification the amplicons were purified using polystyrene beads. The purified amplicons were fragmented to <100 bp fragments using DNAse I and 3' biotinylated using a terminale deoxynucleotidyl transferase (see Figure 2.1). The labeled DNA was hybridized allele specific to 25-mer oligonucleotide probes located on the Genome-Wide Human SNP Array 5.0. Each of the two alleles of an SNP is represented by 10 to 14 oligonucleotides, in summary one probe set. The biotinylated targets were stained using streptavidin R-phycoerythrin (Matsuzaki H 2004).

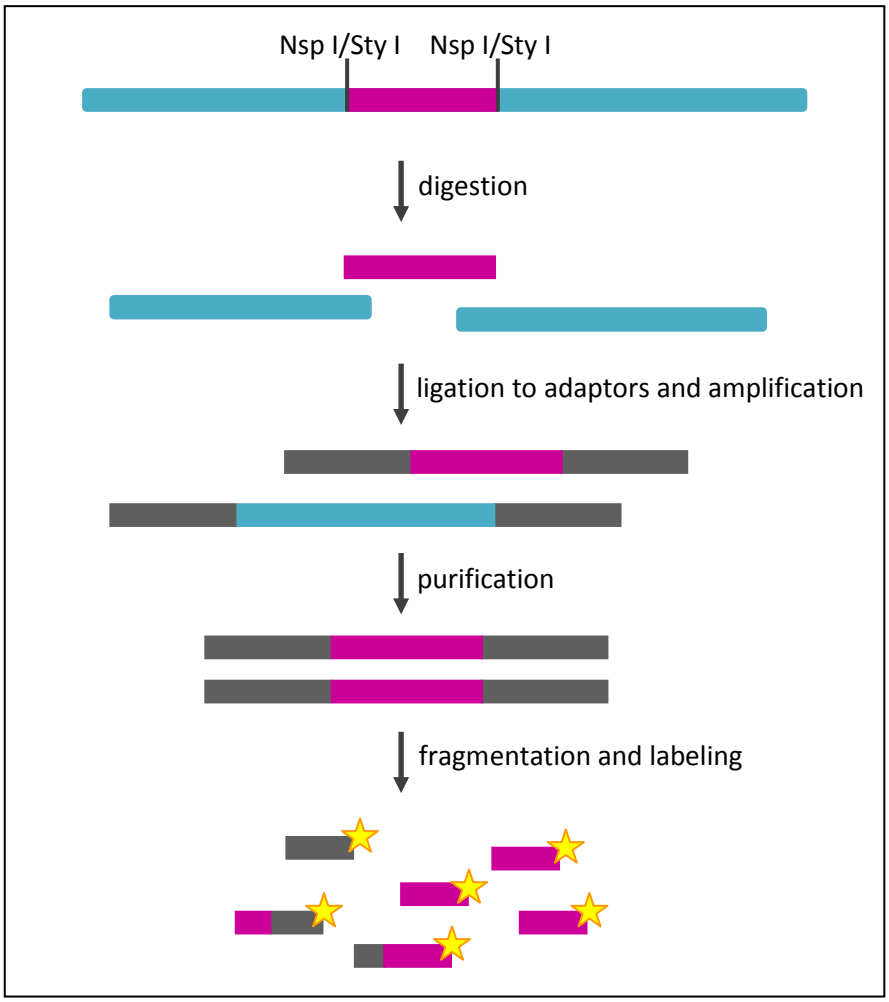


Figure 2.1: The principle of the Affymetrix Genome-Wide Human SNP Assay.

Hybridized, stained and washed arrays were scanned using GeneChip® Operating Software (GCOS) generating raw data (.CEL files). Genotype calling was performed using the updated version of the "Bayesian Robust Linear Model with Mahalanobis distance classifier" algorithm (BRLMM-P) and the genotype calls were extracted as text files for further analysis (Rabbee N 2006; Affymetrix 2007). Control samples for replication were genotyped using the GeneChip® Human Mapping 500K Array Set, an analog technology to the Affymetrix® Genome-Wide Human SNP Array 5.0 unless a separate Nsp I and Sty I digestion for two arrays. Genotypes have been determined using the software BRLMM version 1.4.0 with standard settings proposed by Affymetrix®.

2.2.1.4. TaqMan® SNP genotyping

Candidate SNPs were validated in replication case samples using TaqMan® technology in the form of ready-to-use TaqMan® SNP genotyping assays. Each assay contains sequence-specific forward and reverse primer to amplify the polymorphic sequence of interest and two TaqMan® minor groove binder (mgb) probes, one probe labeled with VIC® dye (detecting allele 1 sequence) and one labeled with FAM™ dye (detecting allele 2 sequence). The 5' exonuclease activity of the used TaqMan® Genotyping Master Mix containing DNA polymerase (AmpliTaq Gold) degrades the probe that has annealed to the template. In that case, the quencher and reporter dye lose their proximity allowing fluorescence of the particular reporter dye (see Figure 2.2).

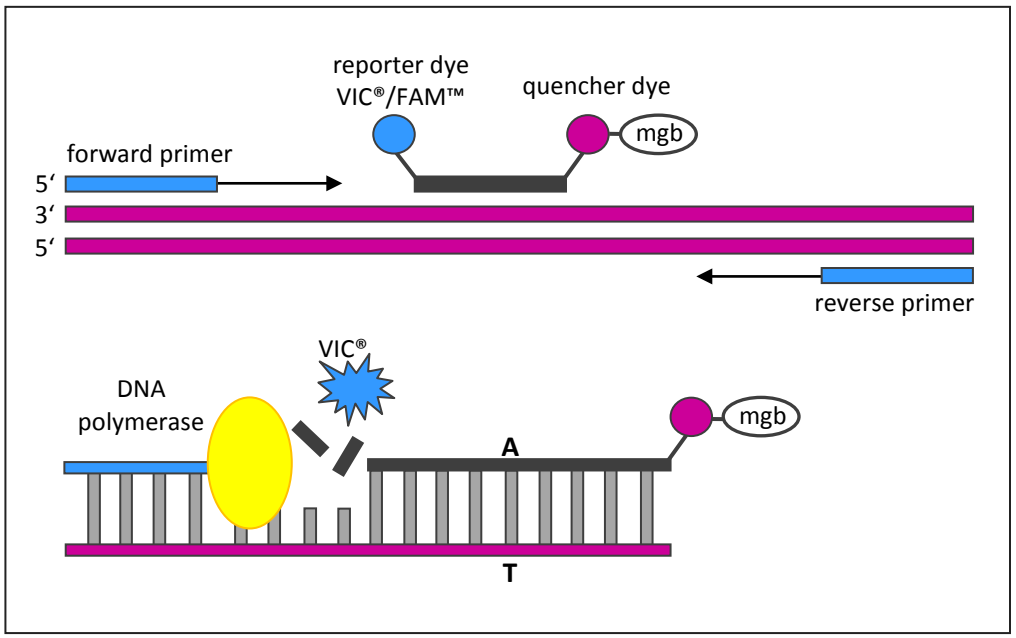


Figure 2.2: The principle of TaqMan® SNP genotyping. The determination of a T/G polymorphism is shown exemplary. In this case, the T allele is detected matching the VIC® dye labeled probe.

TaqMan® PCR processing was performed using the 7900HT Fast Real-Time PCR System under standard PCR conditions (95 °C /10 min denaturing; 40 cycles: 95 °C/15 s denaturing, 60 °C/1 min annealing and elongation). 2 ng genomic DNA served as PCR template in a final PCR reaction volume of 5 µL. After end-point fluorescence measurement, the DNA samples were called (allele discrimination, see Figure 2.3) and exported as text files for further analysis using SDS 2.3 software.

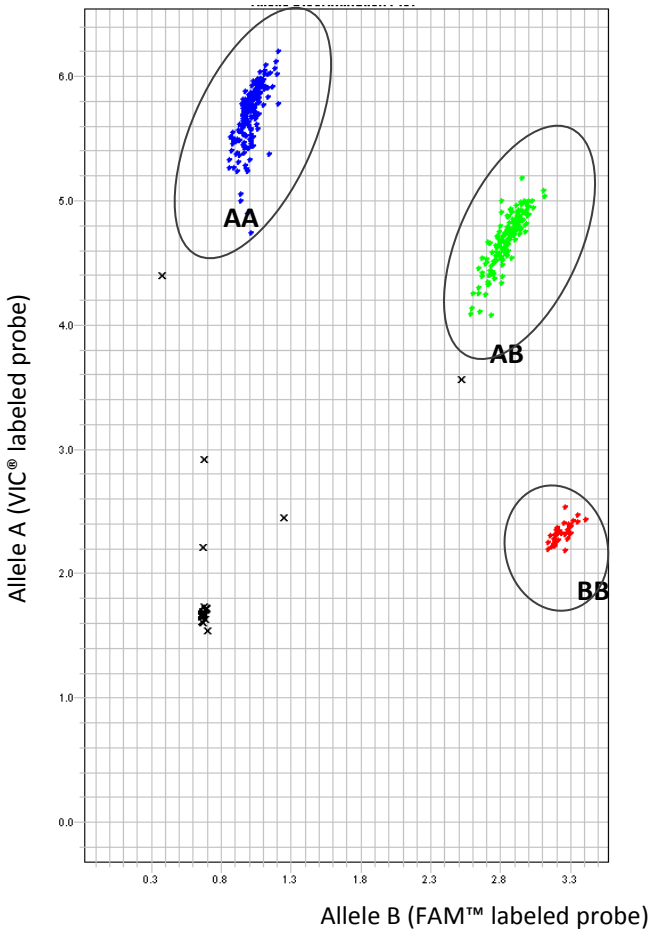


Figure 2.3: Allelic discrimination plot for an exemplary SNP. One of the three possible genotypes (homozygous AA or BB and heterozygous AB) is displayed for each sample.

2.2.1.5. PLINK data formats

BRLMM-p and BRLMM text files (genotype call code: 0 = AA, 1 = AB, 2 = BB, -1 = not called) and TaqMan® calling exports were recoded for usage in the PLINK toolset using Microsoft® Office 2007 Access and Excel respectively. Required allele information were available using the Affymetrix annotation file downloaded from www.affymetrix.com. SNP information was recoded in two basic PLINK data formats: PED and MAP files. For optional use, covariate files were generated containing the available phenotype data (see Table 2.2).

Table 2.2: Composition of the three applied PLINK data formats, all created as tab-separated text files and provided with the corresponding ending like *.ped or *.map.

PED file	family ID	not necessary, same as ind. ID
	individual ID	unique identifier for each sample
	paternal ID	not necessary, 0
	maternal ID	not necessary, 0
	sex	1 = male, 2 = female, -9 = missing data
	phenotype	1 = unaffected, 2 = affected, -9 = missing
	genotypes	2 characters per sample: 1, 2, 3, 4 (A, C, G, T)
MAP file	chromosome	1-22, X or Y
	rs# or SNP identifier	Affymetrix® probe set ID (SNP_A-…)
	genetic distance	not necessary, 0
	base-pair position	in bp units
covariate file	family ID	see PED file
	individual ID	see PED file
	covariate	binary or contonious, one per column

2.2.2. Examination of linkage disequilibrium and haplotype blocks

The dissection of candidate SNP regions took place exploiting the block structure of the human genome.

2.2.2.1. Linkage Disequilibrium

Genotype information of interesting SNP region for a Caucasian population were downloaded from the homepage of the International HapMap Project (The International HapMap Consortium 2007). Dumped genotype data (HapMap Data Rel. 24, Phase II, Nov08, on NCBI36 assembly, dbSNP b126) was analyzed using the HaploView software package (Barrett JC 2005). The calculated linkage disequilibrium (LD) pattern (see Figure 2.4) of the downloaded region was displayed using the standard color scheme. The color coding of the standard scheme is displayed in Table 2.3.

Table 2.3: Standard color scheme of the HaploView LD diagram. The normalized disequilibrium coefficient D' (disequilibrium coefficient D normalized for the maximum of Lewontin´s LD measure D_{max}) and the measure of confidence in the value of D' (log of the odds, LOD) yield in a color coding.

	D' < 1	D' = 1
LOD < 2	white	blue
LOD ≥ 2	shades of red	bright red

An example of a resulting LD pattern is given in Figure 2.4.

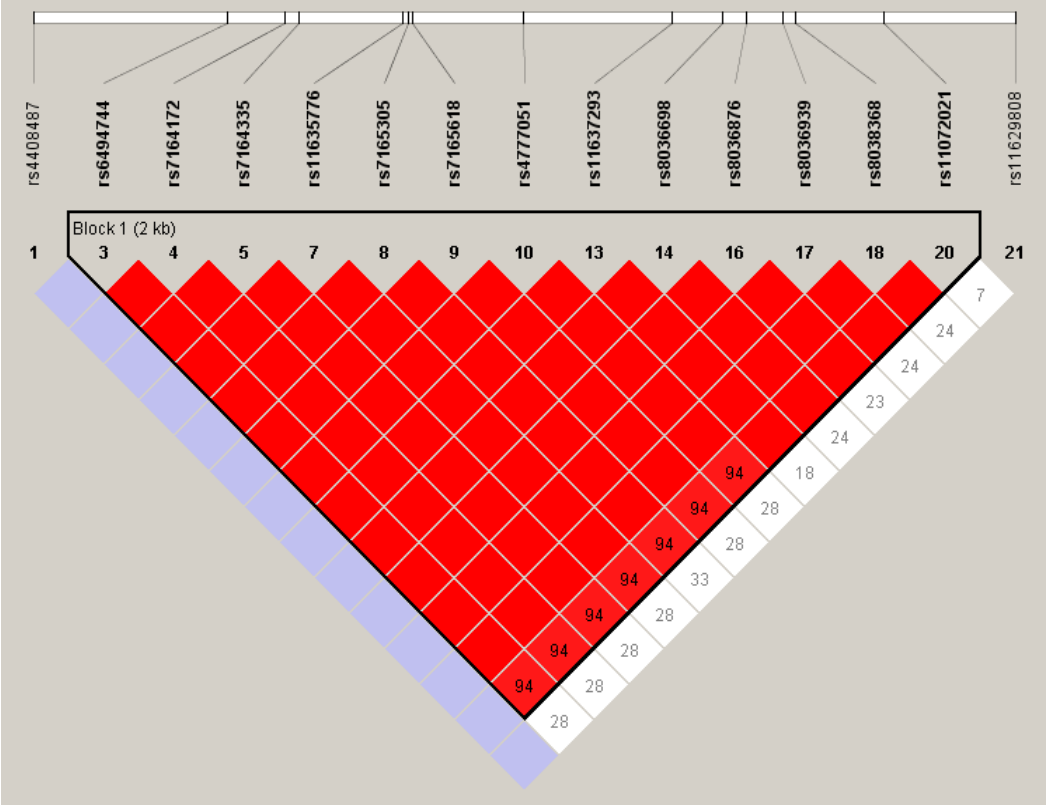


Figure 2.4: Linkage disequilibrium pattern of an exemplary region in the human genome downloaded from the International HapMap Consortium for a Caucasian population and displayed in HaploView. 15 polymorphic SNPs are located within this region. D' and LOD was calculated between each pair of SNPs and displayed as colored rhomb.

2.2.2.2. Haplotype Estimation and tagging SNP selection

For haplotype estimation, HaploView uses a two marker estimation-maximization (EM) algorithm (Barrett JC 2005). Haplotype blocks were selected manually with the aid of the color coding described above. In Figure 2.4, a manually chosen black framed block is shown. Figure 2.5 shows the resulting haplotype structure of this block.

Figure 2.5: Haplotype structure of the haplotype block of 13 SNPs estimated by HaploView using the downloaded genotype information of the International HapMap Consortium. The block is arranged in four possible haplotypes and the frequencies for each haplotype are given. Block dissolving tagging SNPs were identified by HaploView and are labeled using arrowheads.

The principle of choosing tagging SNPs is displayed in Figure 2.6. A sufficient number of SNPs tags the possible haplotypes in a chromosomal region.

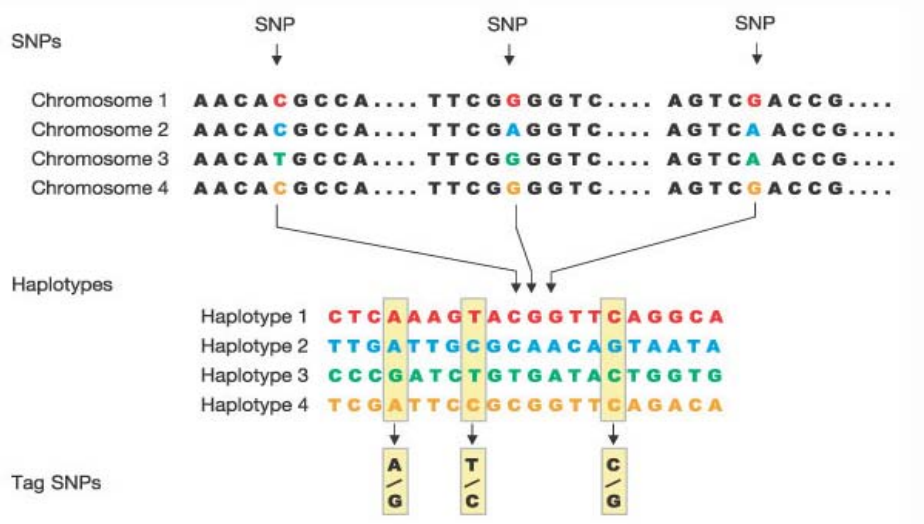


Figure 2.6: Four possible chromosomal regions are displayed and the resulting haplotype structure assuming linkage disequilibrium. The four haplotypes are tagged by genotyping three SNPs, the tagging SNPs (Figure: (The International HapMap Consortium 2003)).

2.2.3. Linkage Analysis and QTL mapping

All methods related to animal breeding, and genotyping outlined below took place in the group of Prof. Dr. med. Reinhold Kreutz, Institute of Clinical Pharmacology and Toxicology, CCM/CBF, Charité Universitätsmedizin Berlin. Statistical and bioinformatical analysis was performed as part of this thesis.

2.2.3.1. Parental strains and F_2-intercrossing

Normotensive parental inbred strain Fischer (F344) and spontaneously hypertensive rat, stroke prone (SHRSP) were obtained from existing colonies at the Charité Berlin, Prof. R. Kreutz, and were maintained under normal conditions. A F_2-intercross population (F344 x SHRSP) was generated and 232 male F_2-animals were included for linkage analysis. Salt loading was performed in all animals at the age of six weeks using a 4 % salt by weight diet for eight weeks.

2.2.3.2. Rat Phenotyping

At 15 weeks of age, systolic blood pressures were measured by non invasive tail-cuff method (McKee PA 1971). The rats were killed under ether anesthesia and the body weight and the weight of the left ventricle (including septum) were determined. For further analysis, left ventricle weight was normalized for body weight. Systolic blood pressure and the left ventricle / body weight (LVBW) phenotype were tested following a normal distribution within the F_2-population.

2.2.3.3. Genotyping of microsatellite markers

A complete genome scan using approximately 10 centiMorgan (cM) - spaced polymorphic microsatellite markers or simple sequence length polymorphisms (SSLP) was performed in two steps. First, 46 F_2-animals exhibiting extreme phenotypes were screened for genomic regions with significant results. Subsequent, these regions were genotyped using all 232 available F_2-animals. The required DNA was isolated using standard methods. The appropriate microsatellite markers were selected using the Rat Genome Database (RGD). PCR amplification of the microsatellite marker region was performed using 5´-radioactive labeled pairs of primer. By means of denaturating polyacrylamide (PAA) gel electrophoresis and autoradiography the amplicons were visualized.

2.2.3.4. Linkage Analysis

Using MAPMAKER/EXP linkage analysis was performed building up a genetic linkage map, QTL were identified by means of parametric linkage analysis using MAPMAKER/QTL 3.0b (Lander E 1987). Genetic distances were calculated considering recombination frequencies using the *Kosambi* algorithm. The threshold for significant linkage was defined as a LOD score of 4.3, suggestive 2.8 (Lander E and Kruglyak L 1995). The LOD score at a particular locus is defined as follows:

$$LOD = Log\left(\frac{likelihood\ for\ a\ present\ QTL}{likelihood\ for\ no\ present\ QTL}\right).$$

The determination of the 95 % confidence interval (CI) of significant QTL was based on the drop of one LOD unit from the peak (Rapp JP 2000).

2.2.3.5. Comparative mapping

Comparative mapping of the 95 % CI of the determined QTL was performed in our group using the Rat Genome Browser available on the Rat Genome Database. The required genomic positions of the appropriate microsatellite markers were received using the Rat Genome Database as well. The mapping approach was based on the rat genome assembly v3.4 and the human annotation is from the March 2006 (hg18) assembly. The UCSC (University of California Santa Cruz, http://genome.ucsc.edu/) derived alignment net track shows the best rat/human chain for the requested region in the rat genome. The chain track was determined using a gap scoring system.

2.2.4. Statistics

2.2.4.1. Hardy-Weinberg Equilibrium

The Hardy-Weinberg law describes the genetic equilibrium within an ideal population, an infinite random-mating population with constant proportions of homo- and heterozygote alleles by means of an algebraic equation.

Assuming two alleles A and a at one locus three genotypes are possible: AA, Aa and aa (see Table 2.4) for a diploid individual.

Table 2.4: Hardy-Weinberg law for two alleles at one single locus.

Parental gametic frequencies for alleles A and a	p(A)	q(a)
p(A)	AA	Aa
q(a)	Aa	aa

The equilibrium frequencies for allele A (p) and a (q) are given by

$$(p+q)^2 = p^2 + 2pq + q^2.$$

If the allele frequencies in a population are known they can be tested for statistically significant deviation, that is in principle the proportion of observed and expected allele frequencies. Using chi-squared testing (see 2.2.4.2) Hardy-Weinberg deviation is tested generally, for large-scale studies of SNP data Fisher´s exact testing is used. The probability of the appearance of heterozygous samples using allele frequencies is given by

$$probability[n_{12}|n, n_1] = \frac{2^{n_{12}}\, n!}{n_{11}!\; n_{12}!\; n_{22}!} * \frac{n_1!\; n_2!}{(2n)!}$$

where n_{11}, n_{12}, n_{22} are the number of observed genotypes AA, Aa and aa and n_1 the number of A alleles (Emigh TH 1980). Disproportions in Hardy-Weinberg equilibrium indicate genotyping problems or population structure or, in case samples a possible association between the examined phenotype and the observed marker. An efficient implementation of exact test for HWE as used in the PLINK software package is described by Wiggington et al. (Wiggington JE 2005).

2.2.4.2. Basic allelic association testing

The comparison of allele appearance in case and control samples may lead to new risk factors for the observed phenotype. In genome-wide association studies, for example, the allele status for a multitude of SNP loci is detected in cases and controls. Three genotypes at one biallelic (A and a) locus are possible where a,…,f is the number of observations:

	AA	Aa	aa	∑
case sample	a	c	e	a + c + e
control sample	b	d	f	b + d + f
∑	a + b	c + d	e + f	n

Alternatively, the number of alleles can be noted at each locus in the two analysis groups:

	A	a	∑
case sample	2a + c	2e + c	2a + 2c + 2e
control sample	2b + d	2f + d	2b + 2d + 2f
∑	2a + 2b + c + d	2e + 2f + c +d	2n

An allelic chi-square (χ^2) test as implemented in the PLINK software package with one degree of freedom was performed to compare the frequency distribution of the two alleles in cases and controls, hence identifying a possible association with the observed phenotype. Assuming the underlying null hypothesis of no association between the variables in the case and control group the chi-square test is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}.$$

By comparing the χ^2 -value to a chi-square distribution, a *p*-value was calculated for each chi-square statistic. A *p*-value ≤ 0.05 was consid ered to reject the null hypothesis, indicating an association between the SNP and the observed phenotype (Bickeböller H 2007). Missingness testing, implemented in the PLINK toolset as well, was performed similarly.

2.2.4.3. Logistic and linear regression

Regression models allow the description of the relation of two or more variables in form of an algebraic equation. The equation is fitted at the best to observed data by means of parameter estimation. In this case the examined phenotype (qualitative or quantitative) is the dependent variable Y, all further genetic and non-genetic variables are characterized as independent variables (covariables) X_i, i=1,…,n.

Linear Regression:

In the case of quantitative data as the dependent variable (i.e. left ventricular mass (LVM)), the relation between the dependent and independent variable (i.e. age) can be described by means of a simple linear regression fitting a straight line:

$$E(Y|X) = \hat{a} + \hat{b}X \, .$$

The y-intercept is given by the estimated parameter $\hat{a}$, the slope of the line by the estimated parameter $\hat{b}$ (also named regression coefficient). The dependent variable was checked to be a normal variable using the STATA software; otherwise it was transformed through an appropriate method to achieve a normal “Gauss” distribution.

Logistic Regression:

Logistic regression was applied to describe the correlation between a binary dependent variable (i.e. left ventricular hypertrophy (LVH) affected or unaffected samples) and an independent variable:

$$E(Y|X) = \frac{\exp(\hat{a} + \hat{b}X)}{1+\exp(\hat{a} + \hat{b}X)}.$$

Examine one dependent variable Y and multiple independent variables X_i, i=1,…,n is called multivariate linear (or analog logistic) regression:

$$E(Y|X) = \hat{a} + \hat{b_1}X_1 + \hat{b_2}X_2 + \cdots + \hat{b_n}X_n.$$

Modeling, i.e. the influence of a particular genotype (X_1) on LVM (linear regression) or LVH (logistic regression) considering the age (X_2) and sex (X_3) of the patients, the age and sex adjusted genotype influence on the dependent variable LVM or LVH is reflected in the regression coefficient $\hat{b_1}$ of X_1 or the odds ratio (OR, = $e^{\hat{b}_1}$) respectively (Bickeböller H 2007). In this case, using PLINK software package, the direction of the regression coefficient or OR represents the effect of the tested minor allele. For each calculation a model assuming full dominance or recessiveness was specified. For this purpose, the genotype variable was coded in a binary fashion:

dominant model → one or two minor alleles have an effect,

recessive model → only two minor alleles have an effect.

Subsequent, *t*-statistics or Wald statistics (see 2.2.4.4) was performed to determine the significance of the calculated regression coefficient or OR (Purcell S 2007).

2.2.4.4. Student´s *t*-test of slope

A *t*-test, that is a statistical hypothesis test following a Student´s *t*-distribution if the null hypothesis is true, is frequently used to prove if the slope of a regression line differs significantly from 0.

The *t*-value is given by

$$t = \frac{\hat{b} - b_0}{SE_{\hat{b}}}$$

where $\hat{b}$ is the calculated regression coefficient, b_0 the null hypothesis regression coefficient and $SE_{\hat{b}}$ the according standard error. By means of the *t*-value a *p*-value was calculated (Kohler U 2005).

2.2.4.5. Multiple-testing correction

Testing multiple independent null hypotheses in one sample, i.e. testing multiple SNP associations in one set of case and control samples, results in an inflation of type I errors (rejects the null hypotheses although it is true). For 20 independent statistic test carried out at the significance level α = 5%, the probability β of erroneously dropping one of the null hypotheses is given by

$$\beta = 1 - (1 - 0.05)^{20} = 0.6415.$$

To deal with the problem of multiple testing of large-scale testing in genome-wide association studies, correction of *p*-values was performed using the false discovery rate (FDR) method by Benjamini-Hochberg using the PLINK software package (Benjamini Y and Hochberg Y 1995).

2.2.4.6. Calculation of combined *p*-values

Unweighted combined *p*-values were calculated using the Fisher method. The approach combines *p*-values from a variety of independent test into one statistic having a χ^2 distribution using the formula:

$$\chi^2_{2k} = -2 \sum_{i=1}^{k} ln\ (p_i).$$

The resulting combined *p*-value can be calculated from a χ^2 table using 2k degrees of freedom where k is the number of the combined tests (Fisher RA 1948) available on the homepage of SISA (Simple Interactive Statistical Analysis).

3. Results

3.1. A genome-wide association study of left ventricular hypertrophy as a hypertensive end organ damage

The first part of this work was the dissection of a complex trait like the hypertensive end organ damage left ventricular hypertrophy (LVH) by means of a genome-wide association study (GWAS). A case control design was chosen in which allele frequencies in cases (ESTher samples) with the phenotype of interest were compared to those in a control group (PopGen samples). In a first step, the case and control screening samples were genotyped using the Affymetrix® SNP Array 5.0. A detailed description of the study samples is given in chapter 2.2.1.1.

3.1.1. Quality control and SNP marker recovery

Quality control of the used screening samples (492 cases, 476 controls) and the resulting genotype calls was a fundamental part of the GWAS analysis. Overall, 440 799 SNP markers were available after Genome-Wide Human SNP Array 5.0 processing. The data was recoded for use in the PLINK toolset (see 2.1.3) as described. Permutation testing was performed resulting in a genomic inflation factor (based on median χ^2) of 1.1615. Hence, population stratification was excluded. A minimum rate of 90 % successfully genotyped SNPs per sample was applied resulting in 5378 zeroed markers. In addition, a minor allele frequency threshold of 5 % was imposed to avoid genotype errors of rare and difficult to call SNPs thus eliminating 96 485 SNPs from the input data set. Furthermore, a Hardy-Weinberg Equilibrium (HWE) *p*-value of at least≥ 0.001 in control samples was selected as a cut-off value for the respective SNPs (1895 markers failed).

Disproportions in Hardy-Weinberg equilibrium may indicate genotyping problems or population stratification. A missingness test (χ^2 test) was performed to compare genotyping rates between cases and controls. A significant (p-value $\leq$ 0.05) deviation was considered to give evidence of differing sample quality and the involved SNPs were excluded from analysis. After data pruning, 310 417 SNPs with an average call rate of 99.19 % were available for following association analysis.

3.1.2. Genetic associations in the screening cohort

In a first step, the analysis was restricted to severe cases of LVH derived from the ESTher samples according to the guidelines of de Simone. Following the definition for LVH with partition values of 50 $g/m^{2.7}$ (male) or 47 (female) $g/m^{2.7}$ (see chapter 2.2.1.1), 204 LVH affected cases were selected from the pool of 492 genome-wide genotyped ESTher samples. A detailed description of sample phenotype composition is given in Table 3.1. 476 PopGen samples served as control samples.

Table 3.1: Phenotype characteristics of cases (ESTher cohort, n=204) and controls (PopGen, n= 476) for the first screening of genetic association using extreme cases. Mean age, left ventricular mass (LVM) and 24h mean systolic blood pressure (s. BP) are given with the respective standard deviation (SD). Missing data is indicated by "/".

Cases and controls of the screening step (n = 680)		
Characteristics	Cases (n = 204)	Controls (n = 476)
sex (female/male)	42 / 162	223 / 253
female [%]	20.59	46.85
age (SD) [years]	60.04 ± 10.52	39.59 ± 11.17
LVM (SD) [g]	274.10 ± 66.41	/
arterial hypertension	all	no
24h mean s. BP [mm Hg]	125.38 ± 15.92	/

All 204 screening cases exhibit arterial hypertension accompanied by a significant LVH as typical hypertensive end organ damage. There were no left ventricular mass (LVM) data available for the 476 PopGen control samples and hence no LVH status calculable. Blood pressure values were not collected as well but a normotensive status of control samples was affirmed by the PopGen database operators within the NGFN network. There were considerable differences in age between the case and the control group: having approximately equivalent standard deviations the mean difference was 20.81 years. The sex of the two analysis groups was yet another asymmetry of the underlying screening data. The percentage share of women in the affected samples was 20.59 %, in the unaffected control samples 46.85 %. Both, sex and age varieties between the two groups were considered in chapter 3.1.2.2 as far as a confounding of allelic association was concerned.

3.1.2.1. Basic allelic association

The analysis of a standard allelic association testing for LVH was performed using an allelic χ^2 test to compare the extreme cases of the screening cohort and the respective control samples described above. 310 417 SNP markers were available for this test after quality control and a *p*-value was calculated for every SNP reflecting the significance of association. Table 3.2 summarizes the distribution of the resulting *p*-values (unadjusted and adjusted for multiple testing, 310 417 tests) over different groups of minor allele frequencies.

Table 3.2: Overview of the *p*-values calculated by a standard allelic association testing between ESTher extreme cases (n = 204) and PopGen controls (n = 476) of the screening cohort. The minor allele frequencies (MAF) are given and the counts of SNPs with the respective *p*-values (unadjusted and adjusted for multiple testing (310 417 tests) using false discovery rate (FDR)).

	unadjusted *p*-value			**adjusted *p*-value (FDR)**		
MAF SNPs	$\leq 5*10^{-2}$	$\leq 5*10^{-4}$	$\leq 5*10^{-6}$	$\leq 5*10^{-2}$	$\leq 5*10^{-4}$	$\leq 5*10^{-6}$
> 5% - ≤10%	2 723	218	78	164	54	13
> 10% - ≤20%	5 038	163	29	87	12	/
> 20% - ≤30%	4 362	59	1	17	/	/
> 30% - ≤40%	3 923	40	/	9	/	/
> 40% - ≤50%	3 645	39	/	4	/	/
∑ (SNPs)	19 691	519	108	281	66	13

Most of the significantly associated SNPs were found in the group of markers having a minor allele frequency of 5 % to 10 %. By trend, the number of significant SNPs decreased with increasing minor allele frequencies. Correction for multiple testing to avoid false positive SNP associations reduced the amount of SNPs with high significance rapidly. This step was necessary because of 310 417 applied χ^2 tests to determine allelic association. Due to the small number of case samples in this

screening step, only markers having a minor allele frequency of 10 % or higher were considered to be reliable. The resulting unadjusted *p*-values (≤ 0.05) of the allelic association test for all markers having a minor allele frequency of > 10 % are displayed in Figure 3.1 by means of a Manhattan plot.

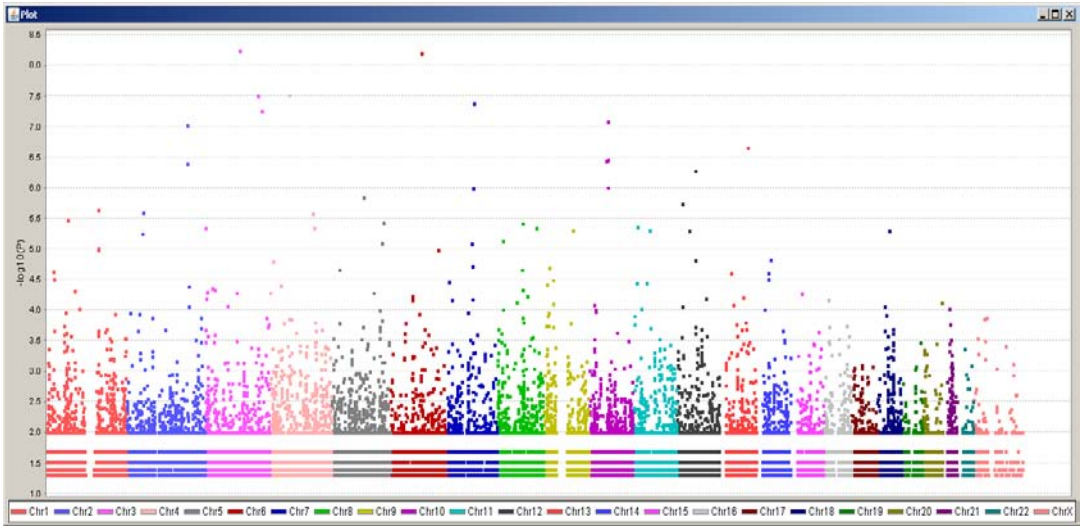


Figure 3.1: Manhattan plot of the significantly associated SNPs (n = 16 968, *p*-value unadjusted ≤ 0.05, minor allele frequency > 10 %) resulting from the first genome-wide scan of extreme LVH cases and PopGen control samples. The total base pair position for each chromosome (color coded) of each SNP is plotted against the negative decadic logarithm (-log) of the allelic association *p*-value.

A Manhattan plot is a good way to annotate genome-wide significance distribution. The disproportion of many markers with relatively low significance (-log (*p*-value) from 1.25 to 3) to some markers having a strong association (-log (*p*-value) greater than 5) is visible. Only 13 SNPs showed a genome-wide association of a *p*-value less than $5 * 10^{-6}$ (ranging between a *p*-value of $2.19 * 10^{-10}$ and $1.21 * 10^{-6}$) but exhibiting minor allele frequencies less than 10 % hence not being displayed in Figure 3.1. Overall, 36 SNPs exhibiting a minor allele frequency greater than 10 % and an unadjusted p-value less than $1 * 10^{-5}$ were selected for a closer examination in Table 3.3.

Table 3.3 (next page): 36 significantly associated SNP markers (adjusted *p*-value less than $6 * 10^{-3}$, unadjusted p-value less than $1 * 10^{-5}$) resulting from the genome-wide scan of LVH cases (extreme cases of ESTher samples, n = 204) and control samples (PopGen samples, n = 476) sorted by *p*-value. Each SNP is characterized by a unique identifier, the probe set ID resulting from the design on the SNP Array 5.0. The absolute position of each SNP (bp) is given and the respective chromosome. χ^2, *p*-value (unadjusted and adjusted using false discovery rate (FDR)) and the estimated odds ratio (OR) for the minor allele arose from the allelic association test with one degree of freedom. The minor (A1) and major (A2) allele of the respective SNP is given using a numeric code (1 = A, 2 = C, 3 = G, 4 = T). Genotype counts for the two alleles in the control sample group are shown for the possible two homozygous and one heterozygous allele combinations (A1A1/A1A2/A2A2). In addition, the overall determined minor allele frequency (MAF) is assigned.

probe set ID	chromsome	bp	χ^2	*p*-value	*p*-value (FDR)	MAF	A1	A2	OR	genotype (controls)
SNP_A-2153126	3	106747807	33.9	5.79E-09	1.53E-05	0.1752	3	1	2.372	11/102/344
SNP_A-2186482	6	97224557	33.73	6.33E-09	1.61E-05	0.1075	3	1	2.724	1/70/405
SNP_A-2065547	4	60243814	30.7	3.01E-08	6.03E-05	0.1401	3	2	2.381	10/81/385
SNP_A-2268977	3	163972160	30.62	3.14E-08	6.22E-05	0.1069	1	3	2.611	1/71/404
SNP_A-1849414	7	88165999	30.08	4.14E-08	7.71E-05	0.1127	3	2	2.539	2/74/400
SNP_A-2063755	3	175492037	29.54	5.48E-08	9.53E-05	0.1243	2	3	2.432	4/80/392
SNP_A-1950636	10	57665797	28.73	8.32E-08	1.29E-04	0.1237	1	3	2.409	3/82/391
SNP_A-1834522	2	188017098	28.74	9.50E-08	1.43E-04	0.1046	3	2	2.546	2/68/406
SNP_A-2057907	13	88080858	26.83	2.22E-07	2.86E-04	0.1237	3	4	2.343	8/73/395
SNP_A-2001186	10	57666414	25.98	3.44E-07	4.16E-04	0.1221	4	2	2.321	3/82/391
SNP_A-2274638	10	52984415	25.93	3.54E-07	4.25E-04	0.1391	1	2	2.284	5/87/360
SNP_A-1889369	2	188016885	25.7	3.98E-07	4.59E-04	0.1019	4	2	2.461	2/67/406
SNP_A-2064767	12	58459673	25.17	5.26E-07	5.73E-04	0.1156	4	2	2.334	5/73/398
SNP_A-4252877	10	57665722	23.95	9.86E-07	9.67E-04	0.1195	4	2	2.267	3/81/392
SNP_A-2157397	7	85169965	23.9	1.01E-06	9.83E-04	0.1541	4	3	2.104	3/111/362
SNP_A-2294633	5	100111301	23.25	1.42E-06	1.27E-03	0.1064	2	4	2.375	0/72/381
SNP_A-4218138	12	18258531	22.75	1.85E-06	1.55E-03	0.1059	2	1	2.305	3/70/403
SNP_A-1935686	1	160576937	22.37	2.25E-06	1.82E-03	0.2673	2	4	0.4966	46/193/229
SNP_A-2101014	2	52831698	22.13	2.55E-06	2.00E-03	0.1269	1	3	2.166	2/90/382
SNP_A-2074123	4	132182895	22.09	2.60E-06	2.02E-03	0.1125	1	3	2.235	3/76/397
SNP_A-1835958	1	67114572	21.65	3.28E-06	2.48E-03	0.1159	1	3	2.243	7/65/375
SNP_A-4265373	5	160998740	21.45	3.64E-06	2.71E-03	0.1637	4	2	1.997	11/105/360
SNP_A-2158385	8	78611923	21.32	3.88E-06	2.82E-03	0.1252	3	1	2.143	3/87/384
SNP_A-4260091	11	14680288	21.11	4.34E-06	3.11E-03	0.1047	4	2	2.251	4/68/404
SNP_A-2107642	4	138261700	21.05	4.48E-06	3.19E-03	0.1108	3	4	2.21	5/71/399
SNP_A-1917004	3	2329035	21.02	4.55E-06	3.23E-03	0.1024	2	4	2.262	2/70/404
SNP_A-2048869	8	122183583	21.01	4.57E-06	3.24E-03	0.1049	4	2	2.26	2/71/396
SNP_A-2011027	11	51285867	20.86	4.93E-06	3.43E-03	0.1105	3	1	2.202	4/73/399
SNP_A-1931236	9	90569830	20.87	4.93E-06	3.43E-03	0.1554	4	1	2.003	8/104/364
SNP_A-2291035	12	39000599	20.85	4.97E-06	3.45E-03	0.1625	2	4	1.983	10/106/359
SNP_A-2140243	18	38926416	20.8	5.10E-06	3.52E-03	0.1854	3	4	1.933	12/120/336
SNP_A-1858693	2	50341732	20.61	5.63E-06	3.82E-03	0.1495	3	2	2.014	5/105/366
SNP_A-4295370	8	19576480	20.06	7.51E-06	4.76E-03	0.3905	1	4	0.5683	87/233/154
SNP_A-1889406	5	155275310	19.93	8.05E-06	5.03E-03	0.1286	3	1	2.133	10/69/361
SNP_A-2196855	7	80634017	19.92	8.08E-06	5.04E-03	0.1006	1	3	2.233	2/69/404
SNP_A-1969214	1	160571784	19.57	9.69E-06	5.84E-03	0.2728	1	3	0.5316	45/203/228

In Table 3.3, the 36 most significant SNPs were documented having a minor allele frequency greater than 10 %. The emerging SNPs were located on the chromosomes 1 to 13 and 18. The odds ratio (OR) for each SNP association is given: the OR of 33 markers was greater than 1 (in the range of 1.933 to 2.724, mean 2.27 ± 0.1847 standard deviation) identifying the minor allele for these SNPs as risk alleles for developing LVH as a hypertensive end organ damage. For three SNPs (SNP_A-1935686, SNP_A-4295370 and SNP_A-1969214) the OR was smaller than 1 and consequently, the minor allele seemed to act as a protective genetic component in disease manifestation. Furthermore, the genotype distribution in the control group was considered. Three allele combinations were possible, homozygous for the minor or major allele and the heterozygous case. 27 out of 36 significantly associated SNPs appeared having less than 10 homozygous samples for the respective minor allele. These SNPs were considered to have low information content and were not qualified for advanced analysis. The remaining nine SNPs (SNP_A-1969214, SNP_A-1889406, SNP_A-4295370, SNP_A-2140243, SNP_A-2291035, SNP_A-4265373, SNP_A-1935686, SNP_A-2153126 and SNP_A-2065547) were possible candidate SNPs associated with the development of LVH as hypertensive end organ damage and were regarded in detail in the following analysis in chapter 3.1.2.3.

3.1.2.2. Age and sex adjustment of the screening cohort

As shown in Table 3.1, considerable differences in age and sex proportion between the case and the control group were present in the screening samples. Using logistic regression and subsequently *t*-statistic to calculate *p*-values, the influence of the detected allelic status and further so called covariates (age and sex) on the phenotype (affected and unaffected samples) were tested. In this way, the required age and sex corrected SNP associations were calculated. Assuming full dominance (or recessiveness) for the minor allele two models were specified and applied on the same dataset as the basic allelic association. The dataset consisted of 204 cases

having a clear LVH and 476 control samples. Quality filtering was performed as described in chapter 3.1.1 and a set of 310 417 SNPs were available for age and sex adjustment. The Manhattan Plots in Figure 3.2 and 3.3 and Table 3.4 and 3.5 show the association results.

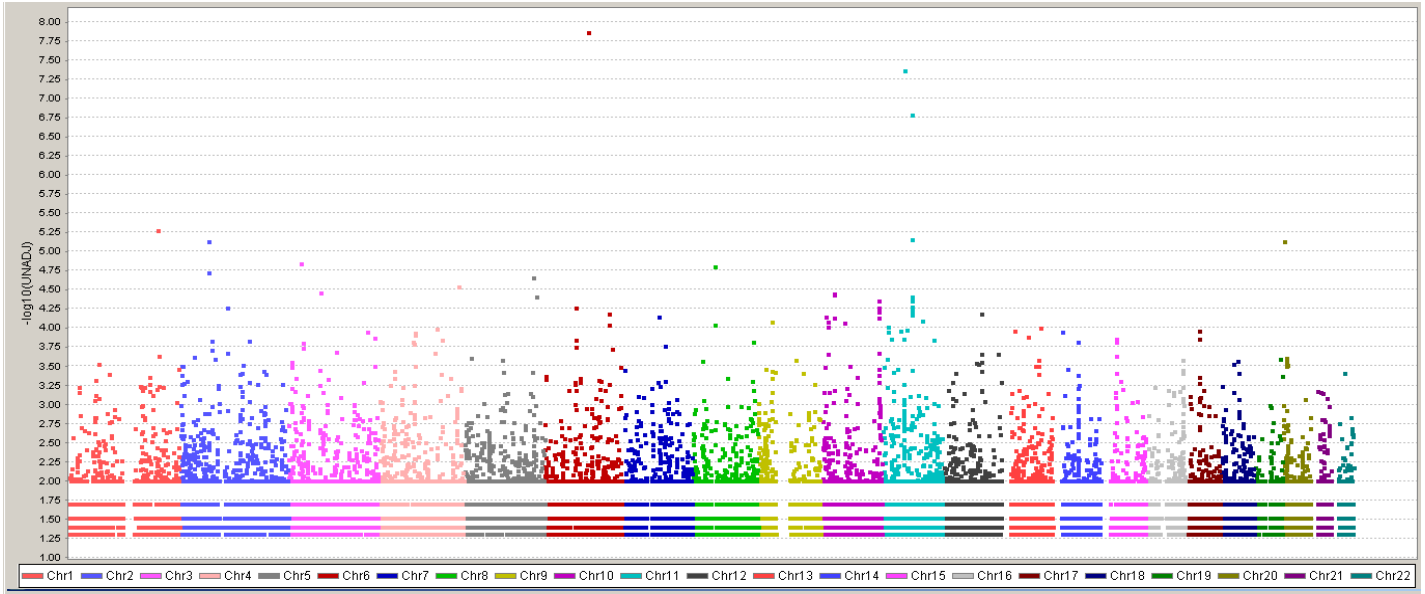


Figure 3.2: Manhattan Plot of the age and sex corrected genome-wide marker association using logistic regression, adjusted for age and sex and assuming full dominance of the minor allele. The –log (*p*-value) of all significant SNP (unadjusted *p*-value ≤ 0.05, n = 18 106) are plotted against the chromosomal position.

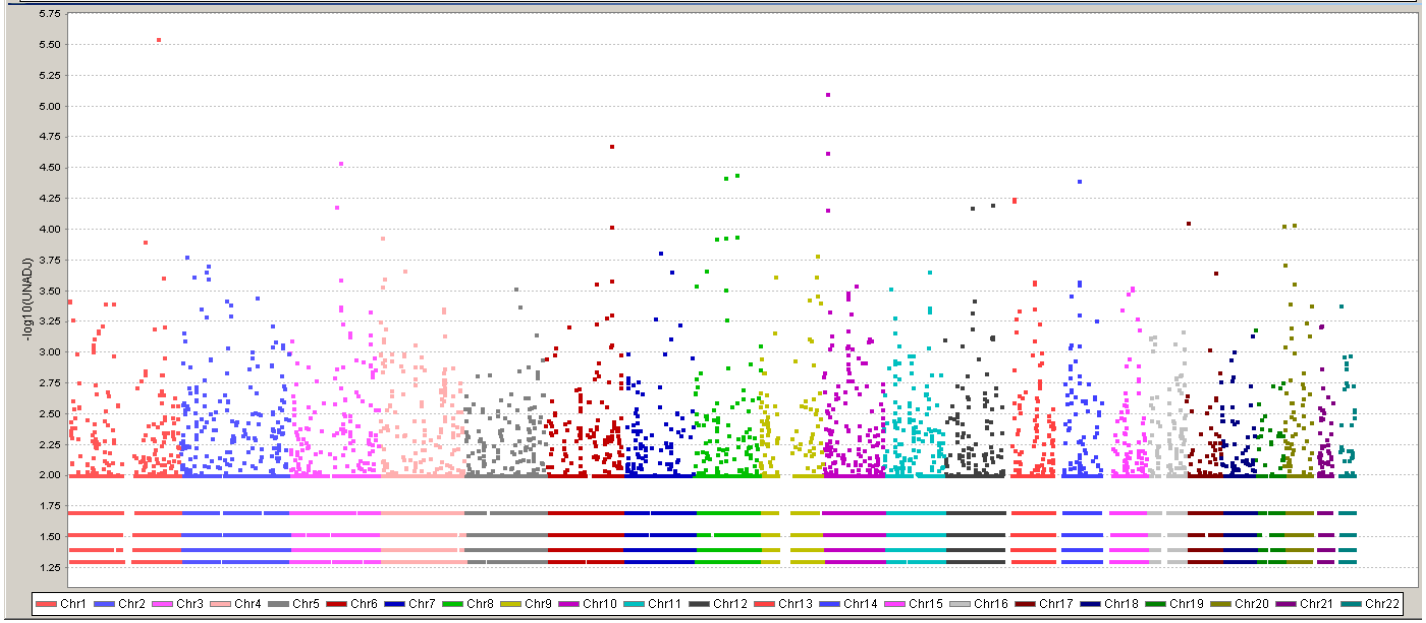


Figure 3.3: Manhattan Plot of the age and sex corrected genome-wide marker association using logistic regression, adjusted for age and sex and assuming full recessiveness of the minor allele. The –log (*p*-value) of all significant SNP (unadjusted *p*-value ≤ 0.05, n = 15 304) are plotted against the chromosomal position.

probe set ID	chromosome	bp	*p*-value	*p*-value (FDR)	MAF	A1	A2	OR	genotype (controls)
SNP_A-2186482	6	97224557	1.37E-08	7.48E-04	0.1075	3	1	5.768	1/70/405
SNP_A-2011027	11	51285867	4.21E-08	1.41E-03	0.1105	3	1	5.306	4/73/399
SNP_A-2064936	11	65741744	1.63E-07	2.95E-03	0.2574	4	2	4.096	24/176/276
SNP_A-2270325	1	199804108	5.24E-06	9.44E-03	0.2912	3	1	0.3011	43/200/233
SNP_A-2079405	11	65711257	6.95E-06	1.08E-02	0.2	2	1	3.311	19/137/320
SNP_A-2086003	2	64251922	7.39E-06	1.12E-02	0.1295	3	1	3.761	9/86/375
SNP_A-2069616	20	2676570	7.41E-06	1.12E-02	0.1213	2	4	3.894	7/86/382

Table 3.4: Results of genome-wide association using logistic regression (correction for age and sex), assuming full dominance of the minor allele (A1). The localization of the respective SNP markers (identified by the probe set ID) is given by the chromosome and the absolute base pair (bp) position. The unadjusted and adjusted (FDR, false discovery rate) *p*-value resulting from logistic regression and subsequently *t*-statistic, the minor allele frequency, the odds ratio (OR) and the genotype appearance in the control samples are given. All SNPs showing an unadjusted *p*-value $\leq 1 * 10^{-5}$ are listed.

probe set ID	chromosome	bp	*p*-value	*p*-value (FDR)	MAF	A1	A2	OR	genotype (controls)
SNP_A-2060349	1	198896407	2.82E-06	1.02E-01	0.1632	4	2	33.08	8/118/350
SNP_A-2262176	10	9662849	7.93E-06	1.81E-01	0.4919	2	4	3.712	105/241/129

Table 3.5: Results of genome-wide association results using logistic regression, assuming full recessiveness of the minor allele (A1). The given characteristics are comparable to those in Table 3.4. The SNP associations are corrected for age and sex. All SNPs showing an unadjusted *p*-value $\leq 1 * 10^{-5}$ are listed.

Figure 3.2 and Figure 3.3 display all significant SNP associations (unadjusted *p*-value ≤ 0.05) of the genome-wide scan between the LVH affected samples of the ESTher cohort (n = 204) and the PopGen control samples (n = 476) corrected for age and sex of the respective samples for chromosome 1 to 22. The sex chromosomes were not tested because the regression method uses genotype data in contrast to the allelic test. The resulting *p*-values assuming full dominance of the minor allele are displayed in Figure 3.2 by means of a Manhattan plot. 18 106 SNPs showed a *p*-value ≤ 0.05 and most of the significant associations lay within a range from 0.05 to 0.0005 (range of negative logarithm from approximately 1.3 to 3.5). The number of associated makers decreased rapidly by increasing significance. Only seven associated SNPs with an unadjusted *p*-value ≤ $1 * 10^{-5}$ after age and sex correction were detected, shown in Table 3.4. The seven SNPs are located on chromosome 1, 2, 6, 11 and 20 and after adjustment for multiple testing using false discovery rate, all SNPs kept a minimum *p*-value ≤ of 0.05. Except for one allele with a protective effect on developing a LVH under high blood pressure (minor allele of SNP_A-2270325), all of the minor variants increased the risk for the manifestation of LVH. For the calculation of the dominant model during logistic regression an effect of one or two minor alleles is assumed. Hence, the genotype distribution of the homozygous minor allele of the controls was negligible. All SNPs are possible candidates and were regarded in detail in the following analysis in chapter 3.1.2.3. In Figure 3.3, the Manhattan plot for the recessive model is displayed. As described for Figure 3.2, the number of SNP markers decreased with increasing significance level of association. In general, fewer SNPs were significantly associated with the examined phenotype in the recessive model (n = 15 304) and *p*-values were smaller than the corrected *p*-values in the dominant model. Only two markers were associated under recessive conditions exhibiting a *p*-value ≤ $1 * 10^{-5}$ (see Table 3.5). Both *p*-values were not significant (threshold ≤ 0.05) after correction for multiple testing. The OR were greater than 1, thus the respective minor allele of the two SNPs is a risk allele for developing LVH. Assuming full recessiveness (only two minor alleles have an effect on the examined phenotype), the genotype distribution of

SNP_A-2060349 was not significant because of only having eight homozygous samples in the controls samples for the minor allele and the marker was rejected for further analysis.

Two SNPs were associated significantly (unadjusted *p*-value ≤ $1 * 10^{-5}$) both in basic allelic association testing (see Table 3.3) and after logistic regression corrected for age and sex (dominant model, see Table 3.4): SNP_A-2011027 (chromosome 11) and SNP_A-2186482 (chromosome 6). The significance level of SNP_A-2011027 was raised due to age and sex correction from *p*-value $4.93 * 10^{-6}$ to $1.374 * 10^{-8}$ and the respective OR was 5.768 after logistic regression (OR 2.202 after allelic association testing). The significance level of SNP_A-2186482 acted reversely, before correction for age and sex the *p*-value was $6.33 * 10^{-9}$ and afterwards $4.21 * 10^{-8}$. However, the OR was increased from 2.724 to 5.306.

3.1.2.3. Left ventricular mass as a quantitative trait

Up to date, genetic association screening was based on a case control approach, typing affected cases having a distinct LVH under high blood pressure conditions (n = 204) against healthy control samples (n = 476). The LVH status was coded binary for cases and controls, whereas the dimension of the left ventricle related to the body height has to exceed a distinct threshold. Another approach for finding SNPs that influence the examined phenotype is to look at the left ventricular mass as a continuous, quantitative trait. Linear regression was applied on the 17 selected SNPs resulting from chapter 3.1.2.1 and 3.1.2.2 dealing this approach using the STATA software package. The genotype data of all 492 screening cases was applied though having a distinct LVH by definition because all of the available genome-wide genotyped ESTher samples exhibit hypertension in combination with increased left ventricular masses. The dependent variable left ventricular mass (LVM, in grams) was checked to be a normal variable using the STATA software and was transformed to normal distribution using the forth root of the raw data. Linear regression and subsequently *t*-statistics yielding a *p*-value was calculated for the 17

candidate SNPs, assuming full dominance or recessiveness of the minor allele (see Table 3.6 and 3.7). The regression coefficient for each association is given indicating the effect of the minor allele on the size of the left ventricle. The effect unit is the forth root of the LVM in grams as well.

probe set ID	regression coeffcient	*p*-value	95 % CI	
SNP_A-2153126	0.0492869	0.12	-0.0129495	0.1115233
SNP_A-2065547	0.0587992	0.059	-0.0021471	0.1197454
SNP_A-1935686	-0.0577603	0.059	-0.1177215	0.002201
SNP_A-4265373	0.0693444	**0.022**	0.0098342	0.1288546
SNP_A-2291035	0.0668625	**0.028**	0.0072932	0.1264318
SNP_A-2140243	0.0642904	**0.028**	0.0069798	0.121601
SNP_A-4295370	-0.0561534	0.057	-0.1139831	0.0016762
SNP_A-1889406	0.0142838	0.65	-0.0474832	0.0760508
SNP_A-1969214	-0.045509	0.124	-0.1034696	0.0124517
SNP_A-2086003	0.0720394	0.034	0.0055242	0.1385546
SNP_A-2079405	0.0164062	0.586	-0.0426997	0.0755121
SNP_A-2064936	-0.0054076	0.853	-0.0628499	0.0520348
SNP_A-2270325	0.0093319	0.751	-0.0483928	0.0670565
SNP_A-2262176	-0.0388983	0.216	-0.1006153	0.0228186
SNP_A-2186482	0.1234248	**0.000**	0.060743	0.1861067
SNP_A-2011027	0.0584496	0.076	-0.006235	0.1231342
SNP_A-2069616	0.0570635	0.689	-0.223325	0.337452

Table 3.6: Results of linear regression for 17 SNPs resulting from the case control approach using genotype data of 492 genome-wide typed ESTher samples assuming full dominance of the minor allele. The regression coefficient ($\hat{b}, \sqrt[4]{lvm\ [g]}$) and the respective 95 % confidence interval (CI) is given showing the effect direction of the minor allele. The *p*-value is given; a significance threshold of≤ 0.05 was applied. Significant *p*-values are highlighted red.

Five of 17 SNPs showed significant (*p*-value ≤ 0.05) associations to LVM. All of them increased the LVM in a range between 0.064 and 0.123. The minor allele of SNP_A-2186482 having a *p*-value of≤ 0.001 in the dominant model affected the LVM most strongly ($\hat{b}$ = 0.1234248). The remaining markers had moderate effects on LVM and the *p*-values were less significant. The effect of the SNP variants on LVM as far as the recessive model is concerned is displayed in Table 3.7. Six significant associations (*p*-value ≤ 0.05) were determined in which two minor alleles

of three loci (SNP_A-1935686, SNP_A-4295370 and SNP_A-1969214) lower LVM in a range between about – 0.09 to – 0.16. The remaining three SNPs increased the LVM (about 0.17) and all in all, the impact of the analyzed SNPs assuming full recessiveness on LVM was greater than under dominant conditions. One marker showed comparable significant results in both models, SNP_A-4265373. In both models, an increasing effect on LVM was detected. Overall, 10 SNPs had a significant effect on the LVM and were considered for further analyses.

probe set ID	regression coeffcient	*p*-value	95 % CI	
SNP_A-2153126	0.1671072	**0.002**	0.0622322	0.2719822
SNP_A-2065547	0.17909	**0.04**	0.007926	0.3502541
SNP_A-1935686	-0.1545467	**0.01**	-0.2718357	-0.0372578
SNP_A-4265373	0.1798286	**0.029**	0.0187055	0.3409516
SNP_A-2291035	0.0865946	0.343	-0.0925197	0.2657089
SNP_A-2140243	0.0107776	0.901	-0.1588736	0.1804289
SNP_A-4295370	-0.0910379	**0.031**	-0.1738216	-0.0082542
SNP_A-1889406	-0.0590533	0.584	-0.2710481	0.1529415
SNP_A-1969214	-0.1559418	**0.01**	-0.2749293	-0.0369542
SNP_A-2086003	0.0422535	0.62	-0.1251929	0.2096998
SNP_A-2079405	0.0254872	0.713	-0.1103671	0.1613414
SNP_A-2064936	0.0178307	0.754	-0.0937827	0.129444
SNP_A-2270325	-0.0084844	0.881	-0.1199909	0.103022
SNP_A-2262176	0.0228788	0.506	-0.0446179	0.0903755
SNP_A-2186482	0.591369	0.066	-0.0390478	1.221786
SNP_A-2011027	-0.1462865	0.317	-0.4332742	0.1407013
SNP_A-2069616	-0.0245482	0.458	-0.089553	0.0404566

Table 3.7: Results of linear regression (recessive model) for 17 SNPs resulting from the case control approach using genotype data of 492 genome-wide typed ESTher samples. The regression coefficient ($\sqrt[4]{lvm\,[g]}$) and the respective 95 % confidence interval (CI) is given showing the effect direction of the minor allele. The *p*-value is given; a significance threshold of ≤ 0.05 was applied. Significant *p*-values are highlighted red.

In the next step, the linear regression findings (10 SNPs) were corrected for additional covariates. For the well phenotyped 492 ESTher screening samples (see Table 2.1) phenotype data were available having a putative effect on the LVM. The age, sex, waist circumference, body mass index, the mean value of blood pressure measurement over 24 hours (systolic), diabetes status and the appearance of

cardiac heart disease per patient were used as covariates and the linear regression (dominant and recessive model) was applied once again on the 10 significant SNPs shown in Table 3.6 and 3.7. Five SNPs with a significant effect on LVM were determined using the dominant model (see Table 3.8) and showed the same trend in regression coefficient expect of SNP_A-1935686. The *p*-value of this marker was 0.059 before and 0.017 after covariate correction. SNP_A-2140243 lost the significant impact on LVM, the *p*-value dropped from 0.028 to 0.122.

probe set ID	regression coefficient	*p*-value	95 % CI	
SNP_A-2153126	0.0568401	0.06	-0.0023495	0.1160296
SNP_A-2065547	0.0174572	0.532	-0.0374219	0.0723364
SNP_A-1935686	-0.0694063	**0.017**	-0.1262214	-0.0125912
SNP_A-4265373	0.0613817	**0.034**	0.0046712	0.1180921
SNP_A-2291035	0.0718743	**0.014**	0.0147586	0.1289899
SNP_A-2140243	0.0436474	0.122	-0.011667	0.0989618
SNP_A-4295370	-0.0493135	0.079	-0.1044368	0.0058099
SNP_A-1969214	-0.0529695	0.059	-0.1079729	0.002034
SNP_A-2086003	0.076422	**0.018**	0.0133527	0.1394913
SNP_A-2186482	0.133517	**0.000**	0.0737778	0.1932561

Table 3.8: Linear regression results assuming full dominance of the minor allele using left ventricular mass (LVM) as the dependent variable. The *p*-values are adjusted for the covariates age, sex, mean blood pressure over 24 hours (systolic), waist circumference, body mass index, diabetes status and the appearance of cardiac heart disease per patient. Significant *p*-values are highlighted red.

As far as the recessive model is concerned, three SNPs with a significant impact on left ventricular mass before multivariate correction were significant after correction as well (SNP_A-2153126, SNP_A-1935686 and SNP_A-1969214). Regression coefficients were almost identical before and after correction. Three markers were no longer significant after multivariate adjustment, SNP_A-2065547, SNP_A-4265373 and SNP_A-4295370. All results for the 10 considered SNPs are displayed in Table 3.9.

probe set ID	regression coefficient	*p*-value	95 % CI	
SNP_A-2153126	0.1681114	**0.001**	0.0682313	0.2679916
SNP_A-2065547	0.0122498	0.825	-0.0962807	0.1207803
SNP_A-1935686	-0.1534794	**0.007**	-0.2655091	-0.0414497
SNP_A-4265373	0.1471912	0.06	-0.0063885	0.3007708
SNP_A-2291035	0.1320233	0.127	-0.0377941	0.3018407
SNP_A-2140243	-0.0309645	0.709	-0.1939044	0.1319753
SNP_A-4295370	-0.0728403	0.072	-0.152099	0.0064184
SNP_A-1969214	-0.1551457	**0.008**	-0.2689728	-0.0413186
SNP_A-2086003	0.0614297	0.445	-0.0964647	0.2193241
SNP_A-2186482	0.5401594	0.078	-0.0600712	1.14039

Table 3.9: Linear regression results assuming full recessiveness of the minor allele using left ventricular mass (LVM) as the dependent variable. The *p*-values are adjusted for adjusted for the covariates age, sex, mean blood pressure over 24 hours (systolic), waist circumference, body mass index, diabetes status and the appearance of cardiac heart disease per patient. Significant *p*-values are highlighted red.

Overall, there were seven SNPs having a significant effect on LVM after multivariate correction, both regarding the dominant model and the recessive model. These SNPs were prime candidate SNPs for an obligatory replication concerned in chapter 3.1.2.4. In Table 3.10, the annotations for the seven probe set IDs are given (annotated using NetAffx™). All of the polymorphisms are located in intronic regions or upstream respectively downstream of the concerned genes. No SNPs located in gene exons were affected. Two SNPs, SNP_A-1935686 and SNP_A-1969214 are located in the same gene NOS1AP.

probe set ID	dbSNP rs ID	gene relationship	gene symbol	Entrez gene ID
SNP_A-1935686	rs10800465	intron	NOS1AP	9722
SNP_A-1969214	rs10919200	intron	NOS1AP	9722
SNP_A-2086003	rs1469943	downstream / upstream	PELI1	57162
SNP_A-2153126	rs629187	intron	ALCAM	214
SNP_A-2291035	rs11564177	intron	LRRK2	120892
SNP_A-4265373	rs1838467	upstream	GABRB2 / GABRA6	2561 / 2559
SNP_A-2186482	rs1475750	upstream / downstream	GPR63	81491

Table 3.10: The annotated probe set IDs of the LVH associated and LVM controlling SNPs. The corresponding dbSNP (SNP database) rs-number, gene relationship, associated official gene symbol and Entrez gene ID (database of the National Center for Biotechnology Information (NCBI)) are given.

3.1.2.4. Replication of genome-wide candidate SNPs

Seven SNPs were selected to be valid candidates for a replication approach in further affected case samples and unaffected controls. The group of available replication samples was described in Table 2.1. Overall, 2969 replication samples including two groups of cases (I and II) and one control group were allocated. The case group I consisted of the 855 remaining DNA samples from patients of the ESTher group exhibiting the same phenotype characteristics as the screening cases. The cases II group (n = 470) were recruited independently, all showing an arterial hypertension and enlarged left ventricles during echocardiography. Genome-wide genotyped (Affymetrix® Genome-Wide Human SNP Array 5.0) KORA control samples (n = 1644) served as replication controls. SNP replication in cases I (n = 855) and cases II (n = 375) was performed using TaqMan® technology. After genotype calling and data recoding for usage in PLINK, association *p*-values were calculated using basic allelic association and logistic regression (dominant and recessive model corrected for age and sex). In a first step, the ESTher replication cases were compared to the KORA control samples and subsequently the second, independent affected data set cases II was compared to the KORA control samples as well. All of the SNPs were checked to be in Hardy-Weinberg equilibrium (*p*-value > 0.001) in the

control samples. Table 3.11 shows the resulting association *p*-values for the seven tested candidate SNPs for the replication in the ESTher subgroup and the KORA samples. None of the previously determined associations were replicated; the only significant *p*-value (0.01866) was detected using logistic regression (recessive model, SNP_A-2086003). Narrowing down the examined phenotype, the same calculations were repeated selecting only LVH affected samples from replication case I (n = 459). None of the previously determined associations of the seven SNPs could be replicated using this approach either (data not shown). Due to the disappointing replication results as far as the loci associations to LVH are concerned, determination of the impact on the left ventricular mass was dropped.

			allelic assoc.		dom. model		rec. model	
probe set ID	**MAF (cases)**	**MAF (controls)**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
SNP_A-1935686	0.264	0.2654	0.9224	0.993	0.7482	0.9708	0.6784	0.9247
SNP_A-1969214	0.2636	0.2651	0.9133	0.9924	0.6076	0.955	0.966	1.008
SNP_A-2086003	0.1221	0.1115	0.2864	1.109	0.7836	1.03	**0.01866**	2.922
SNP_A-2153126	0.1634	0.1521	0.3251	1.089	0.5365	1.064	0.1913	1.466
SNP_A-2186482	0.08082	0.09046	0.2782	0.884	0.2818	0.874	0.509	1.355
SNP_A-2291035	0.1476	0.1453	0.835	1.019	0.9097	1.012	0.5788	1.205
SNP_A-4265373	0.1455	0.1487	0.7784	0.975	0.3976	0.9155	0.7871	1.077
SNP_A-4292544	0.4895	0.507	0.2584	0.9323	0.5912	1.057	0.07462	0.8293

Table 3.11: Replication results of the seven significant LVH-associated SNPs between the screening cases and controls (basic allelic association and logistic regression results) and having a significant impact on LVM. Replication was performed using TaqMan® technology, genotyping took place in ESTher replication samples (cases I, n = 855) and KORA controls (n = 1644). *P*-values and odds ratios (OR) were determined using basic allelic association (allelic assoc.) and logistic regression using the dominant (dom.) and recessive (rec.) model. The minor allele frequencies (MAF) in cases and controls are given. Significant *p*-values are highlighted red.

The repeated replication using another group of affected samples was unsuccessful as well. None of the detected associations in the screening samples could be replicated. The only significant associated SNP (SNP_A-2153126) was formerly associated without covariate correction using basic allelic association.

			allelic assoc.		dom. model		rec. model	
probe set ID	**MAF (cases)**	**MAF (controls)**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
SNP_A-1935686	0.241	0.2654	0.177	0.879	0.2365	0.8672	0.7831	0.933
SNP_A-1969214	0.2417	0.2651	0.1954	0.8835	0.2519	0.8708	0.8676	0.9587
SNP_A-2086003	0.1042	0.1115	0.5685	0.9264	0.6572	0.9362	0.4476	1.688
SNP_A-2153126	0.1732	0.1521	0.1591	1.168	0.6719	1.057	**0.02281**	2.069
SNP_A-2186482	0.07597	0.09046	0.2124	0.8266	0.2695	0.8321	0.6012	0.6666
SNP_A-2291035	0.1507	0.1453	0.7098	1.044	0.9298	1.012	0.5461	1.294
SNP_A-4265373	0.1527	0.1487	0.7893	1.031	0.8872	0.981	0.1048	1.731
SNP_A-4292544	0.5083	0.493	0.4562	1.063	0.1936	1.198	0.9311	1.012

Table 3.12: Replication results of the seven SNPs that were significantly associated with LVH (basic allelic association and logistic regression results) and LVM (linear regression). Replication was performed using TaqMan® technology, genotyping took place in cases II (n = 375) and KORA controls (n = 1644). *P*-values and odds ratios (OR) were determined using basic allelic association (allelic assoc.) and logistic regression using the dominant (dom.) and recessive (rec.) model. The minor allele frequencies (MAF) in cases and controls are given. Significant *p*-values are highlighted red.

3.2. Comparative mapping approach

The background for the comparative mapping approach to dissect the genetic basis of left ventricular hypertrophy (LVH) as a hypertensive end organ damage was the QTL mapping approach performed in the group of Prof. Dr. med. Reinhold Kreutz. The linkage analysis was performed using the spontaneously hypertensive rat, stroke prone (SHRSP) because the inbred strain exhibits a significant LVH under high blood pressure conditions.

3.2.1. QTL mapping

The SHRSP and the contrasting normotensive F344 rats exhibiting normal heart weights were crossed and the parental animals (each with n = 6) and the resulting F_2-animals (n =232) were phenotyped at the age of 14 weeks (see Table 3.13). Only male rats were selected to avoid sex bias. Mean systolic blood pressure of the F_2-animals was 186.85 ± 26.14 mmHg and lay in between the mean systolic blood pressure of the parental rats (with mean values of 132 and 262 mmHg respectively). The relative left ventricular weight of the intercross animals was slightly increased in contrast to the F344 parental animals but far lower than the relative left ventricular weight of the SHRSP parental rats. Over all, the systolic blood pressure values and the relative left ventricular weight (data not shown) followed a normal distribution.

	F344	SHRSP	F_2 SHRSP x F344		
	mean	mean	min	max	mean
systolic blood pressure [mmHg]	132 ± 8	262 ± 23	139	284	186.85 ± 26.14
rel. left ventricular weight [mg/g]	2.23 ± 0.55	4.27 ± 0.55	2.09	3.88	2.63 ± 0.31

Table 3.13: Phenotype characteristics of F344, SHRSP (each with n = 6) and F_2-intercross rats (n = 232). The systolic blood pressure (in mmHg) and the relative left ventricular weight (in mg left ventricular weight per g body weight) are given. The data were provided by the group of Prof. R. Kreutz, Charité Universitätsmedizin Berlin.

Using the described total genome scan approach and parametric linkage analysis, a major quantitative trait locus (QTL) for the relative weight of the left ventricle in the 232 male F_2-animals (SHRSP x F344) was identified. No further significant QTL regions were detected on the remaining chromosomes.

Figure 3.4: LOD (Log of the Odds) plot for rat chromosome 1 of the complete genome scan of 232 male F_2-animals from an intercross between SHRSP and F344 rats. The microsatellite marker intervals are given in centiMorgan (cM). The blue colored line represents the linkage results for the systolic blood pressure. The relative heart weight is plotted in light blue (peak LOD 10.51), relative weight of the left ventricle in red (peak LOD 8.38). The dotted (LOD 2.8) and solid (LOD 4.3) line mark the significance thresholds for suggestive and significant linkage. The figure was provided by the group of Prof. R. Kreutz, Charité Universitätsmedizin Berlin.

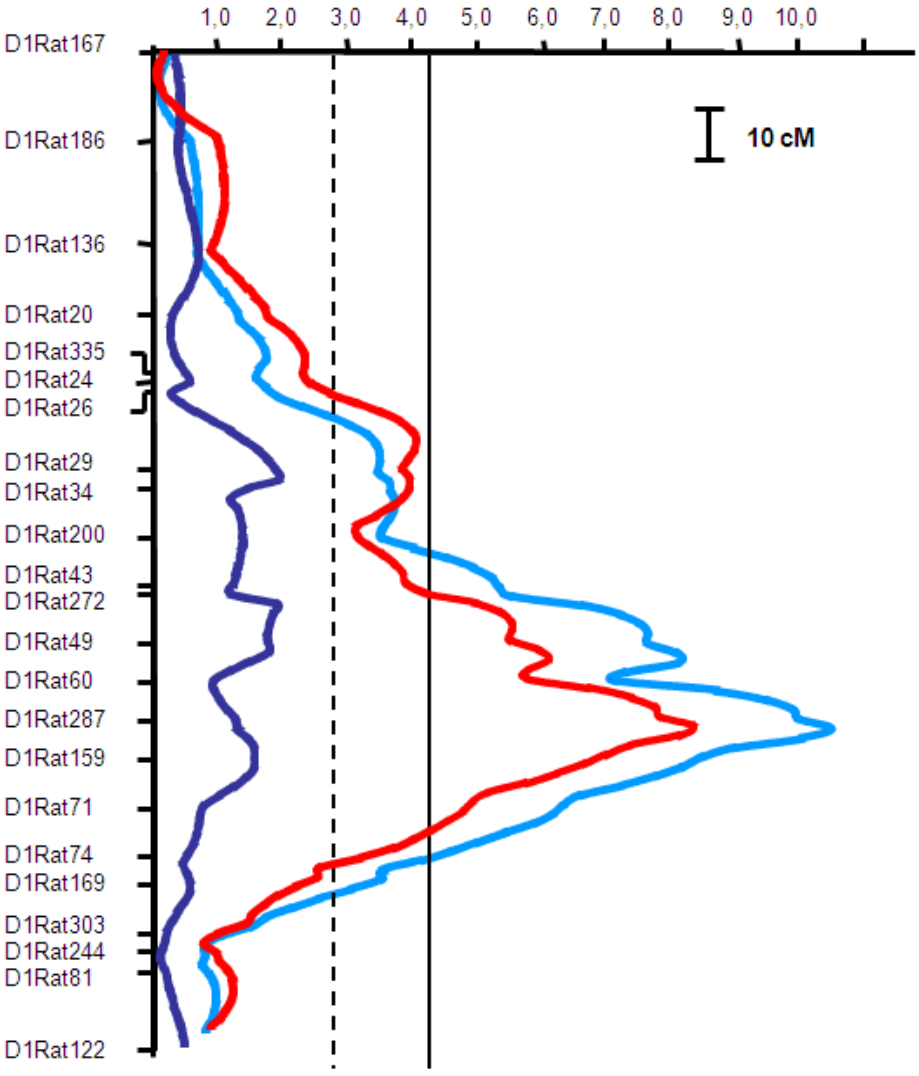


Figure 3.4 shows the resulting LOD plots of the QTL mapping approach for the F_2-intercross animals between the normotensive F344 rats and the hypertensive SHRSP rats on rat chromosome 1. The LOD plots for the three examined phenotypes relative heart weight (light blue), relative left ventricular weight (red) and systolic blood pressures (blue) are shown. Both, the relative heart weight and relative left ventricular weight QTL were significant with LOD scores exceeding the LOD 4.3 threshold. They showed a parallel curve progression and the peak LODs are 10.51 and 8.38 respectively. For systolic blood pressure, no significant or suggestive QTL was detected providing evidence that the relative left ventricular weight QTL was genetic independent from blood pressure in the F_2- animals.

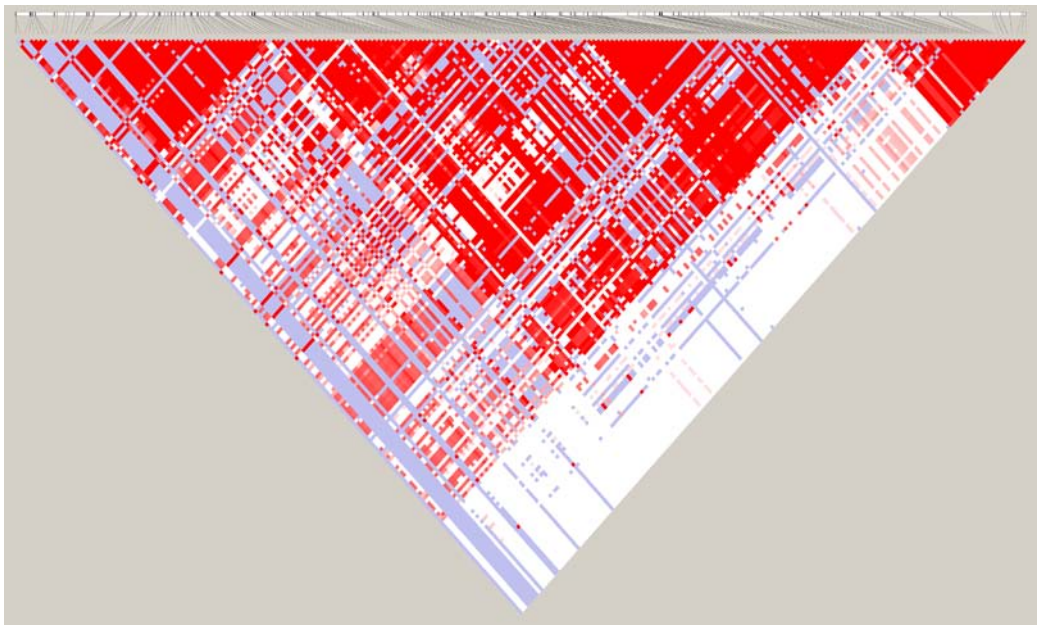
The phenotype of interest for the comparative mapping approach was the relative weight of the left ventricle in the context of hypertension. The corresponding LOD plot of the significant QTL peaked at LOD 8.38 and the peak region is flanked by the microsatellite markers D1Rat287 and D1Rat159. The 1-LOD decrease region from the peak was equal to the region between D1Rat287 and D1Rat159, giving the empiric 95 % confidence interval of the relative left ventricular weight QTL.

3.2.2. Comparative mapping of the QTL region

The resulting 95% confidence interval of the relative left ventricular weight QTL detected in chapter 3.2.1 ranged from D1Rat287 to D1Rat159. The genomic position on rat chromosome 1 for this interval is 190,114,247 - 198,792,612 resulting in an 8,678,365 base pair (bp) region of interest. 52 annotated putative candidate genes (rat genome database) linked to the examined phenotype were localized within this interval using BioMart data management system and the rat data set rat genome assembly 3.4. In addition, 87 potential transcripts were detected within the interval.

Using the Rat Genome Browser available on the Rat Genome Database, the resulting QTL interval was mapped to the human genome (assembly hg18). The best rat/human chain to the mapped rat interval was on the human chromosome 11, from genomic position 58,029,502 to 88,990,550 (Δ 30,961,048 bp). Within this mapped region, 442 annotated genes (Entrez data base) and 1195 detected transcripts were located using BioMart data management system and the GRCh 37 assembly. Dissecting the resulting comparative mapping region, the corresponding HapMap genotype data for Caucasian trios was downloaded from the HapMap Project. The genotyped data was uploaded in HaploView to display the linkage disequilibrium (LD) pattern of the region (Figure 3.5). Figure 3.5 reveals the LD pattern of the comparative mapping region on human chromosome 11. Apparently, the region is divided into three separate LD blocks whereas the two flanking blocks presumptive range over the interval boundaries.

Figure 3.5: Linkage disequilibrium (LD) structure of the human comparative mapping region. The 95% confidence interval of the detected relative left ventricular mass QTL (F_2-intercross of SHRSP x F344 rats) on rat chromosome 1 was mapped on the human genome. HapMap genotype data (Caucasian trios) for the resulting comparative mapping region on human chromosome 11 was downloaded and the LD pattern was displayed using HaploView.



Subsequently, the resulting genomic localization was aligned with the result of the genome-wide association analysis of left ventricular hypertrophy (LVH) described in chapter 3.1. The basis for this alignment were the detected association results of the case control approach using the extremes, LVH affected cases from the ESTher cohort and the PopGen control samples. The genomic positions of significant disease-associated markers of the basic allelic association and logistic regression approach were compared with the genomic position of the rat QTL mapped region on human chromosome 11 (see chapter 3.1.2.1 and 3.1.2.2). Two significantly associated SNPs (logistic model, dominant model, corrected for age and sex, see Table 3.4) were detected in the interval region, both located within the intron 1 and 7 of the gene PACS1 (phosphofurin acidic cluster sorting protein 1, gene ID 55690) on chromosome 11q13.1. PACS1 includes 24 exons and a transcript length of 4488

bp. The exact genomic position of the two SNPs is displayed in Table 3.14. PACS1, a cytosolic sorting protein is located at the center of the LD block displayed in Figure 3.5.

probe set ID	db SNP ID	chromosome	bp	Gene symbol	lokalization
SNP_A-2064936	rs512421	11	65741744	PACS1	Intron 7
SNP_A-2079405	rs580891	11	65711257	PACS1	Intron 1

Table 3.14: Genomic position in base pairs (bp) and annotation data of the two significant associated SNPs resulting from the genome-wide scan of LVH-affected cases and healthy controls corrected for age and sex using logistic regression, dominant model and being located within the comparative mapping interval of the rat relative left ventricle weight QTL 95% confidence interval.

3.2.3. Replication of SNPs identified by comparative mapping

The two SNPs described in chapter 3.2.2 were selected for further replication using TaqMan® technology. For replication, the previously defined replication samples composed of cases group I and II and control samples were used. Replication case group I consisted of 855 remaining DNA samples from patients of the ESTher group. The cases II group was recruited independently, all showing an arterial hypertension and enlarged left ventricles during echocardiography. 375 samples were selected for replication out of 470 existing. Genome-wide genotyped (Affymetrix® Genome-Wide Human SNP Array 5.0) KORA control samples served as replication controls. Detailed phenotype data of the replication samples are given in Table 2.1.

			allelic assoc.		dom. model		rec. model	
probe set ID	**MAF (cases)**	**MAF (controls)**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
SNP_A-2064936	0.2517	0.2333	0.1708	1.106	0.6744	1.04	0.0941	1.674
SNP_A-2079405	0.2159	0.1765	**0.001371**	1.285	**0.009534**	1.28	0.3351	1.234

Table 3.15: Replication results of the two SNPs resulting from the comparative mapping approach for the remaining ESTher samples (n = 855) acting as cases and KORA controls (n = 1644). The minor allele frequencies (MAF) are given for cases and controls. Association with LVH was determined using basic allelic association (allelic assoc.) and logistic regression with adjustment for age and sex using the dominant (dom.) and recessive (rec.) model. Significant *p*-values are highlighted red.

First, the two SNPs were genotyped in the cases group I and the control samples (see Table 3.15). After Hardy-Weinberg quality checking (Hardy-Weinberg *p*-value ≥ 0.001 in controls for both SNPs) basic allelic association was determined for the two SNPs. SNP_A-2079405 was significantly associated with the examined phenotype LVH (*p*-value 0.001371) having an odds ratio (OR) of 1.285. For SNP_A-2064936, no significant association was determined. Subsequently, the calculation was repeated correcting the model for age and sex of the samples using logistic regression. Once again, the association of SNP_A-2079405 detected in the genome-wide scan was replicated (*p*-value 0.009534). Carriers of minor allele of this SNP have a 1.28-fold elevated risk developing of LVH under high blood pressure conditions. The combined *p*-value of the two logistic regression results, for screening and replication samples for this SNP was $1.161 * 10^{-5}$.

The replication approach was repeated using the cases group II consisting of 375 samples (see Table 3.16). The allele frequencies of the two SNPs were checked to be in Hardy-Weinberg equilibrium (*p*-value ≥ 0.001). Once again, the association results of the genome-wide scan for SNP_A-2079405 were replicated both, for a basic

allelic association *p*-value of 0.007924 and a *p*-value after logistic regression of 0.01832. The combined *p*-value (see 2.2.4.6) for genome-wide and replication association after logistic regression, dominant model is $2.149 * 10^{-5}$. In addition, logistic regression using the recessive model resulted in a significant *p*-value for SNP_A-2064936 (*p*-value $3 * 10^{-4}$).

			allelic assoc.		dom. model		rec. model	
probe set ID	**MAF (cases)**	**MAF (controls)**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
SNP_A-2064936	0.27	0.2333	0.05661	1.216	0.892	1.018	**0.00003**	2.578
SNP_A-2079405	0.2213	0.1765	**0.007924**	1.327	**0.01832**	1.364	0.09377	1.61

Table 3.16: Replication results of the two SNPs resulting from the comparative mapping approach for the cases group II (n = 375) and KORA controls (n = 1644). The minor allele frequencies (MAF) are given for cases and controls. Phenotype associations were determined using basic allelic association (allelic assoc.) and logistic regression for age and sex correction using the dominant (dom.) and recessive (rec.) model. Significant *p*-values are highlighted red.

Overall, the PACS1-intronic marker SNP_A-2079405 on human chromosome 11 detected by a genome-wide association study and located within the comparative mapped interval of the rat QTL was the first valid association with significant results in the two independent replication approaches. Using linear regression (data not shown), the impact of the two SNPs on the quantitative trait left ventricular mass was tested. Both, neither SNP_A-2079405 nor SNP_A-2064936 showed a significant result.

Using KORA samples as control samples for replication, it was possible to evaluate whether the detected associations between SNPs and the phenotype left ventricular hypertrophy under hypertensive conditions were independent from high blood pressure to exclude that the two markers were associated with the phenotype hypertension, and thus only indirectly associated with LVH. A new

control sample group was generated, using 770 out of 1644 KORA control samples showing a significant arterial hypertension. Basic allelic association and logistic regression using the dominant model was repeated with both replication cases groups as described above. The resulting *p*-values and OR are shown in Table 3.17 and Table 3.18.

	repl. cases I vs. KORA (all, n = 1644)					
	allelic assoc.		**dom. model**		**rec. model**	
probe set ID	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
SNP_A-2064936	0.1708	1.106	0.6744	1.04	0.0941	1.373
SNP_A-2079405	**0.00137**	1.285	**0.009534**	1.28	0.3351	1.234
	repl. cases I vs. KORA (a. hypert., n = 770)					
SNP_A-2064936	0.09703	1.153	0.4522	1.089	0.1791	1.388
SNP_A-2079405	**0.00347**	1.311	**0.01174**	1.346	0.9276	1.024

Table 3.17: Repeated replication analysis using n = 770 arterial hypertension (a. hypert.) affected controls versus (vs.) replication (repl.) cases group I (n = 855). Basic allelic association (allelic assoc.) determination and logistic regression, dominant and recessive model (dom. and rec., corrected for age and sex) was repeated. Significant *p*-values are highlighted red.

	repl. cases II vs. KORA (all, n = 1644)					
	allelic assoc.		**dom. model**		**rec. model**	
probe set ID	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
SNP_A-2064936	0.05661	1.216	0.892	1.018	**0.00003**	2.578
SNP_A-2079405	**0.007924**	1.327	**0.01832**	1.364	0.09377	1.61
	repl. cases II vs. KORA (a. hypert., n = 770)					
SNP_A-2064936	**0.03367**	1.268	0.6823	1.06	**0.00012**	2.677
SNP_A-2079405	**0.009983**	1.353	**0.0161**	1.406	0.2882	1.386

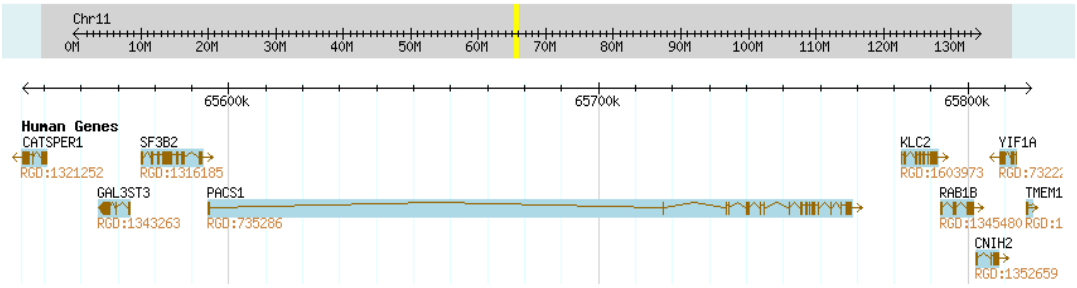
Table 3.18: Repeated replication analysis using n = 770 arterial hypertension (a. hypert.) affected controls versus (vs.) replication (repl.) cases group II (n = 375). Basic allelic association (allelic assoc.) determination and logistic regression, dominant and recessive model (dom. and rec., corrected for age and sex) was repeated. Significant *p*-values are highlighted red.

The levels of significance remained almost identical except for SNP_A-2079405 using logistic regression (dominant model, cases I vs. KORA), the *p*-value increased from about 0.009 to 0.01 but remains significant (*p*-value ≤ 0.05). Using logistic regression assuming full recessiveness and basic allelic association the results for both SNPs had the same dimension after using arterial hypertension affected samples as controls. For replication cases II, the marker SNP_A-2064936 got significant when using arterial hypertension affected KORA controls instead of all KORA samples (from 0.05661 to 0.03367).

3.2.4. Fine mapping of PACS1 region

The two markers (SNP_A-2079405 and SNP_A-2064936) within the introns of PACS1 were the only valid SNPS associated with left ventricular hypertrophy (LVH). Therefore, PACS1 was determined as a new candidate gene for LVH and selected for downstream analysis addressing the attributable genetic variance lying within this genomic region. As described in chapter 3.2.2, PACS1 is located within a block of linkage disequilibrium (LD) on human chromosome 11 as far as HapMap Caucasian genotype data is concerned. The LD block ranges from chromosome 11, base pair position 65,544,242 to 65,817,122 (~ 273 Mb) with the flanking markers rs3829937 and rs479018. The corresponding genomic region is displayed in Figure 3.6. Nine annotated genes are located within the described LD block (see Figure 3.6). PACS1 is located in the middle of the block, it is the largest gene. Downstream of PACS1, there are three further genes (CATSPER1, GAL3ST3 and SF3B2) and upstream five genes (KLC2, RAB1B, Y1F1A, TEM1 and CNIH2).

Figure 3.6: PACS1 is located within a block of strong LD. The genomic structure of this region (chromosome 11, base pair position 65,544,242 to 65,817,122) is shown with the corresponding annotated genes.



Using HaploView, the haplotype composition of the LD block was estimated (see Figure 3.7). The HaploView SNP ID is given above each polymorphism. The SNPs are not numbered consecutively because all SNPs genotyped in the HapMap Project are downloaded for the analysis but only polymorph markers were displayed in HaploView. All haplotypes having a frequency above 3% were examined. Nine haplotypes were detected with frequencies between 17.4 % and 3.1 %.

Figure 3.7: Haplotype structure of the PACS1 containing LD block. For lack of space, the block was divided into two sections. All haplotypes having a frequency above 3 % are displayed. The numbers above each SNP (n = 263) are HaploView internal IDs. Tagging SNPs are labeled by arrows.

Seven tagging SNPs were determined using HaploView and labeled with arrows (Figure 3.7). These seven SNPs (or SNPs having the same information) were selected for a further analysis to dissolve the LD block. The seven tagging SNPs are displayed in Table 3.19. In Figure 3.7, the two underlying SNPs SNP_A-2079405 and SNP_A-2064936 are numbered with 285 and 320 respectively.

dbSNP ID	bp	gene symbol
rs947847	65550518	CATSPER1
rs12576969	65559054	CATSPER1 and GAL3ST3
rs2452680	65576259	GAL3ST3 and SF3B2
rs576740	65609147	PACS1
rs11227408	65625547	PACS1
rs17494956	65749015	PACS1
rs556595	65814658	CNIH2 and YIF1A

Table 3.19: Selected tagging SNPs to dissolve the PACS1-containing LD block. The base pair positions and the proximate genes are given.

Using TaqMan® technology, the seven SNPs were genotyped to resolve the LD structure of the PACS1 containing block on chromosome 11. The replication samples cases I (n = 855) and cases II (n = 470) were used. To avoid possible technology bias comparing genotypes resulting from TaqMan® technology and Affymetrix® SNP Array, 736 KORA samples out of the 1644 available samples served as control samples and were genotyped using TaqMan® technology as well. Unfortunately, genotyping of one out of seven SNPs failed (rs17494956) and the marker was excluded from analysis.

Basic allelic association and logistic regression assuming full dominance or recessiveness (correction for age and sex) was determined for both groups of affected samples in contrast to the KORA control samples, the resulting *p*-values and OR are shown in Table 3.20. Once again, the association analysis was repeated using only hypertensive control samples (n = 372). The allele frequencies of the remaining six SNPs were checked to be in Hardy-Weinberg equilibrium (*p*-value ≥ 0.001). All minor allele frequencies (data not shown) were greater than 5 %.

		repl. cases I vs. KORA (all, n = 736)						repl. cases II vs. KORA (all, n = 736)					
		allelic assoc.		dom. model		rec. model		allelic assoc.		dom. model		rec. model	
gene symbol	**dbSNP ID**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**	***p*-value**	**OR**
CATSPER1	rs947847	0.2656	1.095	0.1377	1.184	0.4034	1.185	0.8997	1.012	0.8077	1.031	0.5309	0.8645
CATSPER1 and GAL3ST3	rs12576969	0.1228	1.161	0.3437	1.123	0.8779	1.9492	0.6642	1.052	0.3729	1.126	0.4132	0.7056
GAL3ST3 and SF3B2	rs2452680	0.1558	1.128	0.07633	1.225	0.3706	1.23	0.9388	0.9922	0.8305	1.027	0.1222	0.6396
PACS1	rs576740	**0.00109**	1.269	0.2344	1.153	**0.0010**	1.59	**0.02383**	1.215	0.0557	1.287	0.05712	1.346
PACS1	rs11227408	0.1818	0.8163	0.1911	0.793	0.8296	0.8589	0.1314	0.7585	0.1987	0.7803	0.9985	1.052
CNIH2 and YIF1A	rs556595	0.4062	0.938	0.5537	0.9232	0.8319	0.9706	0.182	0.8821	0.1001	0.787	0.9881	0.9977
		repl. cases I vs. KORA (a. hypert., n = 372)						repl. cases II vs. KORA (a. hypert., n = 372)					
CATSPER1	rs947847	0.2085	1.136	**0.04735**	1.322	0.5573	1.156	0.6707	1.05	0.4522	1.115	0.41	0.8048
CATSPER1 and GAL3ST3	rs12576969	0.7194	1.043	0.9535	0.9913	0.3694	0.707	0.6657	0.9442	0.9187	1.016	0.1726	0.5367
GAL3ST3 and SF3B2	rs2452680	0.1248	1.177	**0.02701**	1.369	0.5996	1.157	0.7752	1.035	0.4827	1.108	0.1083	0.5965
PACS1	rs576740	**0.0050**	1.288	0.0797	1.292	**0.0134**	1.546	**0.03841**	1.233	**0.04564**	1.354	0.1955	1.266
PACS1	rs11227408	0.7695	0.9453	0.4807	0.8558	0.3796	2.809	0.5506	0.8779	0.5825	0.8825	0.9993	0.412
CNIH2 and YIF1A	rs556595	0.4846	0.9375	0.7351	0.9461	0.6512	0.9265	0.2409	0.8817	0.1338	0.7785	0.9503	0.989

Table 3.20: Replication results of the six tagging SNPs to dissolve the PACS1-containing LD block. Replication was performed using TaqMan® technology, 855 cases I, 470 cases II and 736 KORA controls were genotyped. Phenotype associations were determined using basic allelic association (allelic assoc.) and logistic regression for age and sex correction using the dominant (dom.) and recessive (rec.) model. Association analysis was repeated using hypertensive KORA controls (a. hypert., n = 372) as control samples. Significant *p*-values are highlighted red.

The replication results of the six remaining tagging SNPs are shown in Table 3.20. Two SNPs (rs947847 and rs2452680) were significant (p-value ≤ 0.05) associated with the LVH phenotype after correction for age and sex using the dominant model.

One SNP, rs576740 located within intron 1 of PACS1 (like SNP_A-2079405, see chapter 3.2.3) was associated significantly using basic allelic association showing a p-value of 0.001 for cases I and 0.023 for cases II respectively. Minor allele carriers of this SNP have a 1.215-fold to 1.28-fold elevated risk developing of LVH under high blood pressure conditions. This effect corresponds with the previous findings for the effect of SNP_A-2079405. The repeated calculations using hypertensive KORA controls confirms these findings, the p-values and OR were of the same magnitude as previously reported giving evidence that the detected association is independent of high blood pressure as well. After correction for age and sex, rs576740 became significant using the recessive model in cases group I in comparison to both KORA control groups. No significant associations were determined assuming full dominance except for a slight amendment in cases II compared to hypertension affected control samples resulting in a significant p-value of 0.045. For the remaining tagging SNPs, no significant phenotype association was detected.

Subsequently, genotyping results of the six tagging SNPs were used to reveal LD structure and haplotype association using HaploView. The patterns of LD and haplotype occurrence were determined for both, comparison of cases I (n = 855) and cases II (n = 470) versus all KORA (n = 736) controls and the hypertensive KORA controls (n = 372). The resulting LD pattern and haplotype frequencies are comparable and thus only displayed using all KORA controls (Figure 3.8 and Figure 3.9).

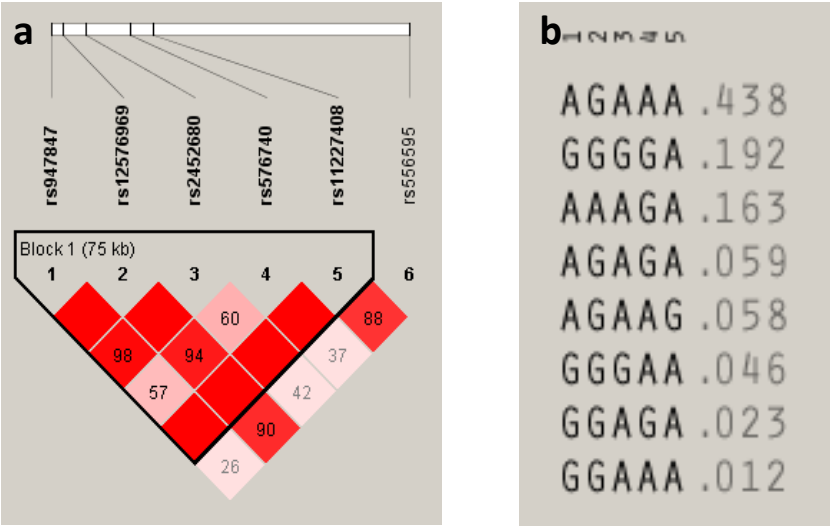


Figure 3.8: LD structure and haplotype frequencies of the PACS1 region using genotype data of six tagging SNPs (see Table 3.19) following from replication in cases I and KORA controls. Haplotype blocks were selected manually (a). Five out of six genotyped SNPs are in LD, the frequencies for the resulting eight haplotypes are given (b).

In Figure 3.8, the resulting LD structure of the six genotyped tagging SNPs in cases I and KORA control samples is displayed. One LD block could be detected, ranging from rs947847 to rs11227408 (a). The frequencies of the resulting haplotypes are given in Figure 3.8 b. The most prominent haplotype AGAAA occurs in 43.8 %. The second common haplotype GGGGA (19.2 %) is the only significantly LVH associated haplotype (*p*-value 0.0281, χ^2 4.82).

In a uniform manner, LD and haplotype analysis was performed using genotype data of cases II and the KORA control samples (Figure 3.9). The detected LD block ranges from rs947847 to rs11227408 as well and the resulting haplotype frequencies are comparable. Two haplotypes are associated significantly in this approach, AGAGA occurring with 6.4 % (*p*-value 0.0475, χ^2 3.928) and GGAGA with 2.7 % (*p*-value 0.047, χ^2 3.947).

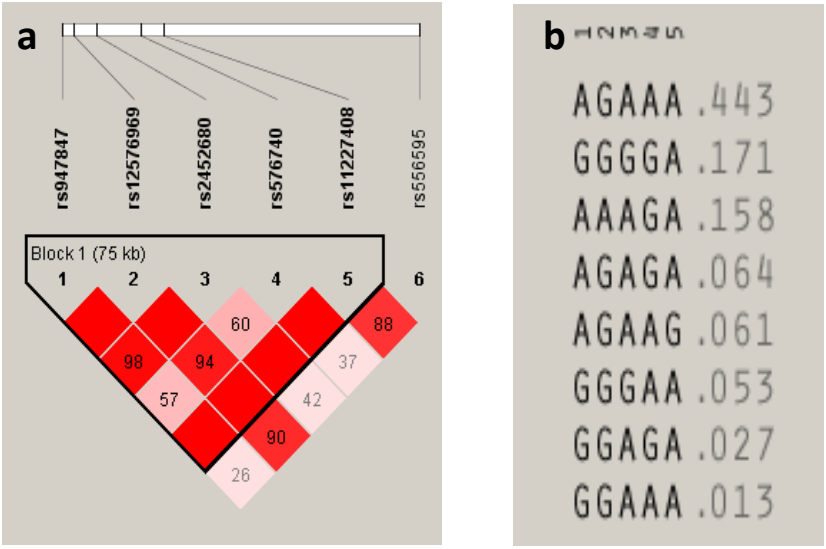


Figure 3.9: LD structure and haplotype frequencies of the PACS1 region using genotype data of six tagging SNPs (see Table 3.19) following from replication in cases II and KORA controls. Haplotype blocks were selected manually (a). Five out of six genotyped SNPs are in LD, the frequencies for the resulting eight haplotypes are given (b).

Unfortunately, the LD structure of the candidate gene region of PACS1 could not be completely resolved in our two studies due to a missing tagging SNP. However, significant haplotype-phenotype associations were detected.

In the future additional block dissolving tagging SNPs should be genotyped. Another possible way resolving the genetic variance underlying the PACS1 region would be targeted resequencing. Further studies involving animal models e.g. zebrafish or knock-out rats will be required to confirm or refute PACS1 as the underlying causative gene.

4. Discussion

The importance to dissect the genetic background of left ventricular hypertrophy (LVH) under high blood pressure conditions is undisputed due to the high LVH prevalence in hypertensives and its role as an independent risk factor for several heart diseases. This challenge requires appropriate methods to investigate the underlying genetic basis. Therefore a genome-wide association study (GWAS) approach was chosen to dissect the global mechanisms of developing LVH. The resulting genotype-phenotype association data was combined with linkage data from a rat model using a comparative mapping approach to provide a proof-of-concept for the analysis of complex traits. In the following section, the chosen GWAS design, its implementation and the resulting associated genetic variants are discussed. Finally, the prospects of comparative mapping approaches for the interpretation of GWAS are addressed and an outlook on future developments in the dissection of the genomic architecture underlying hypertensive LVH is given.

4.1. The genome-wide association study design

The potential of GWAS to uncover modest genetic risk factors in complex disease has been confirmed in a multitude of robustly identified disease associated loci. GWAS is the current method of choice to dissect non-Mendelian multifactorial traits (Johnsons AD and O´Donell CJ 2009). In this thesis a case control study design was chosen for genome-wide analysis of LVH predisposing genetic variants. The case control methodology is straight forward in comparison to other options such as family-based methods e.g. parent-affected-child-trios or cohort study designs. The main advantage is a rapid recruitment of large numbers of affected individuals

through clinical centers and, from a statistical perspective, the best option concerning numbers of individuals needed to achieve sufficient statistical power to detect a moderate effect through association studies. Therefore the approach is relatively cost-effective. On the other hand, the case control design carries the most assumptions. Ascertainment within the same population is implied to avoid population stratification and differences in allele frequencies are supposed to be related to the examined phenotype instead to population differences. Another basic presumption in this study design are well phenotyped and representative case and control samples (Pearson TA 2008).

In our study screening and replication samples for the group of cases and controls were all recruited in Germany and samples selected for genotyping are of Caucasian descent. Consequently, the sample selection for GWAS meets the basic requirements to avoid population stratification, as reflected in the moderate genomic inflation factor and comparable allele frequencies in the two independent study samples investigated. Fortunately, arterial hypertension and LVH are phenotypes which can be diagnosed through non-invasive and standardized methodologies i.e. echocardiography, rendering misclassification of screening and replication cases unlikely. In addition the ESTher survey provided us with plenty available phenotype data for multivariate analysis and to include possible environmental interactions in the genetic analysis. Control samples utilized for screening and replication are usually less well phenotyped and consequently, it is questionable whether control samples are really disease free. The controls from the PopGen program used in the screening step were normotensive, but on average younger than the screening cases; thus age-dependent development of hypertension is possible and could introduce bias into the analysis. The use of hypertensive control samples with clear normal left ventricular mass checked by echocardiography would be more sophisticated, but cost-intensive. Control samples used for the replication were ascertained through the KORA survey and consist of both, normotensive and hypertensive samples, unfortunately without left ventricular mass. Due to a high prevalence of LVH in hypertensive individuals,

affected samples in this control subgroup could not be excluded and a misclassification of cases versus controls might reduce the statistical power of our analysis. Nevertheless, these controls enabled a reduction of misclassified individuals since the KORA controls are currently the best phenotyped population controls available for such analyses.

An established strategy for implementing a GWAS is using a multistage approach to minimize sample size and the amount of required genotyping, primarily the cost-intensive genome-wide scan. Genome-wide interrogation is typically performed in a small fraction of samples in the first step followed by replication steps of the identified associated single nucleotide polymorphisms (SNPs) in larger sample sizes to increase statistical power and to ensure findings for variants with modest effects in complex traits (Hirschhorn JN and Daly MJ 2005). This multistage design was adapted in our study using a relatively small number of screening samples to scan the entire genome for disease associated variants. Subsequently, the identified SNPs were retested in larger sample size and different population, respectively. For high throughput genotyping in stage one, the Affymetrix® Genome-Wide Human SNP Array 5.0 was chosen. At that time, the SNP chip was the best choice having a genomic coverage of 64 % in the CEU HapMap samples and a number of 575 uncovered genes. In the meantime, more comprehensive SNP chip types are available covering for example 93 % of the SNP content of the CEU HapMap samples and only eight genes without SNP coverage like the Illumina Human1M (Li M 2008).

4.2. How to find associations in screening samples

A typical age-related disease like LVH as a hypertensive end-organ damage is a very heterogeneous phenotype due to a multitude of secondary disorders like obesity or type 2 diabetes and therefore hard to analyze. Moreover, LVH is defined as an increase in left ventricular mass in relation to body height and is therefore a quantitative phenotype. A strict exclusion of hypertensive patients with increased left ventricular masses not affected with LVH by definition may lead to a lack of information. In addition, the used screening control samples might be not disease free due to the described lack of phenotype data. Having these limitations in mind, a case control approach was chosen genotyping the extreme cases of the screening samples under stringent conditions. The minor allele frequency threshold was e.g. defined to 10 % to avoid miscalled polymorphisms due to the small number of genotyped samples. As our genome-wide scan revealed no significant associations exceeding the threshold of adjusted p<10-6 as reported in GWAS for other disease traits, a threshold of unadjusted association *p*-value of 0.00001 was chosen to obtain a reasonable number of SNPs. Furthermore adjustment for age and sex was crucial due to phenotype differences in comparison to the control samples as mentioned in 4.1. Subsequently the resulting list of significantly associated SNPs was compared with the results of previously described genetic analysis of LVH introduced in chapter 1.4. Conducting the HyperGen Study population - a family-based study for hypertensives – suggestive linkage was observed on chromosome 2, 4 and 21 for left ventricular mass in whites. None of our significant SNPs were located within these genomic regions (Arnett DK 2009; Tang W 2009). In a pilot case control GWAS, the HyperGen Study identified three SNPs on chromosome 5 and 12 robustly associated with increased left ventricular mass in Whites (Arnett DK 2009). These associations could not be confirmed in our study. Furthermore a gene expression study was published combining linkage and expression analysis of LVH affected rats and humans. A list of differentially expressed putative candidate genes

in humans associated with left ventricular mass is available (Petretto E 2008). None of our significant SNPs is located in one of them or other previously described candidate genes.

As there was no association with known polymorphisms or candidate genes in our analysis we focused downstream analyses on those SNPs, which were on the top of our list of associated SNPs ordered by stringency. A basic premise of our analysis was that variants associated significantly with LVH under high blood pressure conditions should have an impact on the size of the left ventricle. For this reason subsequent steps involved linear regression analyses of quantitative traits affecting LVH on top of simple case control association. Finally SNPs exhibiting significant associations with the LVH phenotype and having an impact on the left ventricular mass were selected as putative candidate SNPs for replication. In this context one of the most interest candidate genes was NOS1AP. Our GWAS analysis exhibited two significantly associated SNPs within NOS1AP. In the recent past four different variants nearby and within this gene were shown to be associated with sudden cardiac death and cardiac repolarization respectively (Eijgelsheim M 2009; Nolte IM 2009).

4.3. Replication – a demanding task

Techniques like correction for multiple testing are insufficient to separate the entire plethora of false-positive associations from the few true associations with disease in GWAS. An essential step in GWAS is replication of association in a larger number of cases and controls or in independent samples. As outlined in the previous sections of this thesis, several case and control samples for the required replication approach were available. For replication purposes, two groups of case samples were analyzed, extension cases of the screening group and independent case samples of another survey each with a sufficient number of individuals. Control samples utilized for replication were different from the screening samples having a gain of information due to a known hypertension status.

Unfortunately replication of genotype-phenotype associations often fails. A lack of reproducibility of associations is a well-known fact, which has already hampered the interpretation of candidate-gene approaches in recent years. This drawback is even greater in GWAS due to the very large numbers of SNPs simultaneously tested. Thus it is an even greater challenge to separate true associations from the blizzard of false positives. In general, a lot of reasons can be specified for a missed replication of association. Commonly discussed problems are e.g. small samples sizes yielding in low power to detect variants of minor to moderate effects. Phenotype differences and heterogeneity in classification of outcomes between the screening and replications samples are problematic as well, since e.g. selection bias between different clinical centers and different methods to assess the disease are likely and hard to control. Another pitfall in replication studies might be a poor study design resulting in populations stratification or differences in exposure to environmental factors (NCI-NHGRI Working Group on Replication in Association Studies 2007).

In this thesis, the initial seven candidate SNPs determined in the screening samples by GWAS were replicated in the described two approaches using extension and independent case samples and much to our disappointment none of the initial valid

associations was replicated. Even though a lot of basic requirements for a successful replication like ascertainment of individuals in the same or similar population were met, versatile reasons for this lack of reproducibility are existent and are outlined in the following. First of all, the sample size of the replication study, particularly for the independent cases, is borderline to ensure sufficient power to detect associations at moderate effect sizes. Selection bias between the screening cases and the independent replication cases is possible due to sample recruitment in different clinical centers. However selection bias cannot be the reason for failed replication in the extension cases of the screening group. As mentioned above no echocardiography data is available for the control samples used in the replication study. Therefore it is questionable whether control samples are really disease free and consequently permit to find a genotype-phenotype association.

Because of the disappointing replication results the reported GWAS design has to be challenged. The drawbacks of the applied GWAS discussed in chapter 4.2 might be too broad to find associations worth replication.

4.4. To learn one´s lesson from comparative genomics

The identification of a robustly replicated SNP-disease association is crucial in identifying disease underlying genetic variants. Unfortunately, we were unable to replicate any of the initial candidate SNPs. Instead of dropping the GWAS results, we decided to implement a comparative genomics approach to dissect the genetic basis of LVH under high blood pressure conditions.

For the comparative mapping approach, a rat disease model for hypertensive left ventricular hypertrophy (spontaneously hypertensive rat, stroke prone, SHRSP) was chosen. The laboratory rat is the most explored experimental animal model by far playing a major role in the study of for example physiology traits. This is due to its pioneer role in domesticating mammalian species for scientific research. The greatest asset using rat models are a lot of well characterized disease models generated by selective breeding. In particular, many inbred models have been developed carrying variation leading to common disease phenotypes like hypertension or diabetes. A long time the mouse has eclipsed the rat as a genetic model (Jacob HJ 1999). In the recent past there has been an increase in rat genomic resources including microsatellite markers, linkage maps, gene expression data and the rat genome sequence followed by a catalogue of about 3 million SNP markers. In addition appropriate database structures for the efficient use of available data like the rat genome database (RGD) were developed (Aitman TJ 2008). These attributes are reflected by a package of 22 significant quantitative trait loci (QTL) for left ventricular mass publically available for different rat strains data mining the RGD. For the comparative mapping approach QTL data of a rat strain was chosen with similar attributes to our human phenotype of interest. The SHRSP rat develops hypertension and a strong LVH phenotype under normal conditions, enhanced by an increased salt intake. Furthermore, the examined locus on rat chromosome 1 shows a remarkable strong linkage to an enlarged left ventricle. The region mapped to a significantly larger region on in the human genome on chromosome 11.

Interestingly, two significantly associated markers dropped primarily due to a lack of impact on left ventricular mass were detected within intron 1 and 7 of the phosphofurin acidic cluster sorting protein 1 (PACS1). These SNPs, SNP_A-2064936 and SNP_A-2079405 were replicated robustly in both, the extension cases of the screening group and the independent case samples and, therefore, were taken forward to a downstream analysis of the PACS1 region. Each of the minor alleles of the two SNPs represents risk alleles for developing LVH. The fine mapping was performed to address the attributable genetic variance within this genomic region. Further genotyping to resolve the complete underlying genomic region revealed another PACS1 intronic SNP (rs576740), which was associated with LVH in hypertensives in downstream analyses. For this reason, we believe that PACS1 is a serious new candidate gene involved in the development of LVH as hypertensive end organ damage.

Interestingly reasons for a failed replication of LVH-associated SNPs discussed in chapter 4.3 were abrogated by confirming the results for PACS1 variants. In all probability the initial GWAS approach is not adequate to find valid associations for LVH under high blood pressure conditions and needs the benefit of comparative genomics to increase power of the findings. Furthermore the question arise as to why variants robustly associated with LVH in hypertensives as those of PACS1 do not have an impact on the left ventricular mass, a prior criterion for exclusion of the study. Maybe, gene-gene interactions are included in the LVH underlying molecular mechanisms that are not detected by GWAS.

To our knowledge, PACS1 was described never before in connection with a genetic association study or heart-associated diseases. PACS1 is a coat protein which localizes membrane proteins in mammalian cells to the *trans*-Golgi network (TGN) (Wan L 1998). As an intracellular sorting protein, it directs the localization of furin and mannose 6-phosphat receptor connecting them to components of the clathrin sorting machinery and it is involved in ion channel trafficking to distinct subcellular compartments (Köttgen M 2005). Recent findings suggest that PACS1 is involved in

the control of ciliary localization of several ion channels and other membrane proteins. Cilia are located on the surface of nearly any mammalian cell. They are microtubule based organelles lacking own protein synthesis. Therefore proteins must be transported into the cilium by several trafficking proteins like PACS1. Deregulation of this protein localization results in a multitude of disease phenotypes called ciliophaties e.g. retinal degeneration or polycystic kidney disease (Jenkins PM 2009). Alteration in cilia function of myocytes due to PACS1 may lead to a change in response to increased biomechanical stress under high blood pressure conditions closing in hypertrophy.

4.5. Outlook

It is in the nature of GWAS to identify a genomic location related to disease but provide little information on gene function. Therefore, functional studies will be required to dissect the molecular mechanisms of the examined phenotype and the putative role of genes with associated SNPs.

In this thesis, PACS1 was identified as a candidate gene influencing development of LVH under high blood pressure conditions due to three significantly associated intronic SNPs. Unfortunately, the attributable genetic variance within the PACS1 gene region could not be completely resolved due to missing tagging SNPs. Additional tagging SNPs genotyped in the near future will help to resolve the genomic region and to extend the finding of haplotype-phenotype associations. Targeted resequencing using next-generation sequencing in a couple of adequate samples provides another possibility to resolve the PACS1 region.

To gain insight into the function of PACS1 *in vivo* a set of methods using animal models is available. An elegant way to quickly functionally assess PACS1 is to conduct Morpholino antisense knockdown experiments using the orthologous loci pacs1 in zebrafish, where cardiac phenotypes can be readily assessed through direct monitoring of the heart in the living animal (Nasevicius A 2000). Another promising approach would be the analysis of PACS1 function in the original dissected model, the SHRSP rat. Thus, different strategies are available but will require a significant amount of time, exceeding the scope and timeframe of this thesis. Generation of consomic and congenic rat strains of the SHRSP rat carrying chromosome 1 or only the region linked to an increased left ventricular mass of a healthy rat are under way. Data not published yet gives evidence that these rats have decreased left ventricular masses. A whole-genome gene expression experiment using Illumina technology is in progress assessing the transcription profile of left ventricle tissue of normotensive F344 control rats and SHRSP rats and the respective consomic strains.

Recently, a gene targeting approach to generate knockout rats was published using zinc-finger nucleases (ZFNs) induce site-specific, double-strand DNA breaks. The ZFNs-encoding mRNA is delivered to rat embryos via microinjection (Geurts AM 2009). This innovative approach closes the gap between mouse and rat functional genetics and enables a PACS1 knockout in our disease model SHRSP to identify it at least as a causative gene or not. For this reason in the end, the wheel of our comparative mapping approach comes to full circle. On one hand it is a powerful tool to facilitate genome-wide analysis in humans like our GWAS. On the other hand comparative genomics provide an opportunity to confirm or to refuse a newly detected candidate gene.

Writing this thesis generated a lot of new ideas to enhance the dissection of the genetic basis of LVH in hypertensives. As mentioned above a multitude of phenotype data is available for the genome-wide genotyped screening samples. Subphenotype analysis is required considering the effect of antihypertensive drugs e.g. the drug type or the time period of drug intake. Moreover other approaches to find associated polymorphisms like haplotype based methods are possible.

5. Literature

Affymetrix (2007). BRLMM-P: a Genotype Calling Method for the SNP 5.0 Array. Affymetrix White Paper.

Aitman TJ, C. J., Cuppen E, Dominiczak A, Fernandez-Suarez XM, Flint J, Cauguier D, Geurts AM, Gould M, Harris PC, Holmdahl R, Hubner N, Izsvak Z, jacob HJ, Kuramoto T, Kwitek AE, Marrone A, Mashimo T, Moreno C, Mullins J, Mullins L, Olsson T, Pravenec M, Riley L, Saar K, Serikawa T, Shull JD, Szpirer C, Twigger SN, Voigt B and Worley K (2008). "Progress and prospects in rat genetics: a community view." Nature Genetics **40**(5): 516-522.

Almasy L, B. J. (2009). "Human QTL linkage mapping." Genetica **136**: 333-340.

Altshuler D, D. M., Lander ES (2008). "Genetic Mapping in Human Disease." Science **322**(5903): 881-888.

Arnett DK, d. l. F. L., Broeckel U (2004). "Genes for Left Ventricular Hypertrophy." Current Hypertension Reports **6**(1): 36-41.

Arnett DK, D. R., Rao DC, Li N, Tang W, Kraemer R, Claas SA, Leon JM, Broeckel U (2009). "Novel genetic variants contributing to left ventricular hypertrophy: the HyperGEN study." Journal of Hypertension **27**: 1585-1593.

Arnett DK, L. N., Tang W, Rao DC, Devereux RB, Class SA, Kraemer R, Broeckel U (2009). "Genome-Wide association study identifies single-nucleotide polymorphism in KCNBI associated with left ventricular mass in humans: The HyperGEN Study." BMC Medical Genetics **10**(43).

Baessler A, K. A., Fischer M, Koehler M, Reinhard W, Erdmann J, Riegger G, Doering A, Schunkert H, Hengstenberg C (2006). "Association of the Ghrelin Receptor Gene Region With Left Ventricular Hypertrophy in the General Population." Hypertension **47**: 920-927.

Barrett JC, F. B., Maller J, Daly MJ (2005). "Haploview: analysis and visualization of LD and haplotype maps." Bioinformatics **21**(2): 263-265.

Benjamini Y and Hochberg Y (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." Journal of the Royal Statistical Society. Series B **57**(1): 289-300.

Bickeböller H, F. C. (2007). Einführung in die genetische Epidemiologie. Berlin Heidelberg, Springer-Verlag.

Carretero OA, O. S. (2000). "Essential hypertension. Part I: definition and etiology." Circulation **101**(3): 329–335.

Classen M, D. V., Kochsiek K, Ed. (2004). Innere Medizin, Urban & Fischer.

Collins A (2009). "Allelic Association: Linkage Disequilibrium Structure and Gene Mapping." Molecular Biotechnology **41**: 83-89.

de Simone G, D. R., Maggioni AP, Gorini M, de Divitiis O, Verdecchia P (2005). "Different Normalizations for Body Size and Population Attributable Risk of Left Ventricular Hypertrophy: The MAVI Study." American Journal of Hypertension **18**: 1288-1293.

Eijgelsheim M, N.-C. C., Aarnoudse AL, van Noord C, Witteman JC, Hofman A, Uitterlinden AG, Stricker BH. (2009). "Genetic variation in NOS1AP is associated with sudden cardiac death: evidence from the Rotterdam Study." Human molecular genetics **18**(21): 4213-4218.

Emigh TH (1980). "A Comparison of Tests for Hardy-Weinberg Equilibrium." Biometrics **36**(4): 627-642.

Fisher RA (1948). "Combining independent tests of significance." American Statistician **2**(5): 30.

Frazer KA, M. S., Schork NJ and Topol EJ (2009). "Human genetic variation and its contribution to complex traits." Nature Reviews Genetics **10**(4): 241-251.

Geurts AM, C. G., Freyvert Y, Zeitler B, Miller JC, Choi VM, Jenkins SS, Wood A, Cui X, Meng X, Vincent A, Lam S, Michalkiewicz, Schilling R, Foeckler J, Kalloway S, Weiler H, Menoret S, Anegon I, Davis GD, Zhang L, Rebar EJ, Gregory PD, Urnov FD, Jacob HJ, Buelow R (2009). "Knockout Rats via Embryo Microinjection of Zinc-Finger Nucleases." Science **325**: 433.

Glazier AM, N. J., Aitmann TJ (2002). "Finding Genes That Underlie Complex Traits." Science **298**: 2345-2348.

Gusella JF, W. N., Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY (1983). "A polymorphic DNA marker genetically linked to Huntington's disease." Nature **306**(5940): 234-238.

Hardison RC (2003). "Comparative Genomics." PLOS Biology **1**(2): 156-160.

Hirschhorn JN and Daly MJ (2005). "Genome-Wide Association Studies for Common Diseases and Complex Traits." Nature Reviews Genetics **6**: 95-106.

Jacob HJ (1999). "Functional Genomics and Rat Models." Genome Research **9**: 1013-1016.

Jenkins PM, Z. L., Thomas G and Martens JR (2009). "PACS-1 Mediated Phosphorylation-Dependent Ciliary Trafficking of the Cystolic-Nucleotide-Gate Channel on olfactory Sensory Neurons." The Journal of Neuroscience **29**(34): 10541-10551.

Johnsons AD and O´Donell CJ (2009). "An Open Acess Database of Genome-wide Association Results." BMC Medical Genetics **10**(6).

Kearney PM, W. M., Reynolds K, Muntner P, Whelton PK, He J (2005). "Global burden of hypertension: analysis of worldwide data." Lancet **365**(9455): 217–223.

Kohler U, K. F., Ed. (2005). Datenanalyse mit Stata. München Wien, R. Oldenbourg Verlag.

Köttgen M, B. T., Simmen T, Tauber R, Buchholz B, Feliciangeli S, Huber TB, Schermer B, Kramer-Zucker A, Höpker K, Simmen KC, Tschucke CC, Sandford R, Kim E, Thomas G, Walz G. (2005). "Trafficking of TRPP2 by PACS proteins represents a novel mechanism of ion channel regulation." The EMBO journal **24**(4): 705-716.

Krawczak M, N. S., Eberstein v. H, Croucher P.J.P., El Mokhtari NE, Schreiber S (2006). "PopGen: Population-Based Recruitment of Patients and Controls for the Analysis of Complex Genotype-Phenotype Realtionships." Community Genetics **9**: 55-61.

Lander E and Kruglyak L (1995). "Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results." Nature Genetics **11**(3): 241-247.

Lander E and Schork NJ (1994). "Genetic Dissection of Complex Traits." Science **265**: 2037-2048.

Lander E, G. P., Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L, (1987). "MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations." Genomics **1**: 174 -181.

Li M, L. C., Guan W (2008). "Evaluation of coverage variation of SNP chips for genome-wide association studies." European Journal of Human Genetics **16**(5): 635-643.

Manolio AT, C. F., Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CH, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE,

Gibson G, Haines JL, Mackay TFC, McCaroll SA and Visscher PM (2009). "Finding the missing heritability of complex diseases." Nature **461**: 747-753.

Matsuzaki H, e. a. (2004). "Genotyping over 100,000 SNPs on a pais of oligonucleotide arrays." Nature Methods **1**(2): 109-111.

McKee PA, C. W., McNamara PM, Kannel WB, (1971). "The natural history of congestive heart failure." New England Journal of Medicine **285**: 1441-1446.

Nasevicius A, E. S. (2000). "Effective targeted gene 'knockdown' in zebrafish." Nature Genetics **26**(2): 216-220.

NCI-NHGRI Working Group on Replication in Association Studies (2007). "Replicating genotype-phenotype associations." Nature **447**: 655-660.

Nisticò L, B. R., Pritchard LE, Van der Auwera B, Giovannini C, Bosi E, Larrad MT, Rios MS, Chow CC, Cockram CS, Jacobs K, Mijovic C, Bain SC, Barnett AH, Vandewalle CL, Schuit F, Gorus FK, Tosi R, Pozzilli P, Todd JA. (1996). "The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. Belgian Diabetes Registry." Human molecular genetics **5**(7): 1075-1080.

Nolte IM, W. C., Newhouse SJ, Waggott D, Fu J, Soranzo N, Gwilliam R, Deloukas P, Savelieva I, Zheng D, Dalageorgou C, Farrall M, Samani NJ, Connell J, Brown M, Dominiczak A, Lathrop M, Zeggini E, Wain LV; Wellcome Trust Case Control Consortium; DCCT/EDIC Research Group, Newton-Cheh C, Eijgelsheim M, Rice K, de Bakker PI; QTGEN consortium, Pfeufer A, Sanna S, Arking DE; QTSCD consortium, Asselbergs FW, Spector TD, Carter ND, Jeffery S, Tobin M, Caulfield M, Snieder H, Paterson AD, Munroe PB, Jamshidi Y. (2009). "Common genetic variation near the phospholamban gene is associated with cardiac repolarisation: meta-analysis of three genome-wide association studies." PLOS One **4**(7).

Pearson TA, M. T. (2008). "How to Interpret a Genome-wide Association Study." JAMA **299**(11): 1335-1344.

Petretto E, S. R., Grieve I, Lu H, Kumaran MK, Muckett PJ, Mangion J, Schroen B, Benson M, Punjabi PP, Prasad SK, Pennell DJ, Kiesewetter C, Tasheva ES, Corpuz LM, Webb MD, Conrad GW, Kurtz TW, Kren V, Fischer J, Hubner N, Pinto YM, Pravenec M, Aitman TJ, Cook SA (2008). "Integrated genomic appraoches implicate osteoglycin (Ogn) in the regulation of left ventricular mass." Nature Genetics **40**(5): 546-552.

Purcell S, e. a. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." The American Journal of Human Genetics **81**: 559-575.

Rabbee N, e. a. (2006). "A genotype calling algorithm for Affymetrix SNP arrays." Bioinformatics **22**(1): 7-12.

Rapp JP (2000). "Genetic Analysis of Inherited Hypertension in the Rat." Physiological Reviews **80**(1): 135-172.

Semplicini A, S. W., Sartori M, Monari A, Naber C, Frigo G, Santonastaso M, Cozzutti E, Winnicki M, Palatini P (2001). "G protein beta3 subunit gene 825T allele is associated with increased left ventricular mass in young subjects with mild hypertension." American Journal of Hypertension **14**(12): 1191-1195.

Sharma P, M. R., Andrew T, Johnson MR, Christley H, Brown MJ, (2006). "Heritability of left ventricular mass in a large cohort of twins." Journal of Hypertension **24**: 231-324.

SMART Study Group (2007). "Rationale and design of the SMART Heart study - A prediction model for left ventricular hypertrophy in hypertension." Netherlands Heart Journal **15**(9): 295-298.

Stoll M, K.-B. A., Cowley AWJ, Harris EL, Harrap SB, Krieger JE, Printz MP, Provoost AP, Sassard J and Jacobs HJ (2000). "New Target Regions for Human Hypertension via Comparative Genomics." Genome Research **10**: 473-482.

Tang W, D. R., Li N, Obermann A, Kitzman DW, Rao DC, Hopkins PN, Class SA, Arnett DK (2009). "Identification of a pleiotropic locus on chromosome 7q for a composite left ventricular wall thickness factor and body mass index: the HyperGEN Study." BMC Medical Genetics **10**(40).

The International HapMap Consortium (2003). "The International HapMap Project." Nature **426**: 789-796.

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

The International HapMap Consortium (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**: 851-862.

Wan L, M. S., Thomas L, Liu G, Xiang Y, Rybak SL, Thomas G. (1998). "PACS-1 defines a novel gene family of cytosolic sorting proteins required for trans-Golgi network localization." Cell **94**(2): 205-216.

Wang X, I. N., Korstanje R, Rollins J and Paigen B (2005). "Indentifiying Novel Genes for Artherosclerosis through Mouse-Human Comparative Genetics." American Journal of Human Genetics **77**: 1-16.

Wichmann HE, G. C., Illig T; MONICA/KORA Study Group. (2005). "KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes." Gesundheitswesen **67**: 26-30.

Wiggington JE, C. D. a. A. G. (2005). "A Note on Exact Tests of Hardy-Weinberg Equilibrium." American Journal of Human Genetics **76**(5): 887–893.

Xu HY, H. X., Wang LF, Wang NF, Xu J (2009). "Association between transforming growth factor beat 1 polymorphisms and left ventricle hypertrophy in essential subjects." Molecular and cellular biochemistry **29**.

Acknowledgment

First of all I would like to express my special gratitude to Moni, my advisor. Thank you so much for supporting my strengths and tolerating my weaknesses at all times. You let me partake in your way of “creative science” and you still gave me the space to develop myself - both unpayable attributes. In the end, you helped me not to get lost during the development of this thesis by supplying me with a multitude of brilliant ideas and corrected proofs in spite of your very busy day. I really appreciate that!

I also want to thank Thorsten Reusch and Joachim Kurtz for supporting my thesis.

Needless to say, that I am also grateful to all of the lovable freaks in our group. Each of you is so special but together, we are an incredible good team. Particularly Steffi and Andreas, who relieved me from the work load during the past few months. Thank you! Frauke, I learned so much from you and I am still inspired of your efforts explaining me the basics of statistic. Hopefully we will keep in touch. And my “girls who play guitars”: Astrid and Tanja. It was very surprising to me to find real friends at work and I have to thank you for all the fun we have! And by the way: Thanks to Paul Smith and Alex Kapranos to keep me in a good mood whenever it was necessary.

Last but not least, I have to thank my family and my friends for creating an environment enabling this thesis. There are so many people I can count on at any time and I cannot describe how much you mean to me.

Abbreviations

BMI	body mass index
bp	base pair
BP	blood pressure
BRLMM	Bayesian Robust Linear Model with Mahalanobis distance classifier
CEPH	Centre d´Etude du Polymorphisme Humain
CHD	cardiac heart disease
CI	confidence interval
cM	centiMorgan
CNV	copy number variation
ECG	electrocardiogram
EM	estimation-maximization
ESTher	Endorganschäden, Therapie und Verlauf
FDR	false discovery rate
F344	Fischer rat
GCOS	GeneChip® Operating Software
GWAS	genome-wide association study
HWE	Hardy-Weinberg Equilibrium
IVSd	end-diastolic interventricular septum thickness
LD	linkage disequilibrium
LOD	log of the Odds
LV	left ventricle
LVBW	left ventricle / body weight
LVEDD	left ventricular enddiastolic diameter

LVH	left ventricular hypertrophy
LVM	left ventricular mass
MAF	minor allele frequency
mmHg	millimeter of mercury
NGFN	National Genome Research Network
NOS1AP	nitric oxide synthase 1 adaptor protein
OR	odds ratio
PAA	polyacrylamide
PACS1	phosphofurin acidic cluster sorting protein 1
PCR	polymerase chain reaction
PWd	left ventricular posterior wall diameter in diastole
QTL	quantitative trait loci
RGD	Rat Genome Database
SHRSP	spontaneously hypertensive rat, stroke prone
SISA	Simple Interactive Statistical Analysis
SNP	single nucleotide polymorphism
SSLP	simple sequence length polymorphisms
TGN	*trans*-Golgi network
ZFN	zinc-finger nuclease

List of publications

Journals

Beetz N, Harrison MD, Brede M, Zong X, Urbanski MJ, Sietmann A, Kaufling J, Barrot M, Seeliger MW, Vieira-Coelho MA, Hamet P, Gaudet D, Seda O, Tremblay J, Kotchen TA, Kaldunski M, Nüsing R, Szabo B, Jacob HJ, Cowley AW Jr, Biel M, Stoll M, Lohse MJ, Broeckel U, Hein L (2009) Phosducin influences sympathetic activity and prevents stress-induced hypertension in humans and mice. Journal of Clinical Investigations 119 (12): 3597-3612

Schnoor M, Buers I, Sietmann A, Brodde MF, Hofnagel O, Robenek H, Lorkowski S (2009) Efficient non-viral transfection of THP-1 cells. Journal of Immunological Methods 344 (2): 109-115

Kreutz R, Bolbrinker J, van der Sman-de Beer F, Boeschoten EW, Dekker FW, Kain S, Martus P, Sietmann A, Friedrichs F, Stoll M, Offermann G, Beige J (2008) CYP3A5 genotype is associated with longer patient survival after kidney transplantation and long-term treatment with cyclosporine. The Pharmacogenomics Journal 8 (6): 416-422

Kreutz R, Schulz A, Sietmann A, Stoll M, Daha MR, de Heer E, Wehland M (2007) Induction of C1q expression in glomerular endothelium in a rat model with arterial hypertension and albuminuria. Journal of Hypertension 25 (11): 2308-2316

Schulz A, Weiss J, Schlesener M, Hänsch J, Wehland M, Wendt N, Kossmehl P, Sietmann A, Grimm D, Stoll M, Nyengaard JR, Kreutz R (2007) Development of overt proteinuria in the Munich Wistar Frömter rat is suppressed by replacement of chromosome 6 in a consomic rat strain. Journal of the American Society of Nephrology 18 (1): 113-121

Wendt N, Schulz A, Siegel AK, Weiss J, Wehland M, Sietmann A, Kossmehl P, Grimm D, Stoll M, Kreutz R (2007) Rat chromosome 19 transfer from SHR ameliorates

hypertension, salt-sensitivity, cardiovascular and renal organ damage in salt-sensitive Dahl rats. Journal of Hypertension 25 (1): 485

Oral Presentations

Sietmann A, Friedrichs F, Schulz A, Kreutz R, Stoll M (2008) Dissection of the development of polygenetic albuminuria in the Munich Wistar Frömter rat through a genomic systems biology approach. SHR Symposium, Prag

Conferences

Sietmann A, Friedrichs F, Huber M, Huge A, Frey N, Weichenhan D, Völler H, Wegscheider K, Schreiber S, Stoll M, Kreutz R (2009) A whole-genome association study in patients with arterial hypertension and heart disease. Dutch-German Joint Meeting of the Molecular Cardiology Groups, Hamburg

Katus HA, Huge A, Sietmann A, Zugck C, Ehlermann P, Friedrichs F, Frey N, Pfeufer A, Kääb S, Ivandic B, Rottbauer W, El-Mokhtari NE, Schreiber S, Stoll M (2008) Chromosome 9q21 as a new major susceptibility locus for dilated cardiomyopathy. NGFN Meeting Heidelberg

Sietmann A, Friedrichs F, Schulz A, Kreutz R, Stoll M (2008) Identification of structural pathways in the development of polygenetic albuminuria through a systems genetics approach. Hypertension, Berlin

Sietmann A, Friedrichs F, Schulz A, Kreutz R, Stoll M (2007) Dissection of UAE QTLs using a systems biology approach. Hypertonie, Bochum

Sietmann A, Friedrichs F, Huber M, Huge A, Frey N, Weichenhan D, Völler H, Wegscheider K, Krawczak M, von Eberstein H, Schreiber S, Stoll M, Kreutz R (2007) A whole-genome association study in patients with arterial hypertension and heart disease. NGFN Meeting, Heidelberg

Sietmann A, Friedrichs F, Urbanczyk C, Huber H, Nassar I, Kreutz R, Stoll M (2006) A new integral systems biology approach to analyze complex diseases. NGFN Meeting, Heidelberg

Sietmann A, Friedrichs F, Urbanczyk C, Huber H, Nassar I, Kreutz R, Stoll M (2006) Genetic analysis of hypertensive target organ damage, a new integrative approach using systems biology. Hypertonie, München