Aus dem Institut für Medizinische Biometrie und Epidemiologie
des Fachbereichs Medizin der Philipps-Universität Marburg
Direktor: Prof. Dr. rer. nat. Helmut Schäfer

Strategies for Genome-Wide Association Analyses of Raw Copy Number Variation Data

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem Fachbereich Medizin
der Philipps-Universität Marburg
vorgelegt von

Ivonne Jarick
aus Cottbus

Marburg, 2013

Angenommen vom Fachbereich Medizin
der Philipps-Universität Marburg am 20.09.2013

Gedruckt mit Genehmigung des Fachbereichs

Dekan: Prof. Dr. Matthias Rothmund
Referent: Prof. Dr. Helmut Schäfer
Korreferent: Prof. Dr. Karl-Heinz Grzeschik

Summary

Copy number variations (CNVs), as one type of genetic variation in which a large sequence of nucleotides is repeated in tandem multiple times to a variable extent among different individuals of one population, have gained much attention with regard to human phenotypic diversity. Recent efforts to map human structural variation have shown that CNVs affect a significantly larger proportion of the human genome than single nucleotide polymorphisms (SNPs). This gave rise to the idea of CNVs playing an important role in explaining some of the large proportion of the phenotypic variance in a population that is due to genetic factors and that could not yet be explained by common SNPs. Current data from SNP genotyping arrays were found to be useful not only for the genome-wide genotyping of SNPs, but also for the detection of CNVs. However, due to the mostly still inadequate accuracy of CNV detection and the rareness of provided methods for association testing, to design a genome-wide CNV association study can be a challenge.

This thesis explored four strategies for the genome-wide association analyses of raw CNV data being derived from the Affymetrix Genome-Wide Human SNP Array 6.0. Initially, the two most commonly used strategic approaches are presented and applied to real data examples for the phenotypes early-onset extreme obesity and childhood attention - deficit / hyperactivity disorder (ADHD). On the one hand, raw intensity values reflecting individual copy numbers are directly tested for an association with the risk of disease, without providing or making use of any information about CNV genotypes. On the other hand, genome-wide CNV analyses are performed as a two-step procedure in first calling individual CNV genotypes and then using these to test for CNV - phenotype associations. Secondly, two extensions of the standard strategies are introduced, which both form its own strategy with a special focus on the intention to overcome problems and weaknesses of the respective widely used strategy. In this sense, one proposed strategy accounts for the fact that thousands of array-provided CNV marker are located in genomic regions without underlying copy number variability, and thus suggests to test only a pre-selected set of relevant and informative intensity values for associations in order to relax the multiple testing issue. Furthermore, the second proposed strategy addresses the known

inaccuracy of CNV calling in especially common CNV regions that is often caused to some extent by the high CNV population frequency and the consequent inadequacy of estimating CNV genotypes relative to sample's mean or median hybridization intensity values. Instead, the use of intensity reference values being estimated in a Gaussian mixture model framework, called MCMR, is investigated in application to data examples for the HapMap and replicate samples as well as to the previously analysed obesity data set. The latter obesity sample has been analysed in use of all four genome-wide CNV analyses strategies which allowed a comparison on the strategy's applicability and performance.

The four strategies were observed to greatly vary in terms of computing efforts and genetic results. Whereas one of the two standard strategies was successful in the identification of rare CNVs at the PARK2 locus being genome-wide statistitically significantly associated with ADHD in children, none of these two strategies detected any CNV - obesity association. Contrarily, alternative MCMR reference intensity values showed improved reliability of CNV calls compared to standard calling in terms of stability, reproducibility and false positive rates. As a consequence, a novel common CNV for early-onset extreme obesity on chromosome 11q11 was identified in application of the proposed analyses strategies. Moreover, a common deletion at chromosome 10q11.22, which was previously reported to be associated with body mass index (BMI), was also replicated in use of one the proposed strategies.

The results suggest that the choice of the genome-wide CNV association analyses strategy may greatly influence genetic results. The presented strategic investigations presented here give an overview on aspects to consider when planning a genome-wide CNV analyses pipeline, but do not allow general recommendations towards an optimal design.

Zusammenfassung

Kopienzahl Variationen (CNVs), als eine Art von genetischer Variation, bei der eine große Sequenz von Nukleotiden im Tandem mehrfach wiederholt ist mit einem variablen Umfang zwischen verschiedenen Individuen einer Population, haben viel Aufmerksamkeit hinsichtlich menschlicher phänotypischer Vielfalt gewonnen. Jüngste Bemühungen die menschliche strukturelle Variation abzubilden haben gezeigt, dass CNVs im Vergleich zu Einzelnukleotid Polymorphismen (SNPs) einen signifikant größeren Anteil des menschlichen Genoms beeinflussen. Dies führte zu der Idee, dass CNVs eine wichtige Rolle spielen könnten in der Aufklärung eines Teils der phänotypischen Varianz in einer Population, die auf genetischen Faktoren beruht und die bisher zum Großteil durch häufige SNPs noch nicht erklärt werden konnte. Aktuelle Daten von SNP Genotypisierungs-Arrays erwiesen sich nicht nur als nützlich für die genomweite Genotypisierung von SNPs, sondern auch zum Nachweis von CNVs. Allerdings kann, aufgrund der meist noch unzureichenden Genauigkeit des CNV Nachweises und der Seltenheit der bereitgestellten Methoden zum Testen von Assoziationen, das Planen der genauen Gestaltung einer genomweiten CNV Assoziations-Studie eine Herausforderung sein.

Diese Dissertation untersucht vier Strategien für genomweite Assoziations - Auswertungen von CNV Rohdaten, welche von dem Affymetrix 6.0 Array gewonnen wurden. Zunächst werden die beiden am häufigsten verwendeten strategischen Ansätze vorgestellt und auf reale Daten Beispiele für die Phänotypen frühmanifeste extreme Adipositas und kindliche Aufmerksamkeits-Defizit / Hyperaktivitäts-Störung (ADHS) angewendet. Auf der einen Seite werden Intensitäts-Rohdaten, welche die individuelle Kopienzahl widerspiegeln, direkt auf eine Assoziation mit dem Krankheits-Risiko getestet, ohne die Bereitstellung oder die Nutzung von Informationen über CNV Genotypen. Auf der anderen Seite werden genomweite CNV Analysen als Zwei-Schritt-Verfahren durchgeführt, in denen zunächst individuelle CNV Genotypen erkannt und anschließend auf CNV - Phänotyp Assoziationen getestet werden. Zum Zweiten werden zwei Erweiterungen der Standard-Strategien eingeführt, die beide eigenständige Strategien darstellen, welche sich besonders auf die Absicht fokussieren Probleme und Schwächen der jeweiligen weit verbreiteten Strategie zu überwinden.

In diesem Sinne befasst sich eine der vorgeschlagenen Strategien damit, dass Tausende der auf dem Array bereitgestellten CNV Marker in genomischen Regionen ohne zugrunde liegende Kopienzahl Variabilität lokalisiert sind, und empfiehlt deshalb nur eine vorab ausgewählte Menge von relevanten und informativen Intensitäts-Werten auf Assoziationen zu testen, wodurch das Problem des multiplen Testens aufgelockert wird. Des weiteren berücksichtigt die zweite vorgeschlagene Strategie die bekannte Ungenauigkeit in der CNV Bestimmung für insbesondere Regionen mit häufigen CNVs, welche oft zu einem gewissen Grad durch die hohe Populations-Häufigkeit der CNVs verursacht wird sowie durch die daraus resultierende Unangemessenheit des Schätzens von CNV Genotypen unter Berücksichtigung von Gruppen-Mittelwert oder Gruppen-Median der Hybridisierungs-Intensitäts-Werte. Stattdessen wird die Verwendung von Intensitäts-Referenz-Werten, welche im Rahmen eines Gaußschen Mischverteilungsmodell geschätzt und als MCMR bezeichnet werden, untersucht im Hinblick auf Anwendungen an Beispieldaten von HapMap- und Replikat-Probanden sowie auch auf den zuvor bereits analysierten Adipositas Datensatz. Dieser Adipositas Datensatz wurde mittels aller vier Strategien zur genomweiten CNV Auswertung analysiert, wodurch ein Vergleich aller Strategien hinsichtlich ihrer Anwendbarkeit und Leistungsfähigkeit ermöglicht wurde.

Für die vier Strategien wurde ein stark unterschiedlich ausgeprägter Rechenaufwand und stark variierende genetische Ergebnissen beobachtet. Während eine der beiden Standard-Strategien seltene CNVs in einem Teilbereich des *PARK2* Gens als genomweit signifikant assoziiert mit ADHS bei Kindern identifizieren konnte, entdeckte keine dieser beiden Strategien auch nur eine CNV - Adipositas Assoziation. Im Gegensatz dazu konnte für alternative MCMR Referenz-Intensitäts-Werte eine verbesserte Verlässlichkeit der geschätzten CNVs im Vergleich zur Standard Detektion in Bezug auf Stabilitäts-, Reproduzierbarkeits- und Falsch-Positiv-Raten gezeigt werden. Als Konsequenz daraus wurde in Anwendung der vorgeschlagenen Analyse-Strategien ein häufiger CNV auf Chromosom 11q11 erstmals als mutmaßlich kausale Variante für frühmanifeste extreme Adipositas identifiziert. Darüber hinaus wurde auch eine häufige Deletion auf Chromosom 10q11.22, für die zuvor bereits eine Assoziation mit dem Körper-Masse-Index (BMI) berichtet wurde, unter Verwendung einer der beiden vorgeschlagenen Strategien repliziert.

Die Resultate deuten an, dass die Strategie-Wahl zur genomweiten CNV Assoziations - Auswertung die genetischen Ergebnisse stark beeinflusst. Die hier vorgestellten Untersuchungen der Strategien geben einen Überblick über Aspekte, die bei der Planung einer genomweiten CNV Analyse-Pipeline zu berücksichtigen sind, sie lassen allerdings keine allgemeinen Empfehlungen bezüglich eines optimalen Designs zu.

The following publications arose from work associated with this doctoral thesis:

(i) **I. Jarick**, A. L. Volckmar, C. Pütter, S. Pechlivanis, T. T. Nguyen, M. R. Dauvermann, S. Beck, Ö. Albayrak, S. Scherag, S. Gilsbach, S. Cichon, P. Hoffmann, F. Degenhardt, M. M. Nöthen, S. Schreiber, H. E. Wichmann, K. H. Jöckel, H. Heinrich, C. M. Tiesler, S. V. Faraone, S. Walitza, J. Sinzig, C. Freitag, J. Meyer, B. Herpertz-Dahlmann, G. Lehmkuhl, T. J. Renner, A. Warnke, M. Romanos, K. P. Lesch, A. Reif, B. G. Schimmelmann, J. Hebebrand, A. Scherag, A. Hinney. Genome-wide analysis of rare copy number variations reveals PARK2 as a candidate gene for attention-deficit/hyperactivity disorder. *Mol. Psychiatry*, doi: 10.1038/mp.2012.161, 2012.

(ii) **I. Jarick**, C. I. Vogel, S. Scherag, H. Schäfer, J. Hebebrand, A. Hinney, A. Scherag. Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum. Mol. Genet.*, 20(4): 840-52, 2011.

In case of both papers, I participated in the data management and in designing the studies. I had the main responsibility for concepting and performing the statistical analyses and also for writing the papers.

Contents

List of Tables

List of Figures

Summary of Commonly Used Notation

Symbols

$\mathbf{a}, \mathbf{b}$	vectors of observed individual covariates like age or sex
$\mathcal{B}(1, p)$	Bernoulli distribution with probability parameter p
$\mathcal{B}(n, p)$	binomial distribution with parameters n and p
c	number of components of a Gaussian mixture model
C_{ind}, C_{ref}	random variables for individual and reference raw copy numbers
$\mathbf{D}_{\cdot} = (D_{\cdot 1}, \ldots, D_{\cdot c})$	vector of individual Gaussian mixture model components
$\mathbb{E}(\cdot\|x)$	expected value conditional on a given observation x
$\mathbb{E}_0(\cdot\|\cdot)$	conditional expected value under the null hypothesis (H_0)
ϵ	minimal assumed variability of a CNV region
$f_\theta(\cdot)$	probability density function (p.d.f.) of a Gaussian mixture model
$\phi(x)$	density of $\mathcal{N}(\mu, \sigma^2)$
H_0	null hypothesis
H_A	alternative hypothesis
$\ln(\cdot)$	natural logarithm function
$\log_2(\cdot)$	logarithm to base 2
n	total sample size (n_a, n_u, n_t: number of cases, controls, trios)
$\mathcal{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathbb{P}(\cdot)$	probability mass function
$\mathbb{P}(\cdot\|x)$	probability mass function conditional on a given observation x

p	total number of available probe sets on the genotyping array
r	number of CNV regions
$v_{\cdot}$	number of probe sets per CNV regions
$\mathrm{Var}(\cdot)$	variance of a random variable
w	total number of probe sets across all CNV regions
$\mathbf{X}_{\cdot} = (X_{\cdot 1}, \ldots, X_{\cdot p})$	vector of random variables for individual pre-processed hybridization signal intensity data
$\mathbf{x}_{\cdot} = (x_{\cdot 1}, \ldots, x_{\cdot p})$	vector of observed individual pre-processed hybridization signal intensity data
$\mathbf{X}_{[\cdot]} = (X_{1\cdot}, \ldots, X_{n\cdot})$	vector of random variables for probe-wise pre-processed hybridization signal intensity data
$\mathbf{x}_{[\cdot]} = (x_{1\cdot}, \ldots, x_{n\cdot})$	vector of observed probe-wise pre-processed hybridization signal intensity data
χ^2	chi-square distribution
$\mathbf{Y} = (Y_1, \ldots, Y_n)$	vector of random variables for the individual disease affection status
$\mathbf{y} = (y_1, \ldots, y_n)$	vector of observed individual disease affection status

Abbreviations

aCGH	array comparative genome hybridization
BIC	Bayesian information criterion
bp	base pair(s)
BMI	body mass index
CNA	copy number alteration or copy number aberration
CNP	copy number polymorphism
CNV	copy number variation or copy number variant
CNVR	CNV region
DSL	disease-susceptibility locus
FBAT	family-based association test
FDR	(tail area) false discovery rate
FWER	family-wise error rate
GLM	generalized linear model
GWAS	genome-wide association study
hg	human genome
HMM	hidden Markov model
HWE	Hardy-Weinberg equilibrium
i.i.d.	independent and identically distributed
kb	kilo base pairs
LD	linkage disequilibrium
lfdr	local false discovery rate
Mb	mega base pairs
MLE	maximum likelihood estimate
p.d.f	probability density function
PS1, PS2	the two proposed strategies for genome-wide CNV analyses
QC	quality control
S1, S2	the two existing strategies for genome-wide CNV analyses
SD	segmental duplication
sd	standard deviation
SNP	single nucleotide polymorphism

1 Introduction

One of the central current goals in human biology is to better understand the genetic contribution to human phenotypes. On the one hand, the study of human genetics is impossible without technological methods providing knowledge of the human genome sequence. On the other hand, partial or even complete information about thousands of individual human genome sequences is worthless without the availability of appropriate statistical methods to classify and evaluate the observed genetic variation.

Recent efforts to map human genetic variation led to the discovery of 38 million single nucleotide polymorphisms (SNPs), 1.4 million short insertion or deletion variants (InDels) and 14 thousand large deletions (Consortium et al., 2012). On average, the genome of each human individual was estimated to differ from the genomic sequence of any other randomly selected human individual in at least 3.7 million SNPs, 350 InDels and 750 large deletions. Thus, with an assumed human genome size of three GB, the genomes of any two non-related people are different at about one in 800 DNA bases and they are less than 99.9 percent the same. To evaluate whether this is a small or a large proportion of variation, one has to keep in mind that the human genome is, on average, only approximately 98.7 percent identical to corresponding sequences in the genome of our closest living relatives, the chimpanzees (The Chimpanzee Sequencing and Analysis Consortium, 2005; Prüfer et al., 2012).

A catalog of genetic differences and similarities in human beings provides a foundation for the study of human genetics. The information is used to screen variants discovered in genomic data from individuals with genetic disorders, cancers or specific phenotypic characteristics. Of special interest are genetic variants with potential functional consequences, such as sequence differences in protein coding regions (i.e. exons of genes) that lead to differences in the encoded protein sequence (Consortium et al., 2010a) or those with high evolutionary conservation (Consortium et al., 2012). Each human individual was estimated to typically harbour $10\,000 - 11\,000$ non-synonymous sites (Consortium et al., 2010a) among which $2\,500$ are at con-

served positions (Consortium et al., 2012). All this variation in DNA and especially its complex combinations offer the potential of being disease causal variants.

For 20 years, family-based linkage studies and small-scale candidate gene studies were successful in the identification of genetic variants leading to monogenic diseases, i.e. to disorders that are caused by mutations in a single gene (also called Mendelian diseases), such as the Huntington's disease (Hardy and Singleton, 2009). In the past few years, genome-wide association studies (GWASs), which investigate more than a million of SNPs in thousands of individuals, have identified hundreds of genetic variants that are robustly associated with complex diseases, such as type 2 diabetes (Hardy and Singleton, 2009). By use of commercial SNP chips, GWASs address the 'common disease - common variant' hypothesis in analysing allelic variants that are present in more than one to five percent of the population. However, most common associated variants were found to have moderate effects (relative risks of $1.1 - 1.5$) and were shown to account for only a small proportion of the trait's phenotypic variance that is due to genetic differences (heritability) (Manolio et al., 2009).

Several potential sources of the 'missing heritability' have been proposed, including i) a yet undetected much larger number of variants with smaller effects, ii) a lesser number of rare variants with possibly much larger effects that are almost undetectable by use of existing SNP chips, iii) structural variants other than SNPs that are only poorly captured by existing SNP chips, iv) gene-gene interactions with low power for being detected, and v) currently neglected environmental factors (Manolio et al., 2009). One specific type of structural variants are copy number variants, (CNVs) which are genomic regions of at least one kilo base (kb) being present in variable numbers across several individuals. CNVs account for a major proportion of human genetic polymorphisms (Redon et al., 2006). Although their role in genetic susceptibility to a variety of human diseases has been predicted to be important, they have not been explicitly examined in most GWASs in the past. With the development of improved methods for CNV detection, this particular type of genetic variation has gained increasing attention throughout the last years.

Several CNVs were found to be associated with many different human diseases, such as autism, schizophrenia, Crohn's disease or psoriasis (Manolio et al., 2009). As with SNPs, disease associated CNVs were so far detected to include rare variants with large association effect sizes as well as a variety of common variants with moderate effects. Due to the strong linkage disequilibrium between SNPs and common CNVs, a large study accounting for several thousand individuals recently concluded that the contribution to human phenotypic variation of most common simple CNVs, those that can be well typed by use of existing SNP chips, was already indirectly

detected in form of phenotypic associations with nearby SNPs (Consortium et al., 2010b). Further large-scale association studies directly addressing the remaining 23% of common CNVs that are of complex, multi-allelic type and that are thus complicated to be classified using SNP chip technology, are still required to comprehensively evaluate the entire phenotypic contribution of common CNVs (Conrad et al., 2010). However, there is a broad consensus on low frequency and rare CNVs as well as any other type of rare variants being promising candidates to explain a large proportion of the 'missing heritability' (Manolio et al., 2009; Conrad et al., 2010).

Due to the widespread availability of SNP chips, as a consequence of GWASs becoming increasingly popular, a special interest on using SNP platform data for CNV analyses has been developed during the past years (Cooper et al., 2008). Currently, high-density SNP arrays have become a convenient tool for the study of CNVs. However, there is still no consensus on the best method for the detection and analysis of such structural variants (Koike et al., 2011; Dellinger et al., 2010).

In this thesis, several approaches for the genome-wide association analysis of CNVs are presented from a statistical point of view. Rather than exploring the variety of genotyping technology, the focus is primarily on the investigation of raw CNV data being derived from the Affymetrix 6.0 SNP genotyping array. Key issues, such as how to account for the uncertainty of CNV genotype calling or which statistical testing methods to choose in the association testing step, are not restricted to any specific genotyping platform. Instead, the major challenges in evaluating the impact of CNVs on phenotype variation are consistently present over different technical methods. Each presented strategy may easily be adapted to variously derived raw CNV data.

Chapter 2 provides a biological overview of CNVs as one particular type of structural variants. Basic characteristics and classifications of CNVs that are essential in understanding the specific challenges in their analyses, are briefly explained.

Chapter 3 describes in detail existing methods for genome-wide association analyses of raw CNV data. This overview addresses the variety of available technical methods, algorithms and software tools for the detection of CNVs as well as statistical aspects of association testing. Two main existing strategies are shown in detail and their applicability is illustrated on real data for the phenotypes obesity and ADHD. Parts of the obtained genetic results were published in Jarick et al. (2012).

Chapter 4 introduces two new statistical strategies for the genome-wide association analysis of raw CNV data. Both proposed strategies involve extensions and modifications of certain parts of the two presented existing strategies into two new

test designs. Key aspects of the two proposed strategies are shortly outlined and compared against each other. In addition to this more general overview, a detailed presentation of both proposed strategies is given in the following two chapters.

Chapter 5 concentrates on the first proposed CNV analyses strategy, which has a special focus on the selection of genomic marker probe sets being tested for an association with the trait of interest. One of the most striking differences between the genome-wide analysis of CNVs and other genomic variants, such as SNPs or short tandem repeats (STRs), is that the locations in which individuals have gained or lost copies of genetic material are a priori unknown. Current genotyping platforms provide SNP probe sets that are designed to reflect the presence or absence of the two SNP marker alleles and (additional) CNV probe sets that are selected for their linear response to copy number changes. The corresponding genome-wide analysis of SNPs is a straight-forward procedure, which includes the assignment of genotype classes AA, AB or BB to each recruited individual and the subsequent association testing at each available genetic marker. Contrarily, any CNV association testing has to additionally address the question of how and with which precision the quantitative continuous measurements produced by genotyping platforms can be transferred into precise DNA copy numbers. Bypassing the genotype calling step and instead directly testing the CNV intensity measurements, does not sufficiently solve this problem. Instead a new problem arises, since not even the existence of CNVs is ensured for any probe sets that might be found to be statistically significantly associated with phenotypic characteristics. Consequently, the introduced strategy involves to restrict association testing on those probe sets with a certain minimal copy number variability. The consideration of this aspect was first proposed by Ionita-Laza et al. (2008). As a first application of the suggested method, a genome-wide CNV association analysis for the binary trait obesity was performed, which was published in Jarick et al. (2011).

Chapter 6 presents the second alternative to standard genome-wide CNV association analyses strategies. The classical two-step procedure is extended by an extensive modeling of the probe-wise copy number neutral intensity measurements prior to the estimation of underlying CNV genotypes and subsequent association testing. The performance of the proposed approach was investigated in comparison to existing methods by application to publicly available HapMap and replicate data. It will be demonstrated that the precision of CNV calling and thus the validity of association testing can be greatly improved by use of sophisticated reference models in terms of stability and reproducibility rates as well as with respect to the percentage of false positive or Mendelian inconsistent CNV calls. Finally, the obesity data set

was re-analysed in application of the second proposed strategy, whereby the CNV-obesity association results can be compared across all four considered strategies. Being aware of the limitations resulting from real data studies, not allowing to make general conclusions about statistical power or type 1 error, this example impressively demonstrates how the choice of the CNV analysis strategy may relevantly change genetic results.

2 Copy Number Variations - Biological Background

Every two humans are genetically different from each other. Genome sequencing of an individual human revealed that a minimum of 0.5 percent variation exists between two haploid genomes, that is to say, only 99.5 percent similarity exists between the two chromosomal copies inherited from each parent (Levy et al., 2007). Due to mutational events, even monozygotic twins have infrequent genetic differences (Bruder et al., 2008).

In the early 1960s, before the availability of sequencing technology, the first observed differences in our genetic architecture, such as aneuploidies, rearrangements, heteromorphisms or fragile sites, were large enough to be identified using a microscope (Feuk et al., 2006a). In contrast to these microscopic structural variants, which are at least three mega base pairs (Mb) in size, submicroscopic structural variants have gained increasing attention in the course of the ongoing technological development. To date, the diversity of genetic variations is classified according to size, structural type and their frequency of occurrence in a population. Accordingly, the spectrum of genetic variants ranges from simple point mutations or more frequent single nucleotide polymorphisms (SNPs) to various repetitive elements of varying size, such as short tandem repeats (STRs), micro- or minisatellites, and of varying structure including insertions, duplications, deletions, inversions or translocations. An overview of human genetic variation is given in Table 2.1 below.

A copy number variation (CNV) is defined as "a segment of DNA that is one kilo base pairs (kb) or larger and is present at a variable copy number in comparison with a reference genome" (Levy et al., 2007). Furthermore, "a CNV can be simple in structure, such as tandem duplication, or may involve complex gains or losses of homologous sequences at multiple sites in the genome" (Redon et al., 2006). CNVs can be classified according to whether they appear in a deleted or duplicated manner, with respect to the number of occurring alleles, that is whether they are bi-, tri-, or multi-allelic, or with regard to their simple or complex structure. Notably, classes of CNVs include insertions, deletions and duplications but not the copy

Table 2.1: Human genetic variation. The table is adapted from Feuk et al. (2006b) and partly updated by recent frequency estimates.

Variation Type	**Definition**	**Frequency (if known) in the human genome**
single nucleotide polymorphism (SNP)	Single base pair (bp) variation found in $> 1\%$ of chromosomes in a given population.	$\sim$ 38 million SNPs in the human population (Consortium et al., 2012)
Insertion/Deletion variant (InDel)	Deletion or insertion of a DNA segment, including small polymorphic changes and large chromosomal aberrations. InDels > 1 kb in size are called CNVs.	$\sim$ 1.4 million bi-allelic InDels in the human genome (Consortium et al., 2012)
Microsatellite or short tandem repeat (SRT)	Sequences containing variable numbers of $1-6$ bp repeats totaling < 200 bp in length.	> 1 million microsatellites in the human genome, $\sim 3\%$ of the sequence
Minisatellite and variable numbers of tandem repeats (VNTRs)	Polymorphic sequence containing $20-50$ copies of $6-100$ bp repeats.	$\sim$ 150 000 minisatellites, of which $\sim 20\%$ are polymorphic
Multisite variant (MSV)	Single nucleotide variant with complex characteristics due to CNV or gene conversion.	The number of MSVs was unknown in 2006.
Intermediate-sized structural variant (ISV)	Gain or loss of a DNA sequence > 8 kb in size also including inversion breakpoints.	297 ISVs were identified using a fosmid library from a single genome.
copy number variation (CNV); copy number polymorphism (CNP); large-scale CNV (LCV)	Copy number change > 1 kb. If the frequency is $> 1\%$, it is called a CNP. LCVs are CNVs ~ 50 kb in size or greater.	$\sim$ 14 000 large deletions (> 500 bp) in the human genome (Consortium et al., 2012)
Inversion	Rearrangement causing a segment of DNA to be present in reverse orientation.	Estimates of microscopically detectable inversion frequencies are $0.12-0.7\%$ (pericentric) and $0.1-0.5\%$ (paracentric)
Translocation	Rearrangement in which a DNA fragment is attached to a different chromosome.	1/500 is heterozygous for a reciprocal translocation and 1/1 000 for Robertsonian translocations
Unbalanced rearrangements	Rearrangements which lead to a net gain or loss of DNA are referred to as unbalanced.	Unbalanced rearrangements occur in $\sim$ 1/1 500 live births.

number neutral translocations or inversions. Individual CNV states are commonly categorized into relative losses or gains. In many cases, the reference genome is assumed to equal the most common genome which harbours exactly two copies of the respective DNA segment - one on each of the two homologous chromosomes. In this case, relative losses can be homozygous or hemizygous deletions with a total of one or none copies of the DNA segment left, respectively. Analogously, relative gains include the presence of a total unphased number of three, four, five or more copies of the respective DNA segment. A graphical representation of exemplary CNV classes with regard to the CNV copy number is given in Figure 2.1 below.

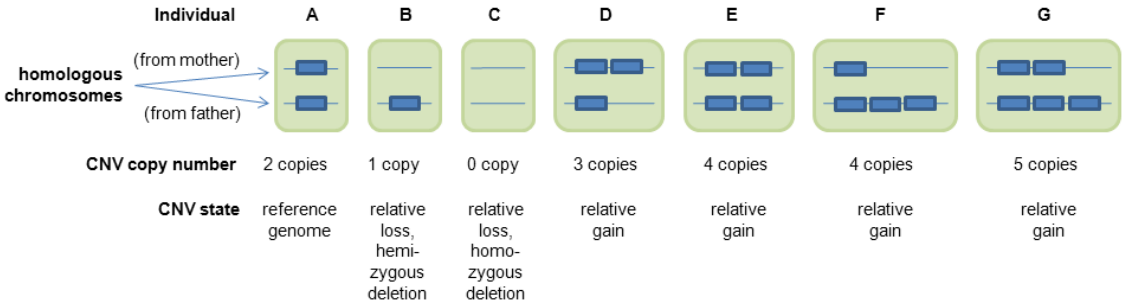

Figure 2.1: Schematic representation of CNV classes based on individual copy number states.

A CNV that is neither inherited from the mother nor from the father is called to appear de novo. Simple CNVs are segments which are deleted or duplicated in tandem. Examples for simple bi-allelic de novo or inherited CNVs are given in Figure 2.2 below.

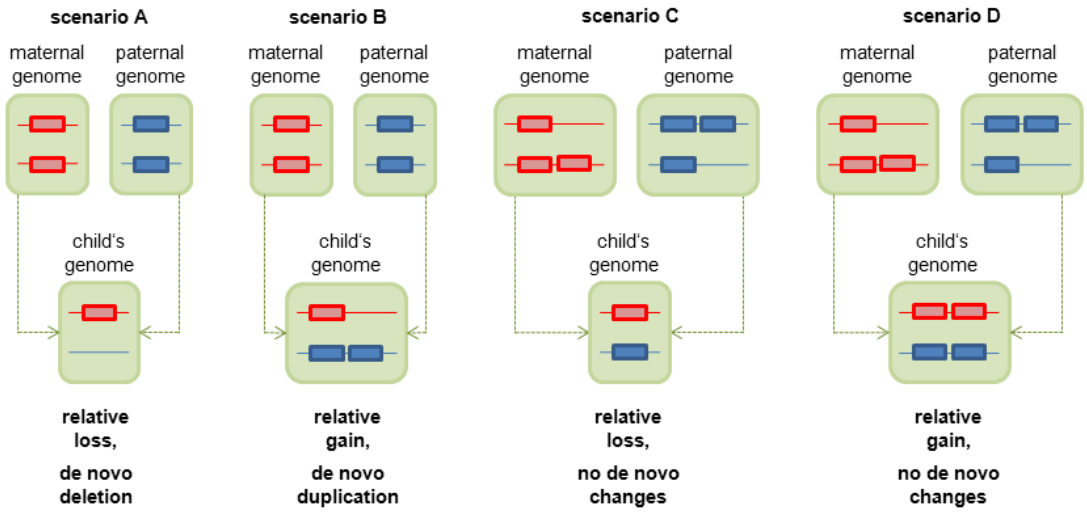

Figure 2.2: Schematic representation of simple bi-allelic CNV inheritance classes. Losses and gains in offspring are defined relative to parental genomes.

In addition to simple CNVs, those that appear at multiple sites in the genome or with a variety of present copies are called complex CNVs. Figure 2.3 illustrates an example of a multi-allelic inherited CNV as well as of a complex de novo CNV.

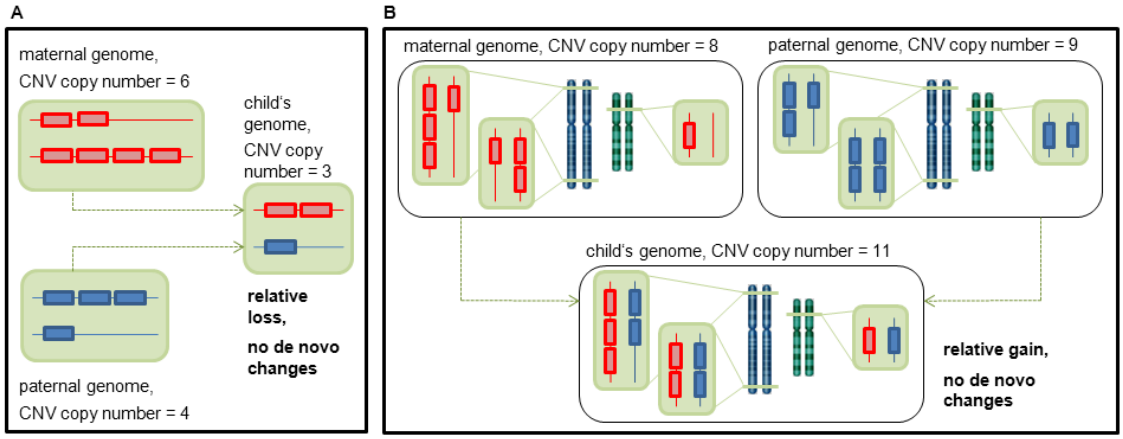

Figure 2.3: Exemplary schematic representation of an inherited multi-allelic CNV (A) and of an inherited complex CNV (B), respectively.

In particular, "a CNV that occurs in more than 1% of the population" (Feuk et al., 2006a) is referred to as a copy number polymorphism (CNP). CNVs that do not affect germline cells, which instead occur in the form of somatic alterations, are more precisely denoted as copy number alterations or copy number aberrations (CNAs). Closely related to CNVs, those duplications reaching fixation in the population are then visible in the genome as segmental duplications (SDs). SDs "are defined as duplicated genomic regions of > 1 kb with 90% or greater sequence identity among the duplicates" (Kim et al., 2008).

Concerning the mechanism of copy number changes, CNVs can arise both meiotically and somatically, as shown by the findings that monozygotic twins can display different DNA copy number variation profiles (Bruder et al., 2008). Moreover, CNV copy numbers can vary across different organs and tissues from the same individual (Piotrowski et al., 2008). In general, there are two mechanisms that cause changes in structure of chromosomes: homologous recombination and non-homologous recombination (Lin et al., 2011). There are at least two main mechanisms for changes in CNV copy number: non-allelic homologous recombination and microhomology-mediated events (Lin et al., 2011). Homologous recombination is the basis of many DNA repair processes. When a damaged sequence is repaired by use of a homologous sequence at the same chromosomal position on the sister chromatid or on the homologous chromosome, there will be no structural changes. Contrarily, repair processes utilizing homologous sequences in different chromosomal positions, which are called non-allelic homologous recombination, can change the chromosome structure.

3 Overview of Methods for the Genome-Wide Association Analysis of Raw CNV Data

This chapter is an introduction to the methodical aspects of genome-wide copy number variation analyses. Currently, most CNV analyses consist of a discovery and of an association testing step. In the subsequent paragraphs, an introduction to recent developments concerning both steps will be given, with a special focus on the analysis of case-control and family-based samples that were previously genotyped with the Affymetrix Genome-Wide Human SNP array 6.0. Two main existing strategies for the genome-wide association analysis of array-derived CNV data will be presented and their application will be illustrated on real data examples for the phenotypes obesity and attention - deficit / hyperactivity disorder (ADHD).

3.1 Technology for CNV Discovery

With regard to the detection of individual copy number variations, there is a variety of methods that assay the genome at either a genome-wide or at a targeted level, with varying degrees of resolution. An overview of those approaches that potentially had the greatest impact on recent CNV discoveries (Feuk et al., 2006a) is given in Table 3.1.

3.1.1 Microarray Technology for CNV Discovery

Besides quantitative, primarily PCR-based assays, array-based analyses are the second main approach for identifying CNVs (Feuk et al., 2006a). Two-channel array-based comparative genome hybridization (aCGH) and SNP genotyping arrays are the two major types of data that serve as the source of CNV discovery using microarrays. Table 3.2 lists the most important, currently available microarray platforms for genome-wide CNV detection.

Table 3.1: Technical methods for the detection of CNVs in the human genome. Adapted from (Feuk et al., 2006a).

Method	**CNVs,** $1-50$ **kb**	**CNVs,** >50 **kb**
Genome-wide scans		
Karyotyping	No	Yes (>3 Mb)
Clone-based array comparative genome hybridization (aCGH)	No	Yes (>50 kb)
Oligonucleotide-based array comparative genome hybridization (aCGH)	Yes (>35 kb)	Yes (>35 kb)
SNP array	Yes	Yes
Sequence-assembly comparison	Yes	Yes
Clone paired-end sequencing (fosmid)	Yes (deletions >8 kb; insertions <40 kb)	Yes (deletions >8 kb)
Targeted scans		
Microsatellite genotyping	Yes (deletions)	Yes (deletions)
Multiplex amplifiable probe hybridization (MAPH)	Yes	Yes
Multiplex ligation-dependent probe amplification (MLPA)	Yes	Yes
Quantitative multiplex PCR of short fluorescent fragments (QMPSF)	Yes	Yes
Real-time quantitative polymerase chain reaction (qPCR)	Yes	Yes
fluorescence in situ hybridization (FISH), including metaphase, interphase and fibre FISH	Yes	Yes
Southern blotting	Yes	Yes

Concerning array-based comparative genome hybridization (aCGH) approaches, labelled fragments from a genome of interest are competitively hybridized with a second differentially labelled genome to arrays that are spotted with cloned DNA fragments. The array can be spotted with different DNA sources. Genomic clones, for example bacterial artificial chromosome (BAC) clones, PCR fragments or oligonucleotides can all be used as array targets. After hybridization, determination of the fluorescence ratio reveals differences in copy number between the test and the reference DNA sample. The first reported application of CGH technology was in 1992 to investigate copy number differences between cancer cells and healthy cells at the chromosome level (Shen and Wu, 2009). In 1997, the first microarray CHG technology was developed with substantially improved resolution as a result of using cloned genomic DNA as probes in a microarray format, which contain sequence

Table 3.2: Major commercial microarray platforms and their current products. Adapted from (Shen and Wu, 2009).

Platform Type	**Company**	**Array platform**	**Median probe spacing (resolution)**	**Probe set number**	**Oligonucleotide probe type**
aCGH					
	Agilent Technologies, Santa Clara, CA				
		4 × 44K CGH array	43 kb	43.000+	60-mer
		8 × 60 K CGH array	41.4 kb	55.000+	60-mer
		2 × 105 K CGH array	21.7 kb	99.000+	60-mer
		4 × 180 K CGH array	13 kb	170.000+	60-mer
		244 K CGH array	8.9 kb	236.000+	60-mer
		2 × 400 K CGH array	5.3 kb	411.000+	60-mer
		1 Million CGH array	2.1 kb	963.000+	60-mer
	NimbleGen, Madison, WI				
		HG18 CGH 4×72 K WG Tiling v2.0	40 kb	72 000	50- to 75-mer
		385K WG Tiling, single array	6.27 kb	385 000 / array	50- to 75-mer
		385 K WG Tiling, 4-set array	1.57 kb	385 000 / array	50- to 75-mer
		385 K WG Tiling, 8-set array	713 bp	385 000 / array	50- to 75-mer
SNP genotyping platform including CN probes					
	Affymetrix, Santa Clara, CA				
		Genome-Wide Human SNP Array 6.0	0.7 kb	906 600 +946 000 (SNP+CN)	25-mer
	Illumina, San Diego, CA				
		HumanCNV370-Quad DNA analysis BeadChip	4.9 kb	320 000 +60 000 (SNP+CN)	50-mer
		Human610-Quad DNA analysis BeadChip	2.7 kb	550 000 +60 000 (SNP+CN)	50-mer
		Human1M-Duo BeadChip	1.5 kb	1.1×10^6 SNP + CN probe sets targeting exons	50-mer

information that permit their specific localization in the human genome. Due to its extensive coverage of the genome, aCGH with BACs is particularly popular (Feuk et al., 2006a). Compared to the use of BACs, aCGH comprising long oligonucleotides ($60 - 100$ bp) with increasing smaller inter-probe spacing can improve the detection resolution, which starts from 50 kb when using BACs and ranges from 30 to 50 kb for most available oligonucleotide arrays (Feuk et al., 2006a).

SNP arrays were explicitly developed to genotype germline encoded single nucleotide polymorphisms. In addition to information about SNP genotypes, current SNP genotyping platforms also provide copy number information in form of hybridization intensity signals that are obtained from spotted oligonucleotides on the SNP arrays. SNP microarrays are a specific type of oligonucleotide arrays, in which the SNP array probes are explicitly designed to indicate the alternative alleles of SNPs. For a test sample, the measured strengths of hybridization to each probe directly reflects the content of nucleic acid in each sample, and can thus be used as a measure of DNA copy number. In contrast to CGH arrays, genotyping arrays do not use a specific control sample. Instead, changes in copy number are detected by comparing individual hybridization intensities with averaged sample hybridization intensities being probe-wise derived from a group of selected control subjects.

3.1.2 Affymetrix Genome-Wide Human SNP Array 6.0

One of the currently most popular genotyping arrays for genomic profiling is the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix 6.0), which combines SNP probes for SNP genotyping with additional CNV probes that are specifically designed for the detection of DNA copy number changes.

In particular, the hybrid array Affymetrix 6.0 contains 909 622 SNP probe sets and additional 945 826 non-polymorphic probe sets for CNV analyses. The latter copy number probes are sequential oligonucleotide probes that do not depend upon the presence of SNPs. Both, SNP and non-polymorphic sites are represented by clusters of identical oligomers of 25 nucleotides immobilized at a specific location on the microarray. Each such cluster of oligomers is commonly referred to as a probe (Chai et al., 2010). Since for most SNPs only two alleles are observed, the examination of a specific SNP is based on several pairs of SNP probes, which differ in just one nucleotide at the position of the SNP locus. Probes targeting the same SNP or non-polymorphic site are commonly referred to as probe sets (Chai et al., 2010). Among the non-polymorphic probe sets, 744 000 probe sets are evenly spaced along the chromosome, and the remaining 202 000 probe sets target 5 677 known CNV regions reported in the Toronto Database of Genomic Variants (DGV) (Affymetrix, 2009). A

Table 3.3: Basic summary of the content of the Affymetrix Genome-Wide Human SNP Array 6.0 according to the current Affymetrix annotation files 'GenomeWideSNP_6.na25.annot.csv' and 'GenomeWideSNP_6.cn.na25.annot.csv' (www.affymetrix.com).

Chr	# SNP probe sets	median SNP probe distance [bp]	mean SNP probe distance [bp]	# CN probe sets	median CN probe distance [bp]	mean CN probe distance [bp]	# CN probe sets in known CNV regions	total # probe sets (SNP & CN)	median probe distance (SNP & CN) [bp]	mean probe distance (SNP & CN) [bp]
NA*	1 224	-	-	20	-	-	-	1 244	-	-
1	71 444	1 306	3 452	73 055	2 166	3 358	18 370	144 499	689	1 698
2	74 103	1 391	3 275	77 799	2 296	3 095	22 942	151 902	722	1 585
3	60 838	1 356	3 276	65 499	2 210	3 019	18 321	126 337	699	1 565
4	56 134	1 429	3 405	62 799	2 209	3 021	15 865	118 933	705	1 595
5	56 569	1 329	3 191	57 764	2 263	3 103	13 719	114 333	702	1 567
6	56 400	1 288	3 026	55 040	2 143	3 078	13 504	111 440	665	1 520
7	47 144	1 395	3 366	52 674	2 011	2 989	18 249	99 818	662	1 577
8	48 753	1 154	2 998	48 287	2 166	3 004	15 609	97 040	643	1 495
9	41 521	1 083	3 376	39 515	1 883	3 522	12 819	81 036	577	1 718
10	48 284	1 151	2 800	44 047	2 153	3 047	13 727	92 331	627	1 453
11	44 624	1 241	3 009	43 671	2 128	3 049	13 369	88 295	655	1 508
12	42 670	1 309	3 100	43 539	2 210	3 013	11 504	86 209	681	1 522
13	34 362	1 199	2 799	30 948	2 423	3 084	9 293	65 310	662	1 461
14	28 160	1 305	3 092	28 179	2 360	3 108	8 270	56 339	724	1 555
15	26 120	1 257	3 135	26 690	2 171	3 048	8 813	52 810	687	1 540
16	27 772	1 042	3 193	25 557	2 056	3 450	10 264	53 329	626	1 653
17	20 693	1 615	3 800	25 331	1 990	3 080	9 231	46 024	763	1 695
18	26 620	1 214	2 859	24 890	2 344	3 032	6 487	51 510	665	1 466
19	11 912	2 156	5 338	17 943	1 764	3 528	8 174	29 855	802	2 120
20	22 891	1 187	2 725	20 161	2 165	3 071	5 385	43 052	647	1 438
21	12 606	1 130	2 933	12 181	1 951	3 026	4 425	24 787	589	1 487
22	11 546	1 229	3 043	12 454	1 453	2 798	5 925	24 000	559	1 452
X	36 865	1 679	4 199	49 200	2 183	3 120	14 772	86 065	816	1 784
Y	257	36 643	95 254	8 583	944	6 355	2 479	8 840	912	6 170
MT	110	102	144	0	-	–	0	110	102	144
Total	909 622	1 297	3 262	945 826	2 154	3 144	281 516	1 855 448	681	1 604

* NA = not available.

detailed characteristic reflecting the content of the Affymetrix 6.0 microarray is given in Table 3.3. In summary, due to the availability of non-polymorphic copy number probes in addition to SNP probes, the Affymetrix 6.0 array provides approximately doubled resolution (median probe distance = 681 bp) for CNV detection compared to the exclusive presence of SNP probes (median probe distance = 1 297 bp).

3.2 Software for CNV Discovery

To date, several methods for CNV detection based on genome-wide SNP array data are available. Reflecting just a fractional amount of available tools, Table 3.4 sum-

marizes those tools for the identification of CNVs that are currently most widely used.

Table 3.4: Summary of widely used software for detecting CNVs from SNP array data. Adapted from (Winchester et al., 2009).

Software	Platform	Related Publication	Details
Birdsuite (Birdseye & Canary)	Affymetrix	Korn et al. (2008), McCarroll et al. (2008)	Combined tool set for genotyping of SNPs and known CNPs, integrated in the Genotyping console (GTC) 3.0
CNAT	Affymetrix	Huang et al. (2004), Affymetrix (2007)	'Copy Number Analysis Tool' for proprietary-run in Genotyping Console (GTC) 3.0
CNVPartition 1.2.1	Illumina	Illumina (2010)	Proprietary-run in BeadStudio
dChip SNP	Affymetrix or Illumina	Li et al. (2008)	HMM based stand alone software
GADA & GADA-JRN	Affymetrix or Illumina	Pique-Regi et al. (2008) & Pique-Regi et al. (2009)	'Genome Alteration Detection Algorithm' uses sparse Bayesian Learning
HMMSeg	Multiple	Day et al. (2007)	HMM application tool for any genomic data
ITALICS	Affymetrix	Rigaill et al. (2008)	R package 'Iterative and Alternative normaLIsation and Copy number calling for Affymetrix SNP arrays'
Nexus Biodiscovery, CBS	Multiple	Olshen et al. (2004)	Commercial 'Circular Binary Segmentation' detection tool
PennCNV	Illumina or Affymetrix	Wang et al. (2007)	Guided HMM application, Perl script based
QuantiSNP	Illumina or Affymetrix	Colella et al. (2007)	Guided HMM application, command line based
SCIMM and SCIMM-Search	Illumina	Cooper et al. (2008)	'SNP Conditional Mixture Modelling' algorithm implemented in R
TriTyper	Illumina	Franke et al. (2008)	Identify and genotype SNPs with null allele

The wide range of presently available analysis tools for CNV detection varies in terms of the algorithm used for CNV genotype calling, in the extent to which components for pre- and post-processing of the data are provided and in the applicability across genotyping platforms. The two most prominent commercial SNP array vendors, Affymetrix and Illumina, provide specially designed proprietary software for complete CNV detection analyses, the Genotyping Console and the BeadStudio.

Due to limitations on CNV analyses using proprietary software, a variety of alternative tools were developed. One group of CNV detection tools, such as dChip,

HMMSeg, PennCNV or QuantiSNP, are based on hidden Markov models (HMMs) by assuming that observed intensity values are directly related to the unobserved copy number states via locus-specific emission distributions. Furthermore, copy number states of neighboring loci are assumed to be similarly characterized, i.e. to depend on each other.

Moreover, a variety of methods that were originally developed for CNV analyses based on aCGH data have been adapted to the use with SNP array data. For instance GADA, which can be applied to SNP array data, is a modification of the Circular Binary Segmentation (CBS) algorithm (Olshen et al., 2004), originally designed for aCGH data. The main idea of the segmentation algorithm was to convert noisy intensity values into regions of similar copy number by continuously dividing a region into segments until each segment is differently composed and can clearly be distinguished from neighboring sections.

Alternative approaches for CNV detection from SNP array data involve conditional mixture models. In the SCIMM tool, the observation that copy number losses appear to have unique signal-intensity clusters is explicitly used for the identification of deletions.

Furthermore, ITALICS is based on separating probe sets with abnormal intensities from copy number neutral probe sets. Iteratively, copy number states are estimated for each probe set, and multiple linear regression is used to estimate the non-linear effects on the copy number.

Finally, TriTyper combines CNV prediction and SNP genotyping by use of a maximum likelihood estimation in order to detect deletions. More precisely, SNP genotyping clusters are modeled in a way to additionally incorporate an extra, so called null allele, and deviations from Hardy-Weinberg equilibrium (HWE) are used as an indicator for the presence or absence of this null allele, which is assumed to reflect deleterious events.

In addition to the core step of calculating individual copy number states, any CNV discovery analysis also consists of several further steps, which are schematically represented in Figure 3.1. The detection of CNVs from SNP array data usually starts with pre-processing of the typically noisy raw output data from microarray experiments. At first, hybridization intensity signals of each individual array are extracted and summarized across probe sets. Afterwards, quantile normalization was shown to perform favorably in removing those variation between arrays, which is of potential non-biological origin (Bolstad et al., 2003). The final step in data pre-processing is to calculate individual continuous raw copy numbers per probe sets,

which are subsequently categorized into discrete copy number states in the following CNV genotype calling step. At the end of a CNV detection pipeline, spurious CNV calls, such as for instance singletons or those that cover large gaps between probe sets, are removed in the course of quality control (QC).

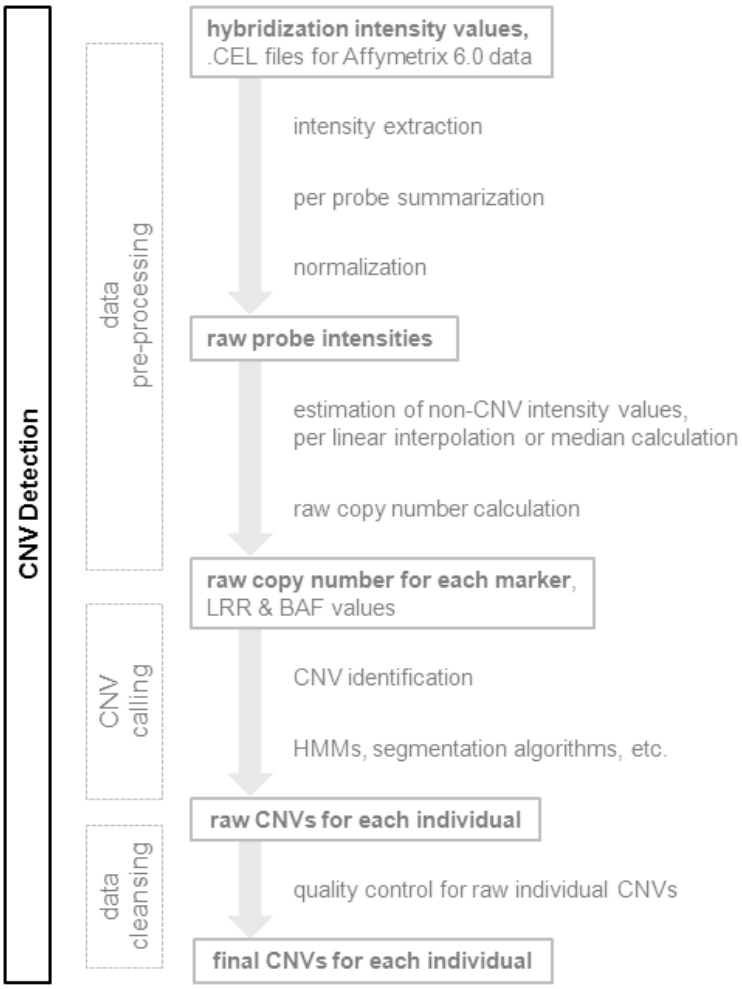

Figure 3.1: Schematic representation of essential steps in CNV detection from SNP array data.

Several recent publications (Dellinger et al., 2010; Koike et al., 2011; Winchester et al., 2009; Zhang et al., 2011) have addressed the question to what extent accuracy of CNV identification depends on the detection program in use and the respective applied parameters. Consistently, programs based on Hidden Markov models, such as PennCNV, QuantiSNP or Birdsuite, were shown to have a better detection performance than other programs with regard to several criteria: Compared to others, HMM-based tools yielded higher reproducibility rates across multiple different arrays of the same individual as well as lower Mendelian inconsistency rates in trio

data (Koike et al., 2011). The recovery rates, i.e. the ability to call CNVs that were previously identified by paired-end sequencing of whole-genome fosmid clones or by aCGH and additional validation procedures, were higher for HMM-based tools, and also positive predictive values of qPCR validated rare CNVs were shown to be higher in comparison to those of alternative tools (Zhang et al., 2011). In simulation studies, QuantiSNP outperformed other methods based on ROC curve residuals over most considered data sets (Dellinger et al., 2010).

In general, comparative analyses of different CNV detection tools demonstrated that there is currently no faultless software available for CNV identification and that academically developed tools seem to be more sensitive and to detect more events than proprietary algorithms (Winchester et al., 2009). Two previous studies measured the similarity of CNVs, being detected in use of each program, as the pairwise sensitivity between programs: The mean observed similarity of detected CNVs across different tools equaled approximately 40% (Koike et al., 2011; Winchester et al., 2009) with a range of 0 to 100% when considering CNVs of only one individual (Winchester et al., 2009) and a range of 4 to 75% for CNVs that were identified in publicly available data from a total of 270 individuals in the International HapMap Project [http://hapmap.ncbi.nlm.nih.gov/] (Koike et al., 2011).

One striking difficulty when judging the quality of CNV detection algorithms is the lack of a gold standard which could be used for the calculation of validity measures like sensitivity or specificity. Theoretically, such a gold standard would reflect the biological presence of CNVs in the genome of certain specified individual for which genetic marker information is, at best, publicly available. In fact, SNP array data for 270 individuals of different ancestry, who were analysed in the context of the International HapMap Project, is accessible to the public (e.g. via download from affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx in case of Affymetrix 6.0 data). However, there is currently no consensus on the set of truly underlying CNVs of those HapMap individuals. Up to now, several technical methods for genotyping of CNVs, such as paired-end sequencing (Korbel et al., 2007; Kidd et al., 2008), by use of Mendelian inconsistent SNP genotypes (Conrad et al., 2006), tiling arrays (Redon et al., 2006; Conrad et al., 2010), aCGH (Perry et al., 2008) and massively parallel sequencing (Park et al., 2010), were applied to all HapMap persons or to a subset of selected HapMap individuals, respectively. Depending on the resolution of the respective technological procedure, several differently composed CNV sets have been published by now. When considering the set of CNVs that were detected for one particular individual (NA15510) in three different studies (Kidd et al., 2008; Korbel et al., 2007; Redon et al., 2006), only

43 CNV events were found in all three studies out of a total of 174, 352 and 11 CNVs being reported in each single study, respectively (Winchester et al., 2009). Consequently, sensitivity of CNV identification software can either be presented relative to a certain CNV detection study, or it can alternatively be given in the form of a range being calculated with respect to results of several studies. Two recent publications, both focussing on a comparison of CNV detection tools, consistently report a mean sensitivity of approximately 15%, with a range from 0.1% to 54%, when all CNV detection algorithms where considered with respect to several other experimental results (Winchester et al., 2009; Koike et al., 2011).

3.3 Association Analyses of CNV Data

Structural variants, such as CNVs, can have an influence on phenotypes (Feuk et al., 2006a). For example, CNVs can modify drug response. Furthermore, CNVs that encompass or overlap a disease-associated gene can predispose to or cause disease in the current or in the next generation. Deletions, duplications or insertions of dosage-sensitive genes lead to reduced or increased gene expression, which can cause diseases. Dosage-insensitive genes can also cause disease, for instance in case that a deletion unmasks a recessive mutation on the homologous chromosome. Additionally, insertions or CNV start and end points can disrupt gene structure or can lead to formations of new transcripts through gene fusion or exon shuffling. In the proximity of dosage-sensitive genes, CNVs can alter gene expression through positive or negative effects. A deletion of important regulatory elements can, for instance, down-regulate gene expression, or a deletion of a functional element could unmask a functional polymorphism within an effector with consequences for gene function. Additionally, as susceptibility alleles in combination with several other genetic factors, CNVs can affect complex phenotypes or a complex disease state (Feuk et al., 2006a).

The first empirical evidence that CNVs can be associated with human phenotypes, came from the observation that sporadic cases of autosomal dominant diseases, like the Charcot-Marie-Tooth disease type 1A (CMT1A) or hereditary neuropathy with liability to pressure palsy (HNPP), result from de novo CNV events (Inoue and Lupski, 2002; Lupski, 2007). Diseases that are caused exclusively by genome structural changes are referred to as 'genomic disorders' (Lupski, 2007). Until very recently, the impact of structural genetic variation was thought to be limited to rare genomic disorders (Ionita-Laza et al., 2009). However, it is now known that CNVs are not exclusively present in patients with sporadic diseases, but that there exists

also widespread common structural variation among unaffected individuals (Conrad et al., 2010). Consequently, the question of whether CNVs influence more common complex human diseases has been addressed and positively answered in several studies, for instance for asthma (Brasch-Andersen et al., 2004) or schizophrenia (Walsh et al., 2008).

3.3.1 Strategies for Genome-Wide Association Testing of CNV Data

As depicted in Figure 3.2, two main strategical approaches can be applied for CNV association analyses (Ionita-Laza et al., 2009): approach S1 and approach S2. A brief summary of the main principles of both approaches is additionally presented

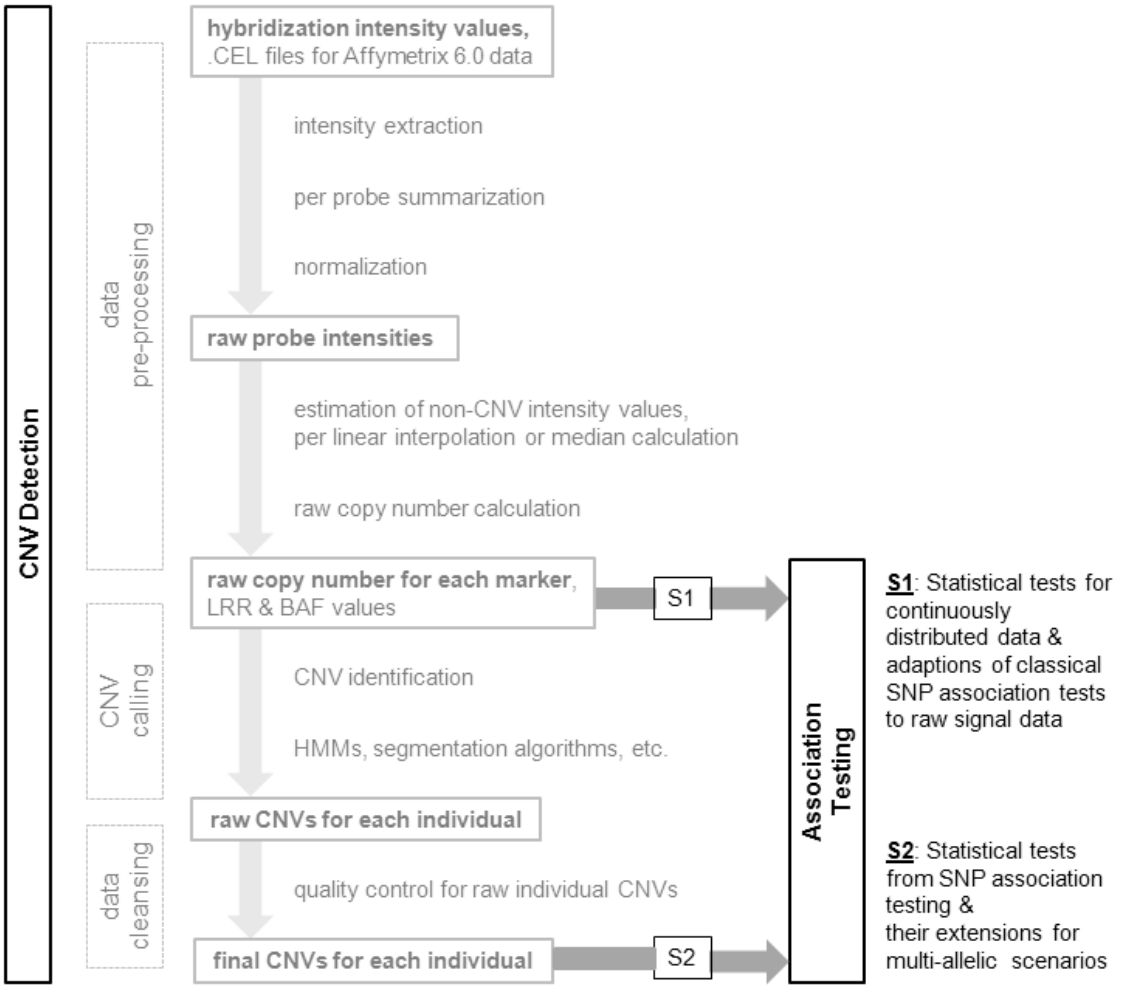

Figure 3.2: Schematic representation of the two main approaches for CNV association analyses.

in Table 3.5. In the following section, approach S1 and S2 will be embedded as key issues into complete analyses strategies.

Approach S1: One methodology aims to avoid the problems that are connected with accuracy in CNV genotype calling. Instead, raw copy number values, which are assumed to reflect the true unknown copy number states, are directly provided for statistical testing. According to the study design, classical statistical methods for continuously distributed data, such as the parametric t-test or the non-parametric Mann-Whitney U test for case-control designs, can be applied. Moreover, methods from SNP association testing can be adapted to the more general scenario of raw signal data when family-based designs are considered. Notably, the approach S1 is limited in the sense of biological interpretation and prognostic relevance.

Approach S2: The second methodology is composed as a two-step procedure. Firstly, CNV genotypes are inferred and afterwards these are incorporated in classical tests of association, which are well explored in the course of previous and ongoing widely-used association analyses accounting for phenotype-association effects of SNPs. Consequently, the performance of the approach S2 depends on the ability to genotype CNVs. Currently, there is no consensus about whether and how to incorporate the uncertainty of CNV genotype calling in the analysis. Moreover, due to the discretisation of continuous CNV copy numbers into CNV copy number classes, substantial information and statistical power may be get lost (Ionita-Laza et al., 2009). Finally, when considering CNVs that harbour more than 2 alleles, the standard association tests provided from SNP association studies are not directly applicable and have to be extended to the multi-allelic scenarios instead.

Table 3.5: Two main approaches for genome-wide statistical association testing of CNV data.

	Genome-wide Statistical Testing Approach S1	**Genome-wide Statistical Testing Approach S2**
Statistical testing is based on ...	individual raw copy number values for each available CNV probe set.	individual CNVs.
Statistical tests are performed at ...	any available CNV probe set.	any genomic region offering at least one CNV.
The number of tests performed equals the ...	number of available CNV probe sets.	number (sub-) CNV regions.

Concerning approach S2, several overlapping individual CNVs with similar but not equal start and end points are often assumed to differ mainly due to technical inaccuracy of the genotyping platform, and are thus transformed into CNVs of exactly the same length and location. Consequently, in this situation it suffices to perform statistical association testing at the CNV level and not at the marker level as it is done when applying approach S1 (see Table 3.5).

3.3.2 Case-Control Association Testing with CNV Data

One of the most simple and commonly used approaches for genetic association analyses is the case-control design, which incorporates one group of unrelated affected probands and another group of unrelated unaffected individuals. One apparent problem in genetic case-control association studies is to distinguish association findings that are based on true biological effects from those that are caused by the underlying structure of the population from which cases and controls were selected.

Depending on the selected methodology for association testing, parametric (for example the t-test) or alternatively non-parametric statistical tests (for example the χ^2 test, the trend test or the Mann-Whitney test) to test for differences in the frequencies of the different copy number classes or for differences in the distributions of continuous raw copy number signals between both groups are applied to the case-control setting.

3.3.3 Family-Based Association Testing with CNV Data

Approaches in which classical genetic association testing is performed within families offer protection from population stratification effects, but this typically comes at the cost of reduced power relative to case-control scenarios. However, simulations showed that the differences between case-control designs and trio designs are generally small in terms of statistical power when the number of trios is assumed to equal the number of cases and likewise the number of controls. Contrarily, with respect to unbalanced case-control scenarios accounting for considerably more controls than cases (i.e. ratio > 3), the number of trios needed to achieve comparable statistical power is substantially increased relative to the number of unrelated cases (McGinnis et al., 2002). On the other hand, trio designs were shown to be more powerful than case-control scenarios when the disease under study is rare (i.e. disease prevalence $\approx 0.1\%$) (Laird and Lange, 2006).

In particular, association testing of CNVs with disease can also in family data be performed on the basis of integers that reflect the individual biological copy numbers

or alternatively on the basis of continuous measurements that approximate the true CNV states. The classical Family-Based Association Test (FBAT) from the SNP analysis setting, has recently been generalized for the application to continuously distributed CNV data (Ionita-Laza et al., 2008).

Compared to case-control studies, family-based designs additionally offer the chance to determine the inheritance status of childish CNVs and to subsequently test for potential relationships with the disease status. However, the assignment of the underlying heritability is frequently problematic, since allelic copy numbers cannot always unambiguously be defined on the basis of the estimated diploid copy numbers. For example, a copy number state of two, i. e. a total number of two copies, could either represent a 1/1 or a 2/0 genotype. Alternatively, CNV data can be regarded as a quantitative trait, which allows the heritability of all types of CNV to be adequately demonstrated (Locke et al., 2006).

3.4 Two Existing Strategies for the Genome-Wide Analysis of Raw CNV Data

Mainly driven by the respective approach for genome-wide association testing, a distinction is made between two main strategies for the whole genome-wide association analysis of raw CNV data. For simplicity, the two existing strategies will be referred to as strategy S1 and S2 in the following. Both complete strategies are shown graphically in Figure 3.3.

Most parts of both strategies were already explained in the previous chapters. In order to complete a genome-wide CNV analysis strategy, a third validation step is usually added to the first two steps of CNV detection and CNV association testing. Due to the previously mentioned technological uncertainties, both strategies, S1 and S2, typically end up with validation experiments and corresponding follow-up analyses accounting for at least those genomic regions with initially observed statistically significant association test results. That is, individual copy number states of interesting (i.e. statistically significant) findings need to be re-determined in the present sample by use of an alternative technological method and subsequently association tests have to be re-evaluated based on the obtained validated CNVs.

In CNV analyses, a distinction is often made between focussing on rare or on common CNVs that appear with a certain frequency (for example $> 5\%$) in a population (McCarroll and Altshuler, 2007). Due to their sparseness, rare CNVs are commonly suggested to be grouped, based on pre-specified criteria, before being tested. Additionally, permutation procedures are explicitly recommended to test for

association of rare CNVs with disease. In order to illustrate the influence of CNV frequency on the design of the genome-wide analysis strategy, real data examples with a special focus on common as well as on rare CNVs will be presented in the following sections.

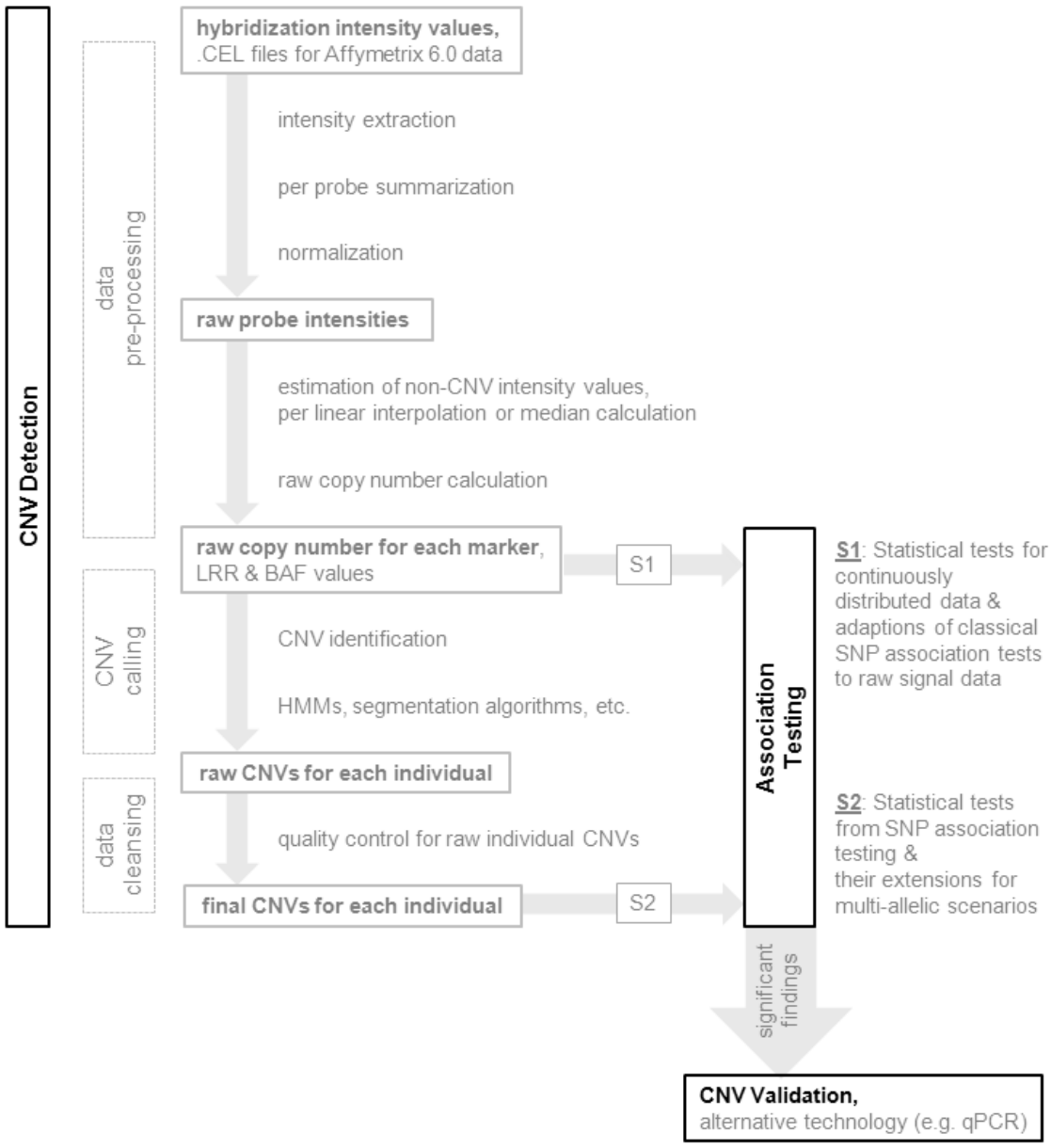

Figure 3.3: Schematic representation of the two main existing strategies for the genome-wide association analysis of raw CNV data.

3.5 Application of Strategy S1 to the Phenotype Obesity

With the dramatical growth of common obesity among adults as well as with the parallel less extreme growth among children, the interest in explaining the origins of the complex trait obesity has increased drastically in the last years. Apart from the obvious impact of environmental factors, common obesity undoubtedly has a large underlying genetic component (Walley et al., 2009). There are several classical twin studies in obesity that have revealed average heritability estimates of $40-75\%$ (Maes et al., 1997) for weight (Stunkard et al., 1986a,b) and for BMI (body mass index) (Turula et al., 1990; Wardle et al., 2008) in both adults and children. Consequently, approximately half of the inter-individual difference in body weight was shown to be explicitly caused by genetic variability.

Nevertheless, in several recent large scale meta-analyses, which incorporated up to $\sim 250\,000$ individuals, only a small proportion of genetic variants that are causal for common obesity could be detected so far (Willer et al., 2009; Speliotes et al., 2010). In sum, by examining associations between BMI and SNPs, such large population-based studies revealed a total of 42 variants at independent genetic loci that were estimated to collectively account for only 1.45% of the variance in BMI, which corresponds to approximately $2-4\%$ of the genetic BMI variance.

On the one hand, the residual variance in BMI with genetic cause may potentially be explained by a variety of further SNPs with even more little effect sizes, which might be detected in samples with considerably greater sample size. On the other hand, rare single base variants with comparably great association effects, which are undetectable in application of the mostly used SNP genotyping arrays, are assumed to make up a large proportion of the still unexplained genetic variability of complex traits, such as obesity (Manolio et al., 2009). Finally, structural genetic variants of type other than SNPs or single point mutations, as for instance CNVs, are another potential source of causal genetic variants (Manolio et al., 2009).

In this chapter, application of strategy S1 to genome-wide raw CNV data of a family-based obesity sample will be presented. The question whether specific common CNVs that are detectable by use of SNP arrays might be a genetic cause for the trait obesity will be further addressed in later parts of this thesis. In the following chapter, application of strategy S2 to the phenotype obesity will be presented. Moreover, in chapters 5.2 and 6.3, the proposed strategies PS1 and PS2 will additionally be applied to the same data set. Finally, advantages, disadvantages and limitations of each implemented CNV analyses strategy will be comparably discussed in detail.

3.5.1 Data Set

The family-based obesity sample was made up of 424 trios, each one consisting of one obese child or adolescent and both biological parents. According to a German reference population (Hebebrand et al., 1994), the measured BMI (in kg/m^2) of each offspring was above the 90^{th} age- and sex-specific percentile. Most of the index cases (93.6%) were extremely obese with a BMI percentile $\geq 97^{th}$. Details on phenotypical characteristics of the obesity trio sample can be found in Jarick et al. (2011) (Supplementary Table S1).

For follow-up analyses, an independent case-control sample comprising 453 obese children or adolescents and 435 normal weight or lean adult controls was adducted. All obese subjects (cases) of the case-control sample had a BMI above the 90^{th} age- and sex-specific percentile. 92.5% cases of the case-control sample were extremely obese with a BMI percentile $\geq 97^{th}$. Phenotypical characteristics of the case-control sample can be found in detail in Jarick et al. (2011) (Supplementary Table S1).

For all 1 272 individuals of the obesity trio sample as well as for all 888 individuals of the case-control sample, genotyping was performed on the Affymetrix Genome-Wide Human SNP Array 6.0 by the ATLAS Biolabs GmbH (Berlin, Germany). Besides $\sim$ 900 000 SNP probe sets ($\sim$ 870 000 autosomal), this SNP genotyping array additionally contains $\sim$ 940 000 non-polymorphic probe sets ($\sim$ 890 000 autosomal) for copy number analyses. More details on the Affymetrix 6.0 chip can be found in chapter 3.1.2. The genotyping procedure was exactly identical for each three members of any trio, i.e. their DNA material was pipetted to the same micro-plate and their hybridization intensity levels were assigned within the same batch.

3.5.2 Methods

Data pre-processing

Corresponding to strategy S1 (see Figure 3.3), statistical association testing was based on the fluorescence hybridization intensities at the autosomal non - polymorphic copy number (CN) probe sets that are a measure of copy number variance. For each individual and each probe set, raw intensity values were extracted from the individual '.CEL' files by use of the R-package 'affxparser' (Bengtsson et al., 2008a). Afterwards, the FBAT approach was genome-wide applied to the family-based obesity sample, that is to each of the 888 023 autosomal CN probe sets, in order to test the locus-specific CNV characteristics for an association with obesity.

Association testing

Since the offspring's genotyping procedure was identical to those of the parents, inter-familial differences in hybridization intensity measurements should solely be derived from CNV inheritance or from de novo CNV events, but not from technical artefacts. Applied to a binary trait, the FBAT is equivalent to a score test with a test statistic equaling the standardized sum of within-family components (see chapter 5.1.2.1 for a detailed description). Hence, a normalization of raw intensity values prior to the association testing is dispensable here. Consequently, the raw hybridization intensity measurements were directly tested without being transformed into raw copy number measurements (see chapter 6.1.1). The latter makes use of the fact that the FBAT approach is invariant under linear transformation (see chapter 5.1.2.1).

To account for multiple comparisons in testing multiple hypotheses (n = 888 023), the empirical Bayes method of local false discovery rates (lfdr) as proposed by Efron et al. (2001) was applied. The lfdr approach is motivated by the tail area false discovery rate (FDR), which was introduced by Benjamini and Hochberg (1995) in a frequentist framework. When a collection of hypotheses is tested simultaneously, the FDR equals the expected proportion of erroneously rejected null hypotheses among all rejected null hypotheses using a given rejection rule (Benjamini and Hochberg, 1995). Closely connected to a local version of the FDR, the lfdr is defined as the posteriori probability that a single null hypothesis is true given the observed value of the respective test statistic (Efron et al., 2001; Efron, 2004, 2007a,b).

In more detail, the lfdr method is based on a Bayesian two-class model that divides all test cases into two classes, 'null' or 'non-null', corresponding to whether or not they are generated according to the null hypothesis and with prior probabilities p_0 and $p_1 = 1 - p_0$, and with associated test statistic densities f_0 and f_1. The test statistic density f can then be written as a mixture density $f = p_0 f_0 + p_1 f_1$. According to the Bayes theorem, the lfdr for an observed test statistic value z is given as the posteriori probability $\text{lfdr}(z) = \mathbb{P}(\text{'null'}|z) = p_0 f_0(z)/f(z)$.

Using the 'locfdr' R package (Efron et al., 2011), lfdr estimates were obtained on the basis of empirical non-parametric estimates using central matching for the null distribution density $\hat{f}_0$, the mixture density $\hat{f}$ and the factor $\hat{p_0}$. Thus, the applied lfdr methodology especially accounts for the fact that the null sub-density f_0 might differ from the theoretical null distribution. The natural choice for f_0 would be the standard $\mathcal{N}(0,1)$ density in the underlying FBAT context with test statistics $\sqrt{FBAT_k}, k = 1, \ldots, 888\,023$ (CNV FBAT z-values, cf. equation (5.4) in chapter 5.1.2.1). However, a deviation from the theoretical null distribution can

be caused by several reasons that are listed by Efron (2004, 2007a,b): (1) failed assumptions on the test statistic, (2) unobserved covariates, (3) correlation across probes and genes, (4) a large proportion of genuine but uninterestingly small effects. As shown by Efron (2004, 2007a,b), in all these situations the application of the inappropriate theoretical null results in misclassified FDR and lfdr estimates. Of note, Efron (2007a) emphasized that even if the theoretical null is singularly appropriate for each probe-wise test situation, correlation across probes can lead to an effectively deviated null distribution f_0 compared to the theoretical $\mathcal{N}(0,1)$. Moreover, Efron (2004) point out that the popular permutation methods, which provide a way of avoiding assumptions on an underlying correlation stucture and asymptotic approximations (like normality), do not automatically resolve the question of an appropriate null hypothesis f_0. As shown by Efron (2004), unobserved covariates such as personal characteristics of the analysed study patients (i.e. age, gender or geographical location) are likely to widen or narrow the empirical f_0, and this effect is not detectable in permutationally derived null hypothesis. Efron (2004) point out that a permutation null distribution will not reveal correlation effects of hidden covariates, but will closely match the theoretical null distribution, irrespective of whether or not there are unobserved covariates or other factors influencing the theoretical null distribution. Finally, results of each inference method, FDR, lfdr, Bonferroni, family-wise error rate (FWER), are doubtful if the null hypothesis is not chosen appropriately. Efron (2004) strongly argues to prefer the empirical null hypothesis in observational studies.

Evaluation of statistically significant results, CNV calling and follow-up analyses

As stated by Ionita-Laza et al. (2008), it is challenging to evaluate whether statistically significant association test results are caused by underlying CNV - trait associations or rather by hybridization intensity differences depending on probe-specificity and signal-to-noise properties of the platform, when the CNV FBAT methodology was applied genome-wide without an a priori selection of markers. To address this concern, the HMM CNV detection algorithm implemented in the Affymetrix Genotyping Console (GTC) 3.0 was employed on the raw intensity data. For each of the 1 272 individuals of the family-based obesity sample, CNV calls were estimated by comparing individual signal intensities against a reference sample. Due to computational constraints, the reference sample size was limited to 106 parental pairs of the obesity trio sample. In order to minimize the potential effect of the choice of the reference sample on the CNV calling results, two differently composed reference samples, each comprising 106 parental pairs, were used. One reference

sample (ref_1) was a random collection, whereas the other one (ref_2) was based on those parental pairs with the lowest mean BMI standard deviation scores out of all non-obese parental pairs. Phenotypical details on both reference samples can be found in Jarick et al. (2011) (Supplementary Table S1).

For probe sets with statistically significant CNV FBAT results, follow-up analyses were performed in the case-control sample. Significance of CNV FBATs was determined with respect to a lfdr level of 0.2, which was proposed to be a sensible threshold by Efron (2004). To address potential plate effects, quantile normalization (Bolstad et al., 2003) was applied to the raw intensity signals of the case-control sample. Subsequently, logistic regression with predictors normalized intensities, sex and age was used to test the CN probe sets for an association with obesity.

3.5.3 Results

For the Affymetrix 6.0 chip, a total of 888 023 CN probe sets are available for copy number analysis. Genome-wide CNV FBAT results for the analysis of 424 obesity trios are summarized in form of a Manhattan plot, which is depicted in Figure 3.4. Non of the tested probe sets reached genome-wide significance at a relevant significance level when correction for multiple testing would be performed in a Bonferroni manner (minimal CNV FBAT p-value $= 1.06 \times 10^{-4}$). However, due to correlations across probes, a Bonferroni correction, which assumes simultaneous testing of independent hypotheses, is not appropriate for the underlying situation.

Contrarily, the lfdr method that does not require stochastic independence between probes, yielded 23 probe sets with lfdr values below 0.20 (Figure 3.5, Table 3.6), which is a sensible threshold as proposed by Efron (2004). As shown in Figure 3.5, the empirical null distribution, $\mathcal{N}(-0.182, 0.576^2)$, strongly deviates from the theoretical null, $\mathcal{N}(0, 1)$. However, given the genome-wide correlation structure across probes, this is a non-surprising and well known phenomenon (Efron, 2004; Efron et al., 2001; Eyheramendy et al., 2011).

The majority of lfdr significant probe sets (n = 13) is located in regions (human genome version 18, hg18) without reported CNVs in the Toronto Database of Genomic Variants (DGV) (Table 3.6). For approximately half of the residual ten probe sets (n = 4), no copy number variability was estimated in the affected offspring of the analysed sample of 424 obesity trios. The most promising association test results, that is those with lowest CNV FBAT p-values, showed either no evidence for CNV variability with regard to the DGV or to sample-based estimates of CNV frequencies.

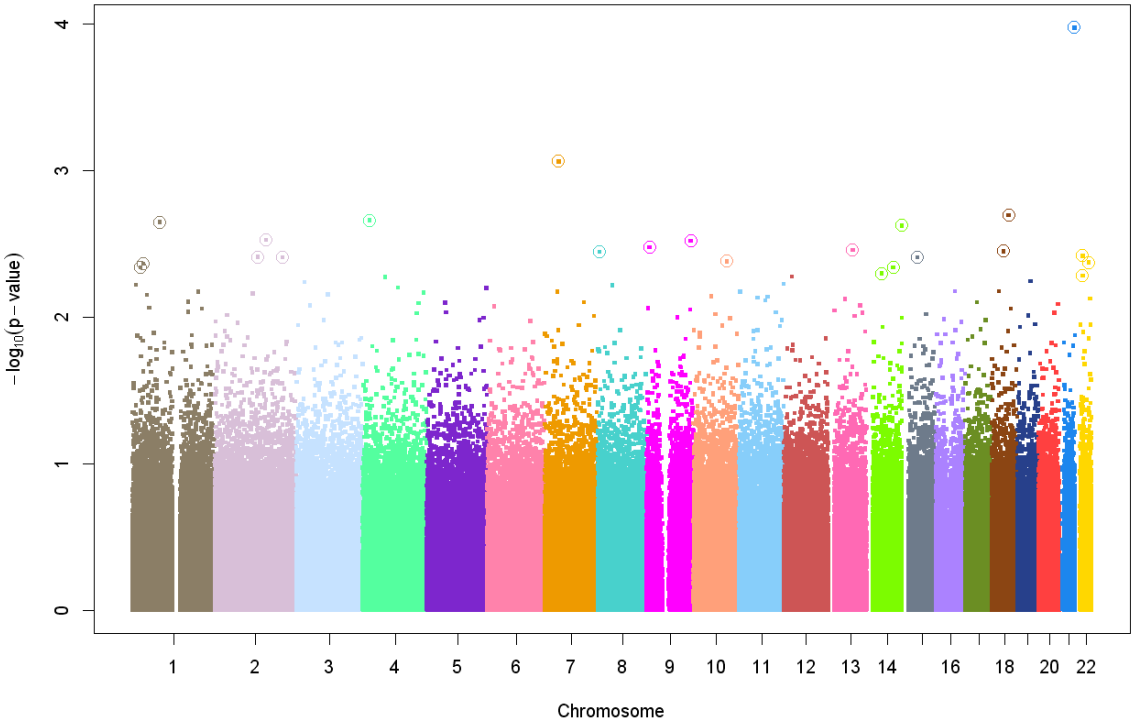

Figure 3.4: Manhattan plot for the genome-wide CNV analysis of 424 obesity trios accounting for 888 023 CN probe sets. For all CN probe sets, the $-\log_{10}$ transformed CNV FBAT p-values are shown relative to their chromosomal position. 23 FBAT results with lfdr < 0.2 are circled.

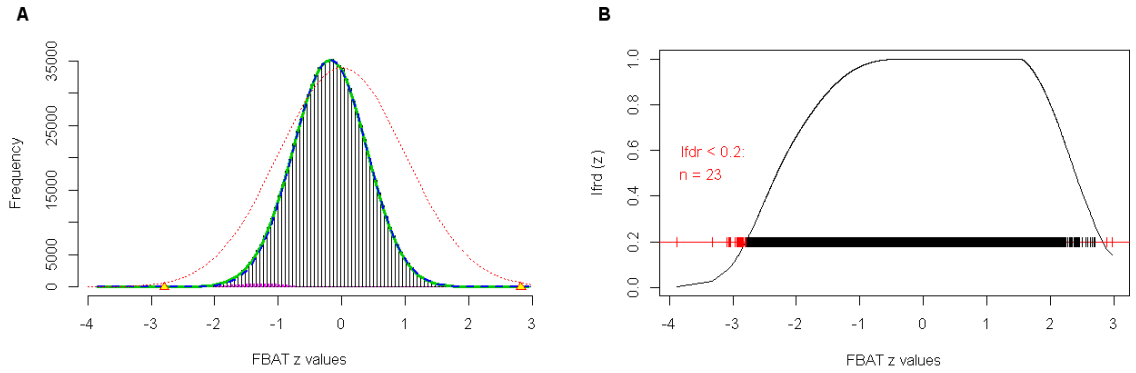

Figure 3.5: Histogram and lfdr curve of CNV FBAT z-values for the genome-wide analysis of 424 obesity trios accounting for 888 023 CN probe sets. Panel A: Histogram. The red dashed curve depicts the standard normal distribution, the dashed blue line is $\hat{p}_0\hat{f}_0$, the empirical null density, $\mathcal{N}(-0.182, 0.576^2)$, and the green line is the empirically estimated mixture density. The small pink bars represent estimated non-null counts. Panel B: Lfdr curve, derived from empirical estimates of f_0, f and p_0 (Panel A). Observed CNV FBAT z-values are illustrated as ticks on the horizontal line at lfdr level 0.2, those with lfdr < 0.2 are printed in red.

Table 3.6: Results for 23 probe sets with lfdr < 0.2 across 888 023 CNV FBATs, each accounting for 424 obesity trios. (Results are sorted by FBAT z-values. Each CN probe accounts for 25 bp.)

CN probe set ID	**Chr: Position [hg18]**	**FBAT z-value**	**FBAT p-value**	**lfdr**	**known CNV [DGV]**	**CNV frequency in offspring [%], ref_1 // ref_2**	**CNV frequency in parents [%], ref_1 // ref_2**	**logistic regression z-value**	**logistic regression p-value**
CN_895557	$21 : 44\,716\,718^{+}$	−3.877	0.0001	0.001	YES	0 // 0	0 // 0	1.518	0.1291
CN_1256115	$7 : 41\,769\,452^{+}$	−3.330	0.0009	0.025	-	0 // 0	0 // 0	1.051	0.293
CN_792740	$18 : 50\,771\,860^{+}$	−3.087	0.0020	0.075	-	0 // 0	0 // 0	0.293	0.7697
CN_1071025	$4 : 20\,652\,044^{+}$	−3.063	0.0022	0.082	-	0 // 0	0 // 0	1.854	0.0638
CN_502773	$1 : 82\,317\,273^{+}$	−3.055	0.0023	0.085	-	0 // 0	0 // 0	1.060	0.2893
CN_648814	$14 : 105\,520\,700^{+}$	−3.038	0.0024	0.090	YES	0.71 // 0.71	1.65 // 2.00	-0.423	0.6722
CN_819604	$2 : 153\,658\,655^{+}$	−2.971	0.0030	0.115	-	0 // 0	0.24 // 0.24	1.978	0.0479
CN_1325083	$9 : 8\,427\,225^{+}$	−2.934	0.0034	0.131	-	0 // 0	0 // 0	-1.401	0.1611
CN_635265	$13 : 73\,536\,303^{+}$	−2.920	0.0035	0.137	-	0 // 0	0 // 0	-0.339	0.7345
CN_779388	$18 : 36\,203\,193^{+}$	−2.914	0.0036	0.140	-	0 // 0	0 // 0	0.578	0.563
CN_369358	$8 : 4\,784\,964^{+}$	−2.912	0.0036	0.140	YES	0.24 // 0.24	0.12 // 0.12	0.229	0.8191
CN_915257	$22 : 22\,661\,955^{+}$	−2.892	0.0038	0.150	YES	2.36 // 3.54	2 // 2.59	0.245	0.8063
CN_663949	$15 : 45\,351\,560^{+}$	−2.886	0.0039	0.153	YES	0.24 // 0.24	0 // 0	0.456	0.6485
CN_816331	$2 : 202\,126\,216^{+}$	−2.886	0.0039	0.153	YES	0 // 0	0.12 // 0.12	-2.401	0.0164
CN_540526	$10 : 99\,065\,374^{+}$	−2.865	0.0042	0.163	YES	0 // 0	0.12 // 0	1.840	0.0657
CN_913545	$22 : 41\,296\,360^{+}$	−2.859	0.0042	0.165	YES	0.71 // 1.18	1.06 // 1.30	-0.765	0.4444
CN_480122	$1 : 32\,876\,668^{+}$	−2.853	0.0043	0.169	-	0 // 0	0.12 // 0.12	1.055	0.2916
CN_484362	$1 : 25\,606\,827^{+}$	−2.834	0.0046	0.178	-	0.47 // 0.47	0.12 // 0.12	-1.216	0.2239
CN_678173	$14 : 80\,138\,675^{+}$	−2.834	0.0046	0.178	-	0 // 0	0.35 // 0.24	-2.643	0.0082
CN_679595	$14 : 45\,136\,094^{+}$	−2.804	0.0051	0.194	YES	0 // 0	0 // 0	2.563	0.0104
CN_915256	$22 : 22\,660\,257^{+}$	−2.793	0.0052	0.199	YES	2.36 // 3.54	2 // 2.59	0.375	0.7078
CN_819097	$2 : 128\,655\,731^{+}$	2.886	0.0039	0.167	-	0 // 0	0 // 0	-2.386	0.0170
CN_1297053	$9 : 131\,449\,649^{+}$	2.966	0.0030	0.144	-	0.71 // 0.71	0.24 // 0.24	-0.990	0.3221

Two of the 23 lfdr significant CN probe sets, CN_816331 and CN_678173, showed an effect direction in logistic regression testing that was identical to CNV FBAT, and additionally yielded a nominal significant logistic regression p-value at a significance level of five percent. None of these two logistic regression test results remained significant after correction for simultaneously testing 23 hypotheses, in neither a Bonferroni nor a lfdr manner.

3.5.4 Discussion

Strategy S1 for the genome-wide analysis of raw CNV data was applied to a family-based sample comprising 424 extremely obese children or adolescents and their biological parents. For association testing, the CNV FBAT methodology as proposed by Ionita-Laza et al. (2008) was adopted. Significance was assessed via the lfdr method that was introduced by Efron (2004). Out of 888 023 tested CN probe sets, 23 achieved genome-wide significant CNV FBAT results. However, a majority of 17 probe sets were located in genomic regions without evidence for copy number variability, neither with respect to previous reports nor to estimates based on the present sample. The logistic regression framework was applied to a case-control sample of 453 obese children and adolescents and 435 lean adult controls for follow-up analyses of significant findings from the family-based sample. Two probe sets showed nominally significant and directionally consistent test results. None of the latter probe sets reached statistical significance after correction for multiple testing.

On the one hand, strategy S1 is easy to implement with regard to the fact that raw hybridization intensity measurements can directly be tested for an association. Since each available array probe set is incorporated into the association testing step without any prior pre-selection, the most complex, work intensive and time consuming part of strategy S1 is the data pre-processing step.

On the other hand, the interpretation of association test results obtained from strategy S1 might be challenging due to the fact that biological plausibility is not necessarily a priori provided. To address this issue in the analysed obesity data set, underlying copy number variability of significantly associated probe sets was assessed with respect to sample based estimates and to estimates from a publicly available catalog of structural variants in healthy control samples (DGV). Out of 23 CN probe sets with significant CNV FBATs, a minority of six (= 26%) probe sets is located in genomic regions that were reported to be copy number variable in the DGV, and that were additionally estimated to be covered by individual, postentially disease causing CNVs in the analysed family-based obesity sample. Of note, the estimated sample copy number variability did not exceed 3.54% for any

of the six probe sets. For the other 17 probe sets, for which a lack of copy number variability was observed, there is no obvious reason to believe that the association test result reached significance due to the presence of underlying CNV association effects. Instead, significance may rather be caused by technical fluctuations for the latter probe sets. These doubts could be resolved in use of altervative technical approaches aiming to validate array based CNV calls or by follow-up analyses in independent obesity samples as it was done here. However, none of the findings could be strengthened by follow-up analyses in an independent case-control obesity sample. Consequently, in application of strategy S1 there is no hint for any association of CNVs with obesity.

It turned out that one of the main advantages of strategy S1, its easy, computationally simple and fast implementation at a genome-wide level without the necessity of previous knowledge on structure, genomic location or even existence of CNVs, simultaneously offers the potential for being one of its greatest weaknesses. Generally speaking, the effective impact of this aspect on genetic results might depend on the quality of raw genotyping data and primarily on those of pre-processed raw copy number data. In more detail, the better appropriate the assignment of individuals to genotyping plates and the more comparable the genotyping pipeline was transposed across batches, the lower is the risk of any bias due to technical concerns. Moreover, data pre-processing, such as normalization procedures, may additionally minimize potentially spurious results. However, particularly the CNV FBAT approach was designed as a robust approach against any such confounding. Keeping in mind that each three members of any considered obesity trio were pipetted to the same micro-plate, the normalization step was thus skipped here.

Another disadvantage of testing probe-specific raw copy number measurements for an association with the trait of interest without knowledge on underling CNVs is that follow-up analyses are canonically performed for significantly discovered probe sets. However, any potentially associated CNV is known to be described by a set of several consecutive array probes with highly correlated characteristics. Consequently, a higher power might be achieved in replication attempts that additionally account for surrounding probes of initial findings. However, in application of CNV analyses strategy S1 no knowledge is provided on how to appropriately extent the follow-up probe set clique concerning this issue. This lack of information might be addressed by a rudimentary CNV calling step, aiming to detect sample-specific CNV breakpoints, as it is proposed in the modified CNV analyses strategy PS2.

Finally, there is no way out of getting to know the true underling CNV genotypes of positive findings by validating the association signal independently by use

of a different technology. Otherwise, no insight can be provided into the biological mechanisms of how the identified CNV influences the disease of interest. The presented application of strategy S1 was finished without technical validation because of having found no positive CNV association effects.

3.6 Application of Strategy S2 to the Phenotype Obesity

In the following chapter, the application of strategy S2 will be demonstrated exemplarily to genome-wide raw CNV data of a family-based obesity sample. For this purpose, the same data set of 424 obesity trios, which has been analysed by use of strategy S1 in the previous chapter, will be re-analysed here. Subsequently, statistical results and genetic conclusions of strategy S2 for the phenotype obesity will be compared to those of strategy S1. Finally, strengths and weaknesses of the genome-wide CNV analyses strategy S2 will be discussed in comparison to adequate characteristics of strategy S1.

3.6.1 Data Set and Methods

Data set

Available genotype data for a family-based sample consisting of 424 nuclear families, each comprising one obese child or adolescent and both biological parents, has been analysed here. All families were previously recruited and phenotypically characterized through the Departments of Child and Adolescent Psychiatry of the Universities of Duisburg-Essen and Marburg. Details on phenotypical characteristics can be found in chapter 3.5.1 and in Jarick et al. (2011) (Supplementary Table S1). For all 1 272 individuals, genotyping was performed on the Affymetrix 6.0 chip by the ATLAS Biolabs GmbH in Berlin (for details see chapter 3.5.1).

CNV calling and association testing

For each of the 1 272 individuals, the CNV detection step was performed in application of the PennCNV software (Wang et al., 2007) by using default parameters. In the course of quality control (QC) for the CNV calling procedure, each CNV call that did not cover more than 20 informative consecutive probe sets was discarded from subsequent statistical analyses. As shown in later chapters, the CNV detection threshold of 20 probe sets per CNV call is the optimal threshold for Affymetrix 6.0

data with regard to CNV's stability and reproducibility rates. The remaining CNVs were tested for an association with the binary trait obesity by use of the FBAT approach with assuming an additive genetic effect model. In more detail, the coding for the different marker genotypes was specified as 0, 1, 2, 3, 4 in concordance with the estimated total unphased number of DNA segment copies. In order to avoid redundancies, only the set of unique CNV's start and end sites but not the whole set of available probe sets was tested for an association with the phenotype obesity. In more detail, overlapping CNVs were at first merged into several CNV containing regions (CNVRs). Secondly, each CNVR was divided into multiple sub-CNVRs. Here, the boundaries of each single sub-CNVR were defined to equal the breakpoints of the maximal intervals with identical CNV configuration across all 1 272 individuals. Thus, the composition of each single CNVR is completely specified by the set of all CNV's start and end sites of any individual CNV (see Figure 6.7 for details).

In order to allow each FBAT to account for a minimal number of informative families, only sites within 244 pre-specified genomic regions that offer a CNV variability of at least five percent in both, the offspring's and the parent's group, were incorporated into the association testing step. Details on how these 244 CNVRs were specified and on their structural characteristics are given in chapters 5.2.2 and 5.2.3. As previously explained in detail, genome-wide significance of simultaneously testing multiple hypotheses was assessed by use of the lfdr method (Efron et al. (2001), see chapter 3.5.2).

3.6.2 Results

A total of 47 796 CNVs were detected in the 1 272 individuals, 15 863 CNVs were observed in the offspring's group and 31 933 CNVs in the parent's group. Out of all detected CNVs, 39 955 CNVs were located in 244 pre-specified CNVRs with a minimal CNV variability of five percent, 13 455 in the offspring and 29 500 in the parents.

For association testing, FBATs were performed at a total of 3 525 unique CNV's start and end sites (Figure 3.6). None of the tested sites reached statistical significance after correction for testing multiple hypotheses (minimal p-value = 0.00071).

3.6.3 Discussion

Application of strategy S2 for the genome-wide analysis of raw CNV data, to a family-based sample of 424 obesity trios, did not reveal any evidence for an association of certain CNVs with the trait obesity. This is in concordance with previous

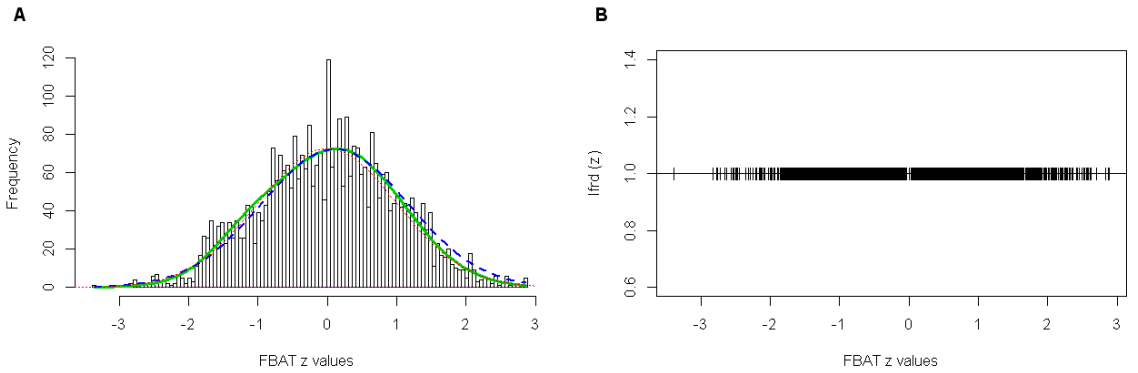

Figure 3.6: Histogram and lfdr curve of CNV FBAT z-values for the genome-wide analysis of 424 obesity trios at 3 525 unique CNV's start and end sites in 244 CNVRs. Panel A: Histogram. The red dashed curve depicts the standard normal distribution, the dashed blue line is $\hat{p_0}\hat{f_0}$, the empirical null density, $\mathcal{N}(0.107, 1.056^2)$, and the green line is the empirically estimated mixture density. Panel B: Lfdr curve, derived from empirical estimates of f_0, f and p_0 (Panel A). Observed CNV FBAT z-value are illustrated as ticks on the horizontal line at lfdr level 1.

results of applying strategy S1 to the same data set. Apart from a true lack of a CNV - obesity association, one potential cause for the negative finding might be seen in a power constraint, which may result from the moderate size of the analysed sample.

In contrast to strategy S1, in which statistical association testing is based on raw copy number measurements, a computational expensive CNV detection step is performed prior to the association testing when applying strategy S2. In the presented example for the phenotype obesity, the academically developed software tool 'PennCNV' was used in the CNV identification step. Compared to alternative software programs, this HMM based program was previously shown to perform comparably well in detecting CNVs from SNP genotyping array data (Winchester et al., 2009; Koike et al., 2011). As outlined in chapter 3.2, there is currently no consensus on the optimal choice of an algorithm or software for estimating individual CNV events with reliable accuracy. When following recent recommendations of using a second algorithm on a single data set to increase confidence in the CNV data (Winchester et al., 2009), the complexity and computing efforts of strategy S2 would even considerably be extended. However, in filtering CNVs by calling results from a second CNV calling software tool it would even become less likely to list all CNVs in a sample. Of note, no CNV tested for an association in strategy S2 can be taken for sure without separate biological validation or replication.

3.7 Application of Strategy S2 to the Phenotype ADHD on rare CNVs

With a worldwide-pooled prevalence of 5.2% (Polanczyk et al., 2007), attention - deficit / hyperactivity disorder (ADHD) represents one of the most common psychiatric disorders in children and adolescents. Although twin studies on ADHD in children and adolescents indicate a strong genetic component with heritability estimates around 75% (Freitag et al., 2010), neither genome-wide association studies nor large scale meta-analyses of GWASs have so far identified potential causal SNP variants (Hinney et al., 2011).

Conversely, genome-wide analyses of CNVs provide evidence that several CNVs might be associated with ADHD in children and adolescents (Elia et al., 2010; Williams et al., 2010; Lesch et al., 2011; Lionel et al., 2011). Especially large ($>$ 500 kb), rare ($<$ 1% frequency) CNVs were found with an increased rate in ADHD patients compared to healthy controls. More precisely, rare CNVs identified in children with ADHD were found to be preferentially located in several candidate regions, such as in the chromosome 16p13.11 region, along the *NPY* and the *CHRNA7* gene (at chr 15q13.3), in metabotropic glutamate receptor genes or in several neurodevelopmental genes, that is in genes reported as candidates in other neuropsychiatric disorders, such as schizophrenia, Parkinson disease or autism.

With regard to the mentioned previous findings, a genome-wide CNV association study was performed by investigating the hypothesis of 'common disease - many rare variants' (Mayo, 2007). Thus, contrary to the previous chapters that aimed to identify common CNVs with association effect to the trait obesity, the focus of this chapter will be on the genome-wide association analysis of rare CNVs with respect to the binary trait ADHD. As in the previous chapter, genome-wide raw CNV data will be analysed in application of strategy S2. Parts of the genetic results of this chapter have been published in Jarick et al. (2012).

3.7.1 Data Set

GWAS discovery sample.

In the discovery step, an available case-control GWAS sample of 489 ADHD patients and 1 285 population-based controls with high quality data was considered. All GWAS cases were previously assessed for the diagnosis ADHD according to DSM-IV (American Psychiatric Association, 1994) by the Department of Child and Adolescent Psychiatry of the University of Duisburg-Essen. The ADHD patients

are all German minors, who were recruited and phenotypically characterized in six psychiatric outpatient units for children and adolescents (Aachen, Cologne, Essen, Marburg, Regensburg and Würzburg). Details on the corresponding ADHD subtypes and basic characteristics are given in Jarick et al. (2012) (Table 1).

Additionally, the GWAS discovery sample consists of 1 285 adult controls that were not screened for ADHD, and that were previously drawn from three German population-based epidemiological studies in adults: (i) the Heinz Nixdorf RECALL (Risk Factors, Evaluation of Coronary Calcification, and Lifestyle) study (Schmermund et al., 2002) (n = 383), (ii) PopGen (Krawczak et al., 2006) (n = 490) and (iii) KORA (Wichmann et al., 2005) (n = 488). The recruitment areas were Western Germany (Essen, Bochum and Mühlheim) for (i), Nothern Germany (Schleswig-Holstein) for (ii) and Southern Germany (Augsburg) for (iii), respectively. The percentage of male GWAS controls was considerably smaller than the percentage of male GWAS ADHD cases (cases: 81.0% males, controls: 50.7% males). Moreover, GWAS controls were older than GWAS ADHD cases (age range cases: $6 - 18$ years, controls: $25 - 75$ years). Details on phenotypical characteristics can be found in Jarick et al. (2012), Table 1.

Genome-wide genotyping of the GWAS discovery sample was performed on the Illumina HumanHap550v3 for the controls group and on the Illumina Human660W-Quadv1 Bead Arrays for the cases group by (i) Illumina customer service, San Diego, CA, USA (all PopGen controls), (ii) the Department of Genomics, Life & Brain Center, University of Bonn, Germany (all ADHD cases and Heinz Nixdorf RECALL study controls) and (iii) the Helmholtz Center of Munich, Germany (all KORA controls). All subjects of both groups met our stringent pre- and post-calling quality control (QC) criteria (for details see the following Methods section and the Supplementary Text of Jarick et al. (2012)).

Replication sample.

For replication analyses of the findings from the GWAS discovery sample, a second available independent ADHD case-control sample consisting of 386 young German ADHD patients and 781 German population-based healthy young controls with high quality data was considered. The cases of the replication sample were previously recruited and phenotypically characterized in two outpatient clinics by and at the Departments of Child and Adolescent Psychiatry, Psychosomatic and Psychotherapy of the Universities of Homburg and Würzburg (Germany). In this context, an ADHD diagnosis was determined for all patients of the replication sample. Patients were only included here, if they were diagnosed with ADHD according to DSM-

IV (American Psychiatric Association, 1994), subtypes and basic characteristics are given in Jarick et al. (2012) (Table 1). Core descriptive statistics, such as the percentage of males (83.7%) or the age range ($6 - 19$ years) were comparable to those of the ADHD GWAS discovery sample.

In addition, a total of 1 063 controls were chosen from available data sets of two ongoing German population-based prospective birth cohorts: (i) the influence of Life-style factors on the Immune System and Allergies Plus environment and genetics (LISAplus) study (Zutavern et al., 2006) and (ii) the German Infant study on the influence of Nutrition Intervention Plus environment and genetics (GINIplus) study (Berg et al., 2010). Briefly, the two birth cohorts consist of healthy full-term newborns, who were recruited between September 1995 and January 1999 in Munich, Wesel, Leipzig and Bad Honnef, and who were follow-up until the age of ten. A detailed description of screening and recruitment has been provided elsewhere (Zutavern et al., 2006; Berg et al., 2010). Any control probands, for whom no questionnaire information was available (n = 111) or who were not previously categorized as being in the normal range at the age of ten on the Strengths and Difficulties Questionnaire on the scales for hyperactivity / inattention as well as on the total difficulties scale (n = 118), were excluded from replication analyses (n = 229) (for details see Jarick et al. (2012), Table 1).

For all subjects of the replication sample, ADHD cases and controls, genome-wide genotyping data of the Affymetrix Genome-Wide Human SNP Array 5.0 was available. Similarly to the ADHD GWAS discovery sample, all analysed subjects of the replication sample met stringent pre- and post calling QC criteria (for details see the following Methods section and the Supplementary Text of Jarick et al. (2012)).

3.7.2 Methods

CNV calling

Prior to any CNV analyses, a standard SNP-based QC procedure was applied to each recruited ADHD patient (n = 504) and control subject (n = 1 361) of the GWAS discovery sample. The cases and controls group of the GWAS discovery sample separately passed this pre-calling QC protocol, which accounts for: i) the genotyping quality, by claiming a SNP call rate $> 97\%$ for each individual, ii) the exclusion of subjects with discrepant sex status with regard to X-chromosomal heterozygosity rates, iii) the exclusion of cryptically related subjects, i.e. those with identical-by-state (IBS) values > 1.65 and iv) population stratification. Nine ADHD patients and 61 controls were excluded in the course of this first QC step.

CNV detection was based on those 535 364 autosomal SNPs, which are common to the Illumina HumanHap550v3 (controls group) and the Illumina Human660W-Quadv1 Bead Arrays (cases group) genotyping chip. CNVs were called in application of the PennCNV software (Wang et al., 2007) by use of default parameters. Afterwards, samples with low CNV genotyping quality were excluded based on the following two post-calling QC criteria: i) a high standard deviation (sd) in genome-wide hybridization intensity levels, i.e. sd(LRR) > 0.3 or ii) an implausibly high number of CNVs, i.e. more than 90 CNV calls. Six ADHD patients and 15 control subjects were excluded in the course of this second QC step. Moreover, CNVs with low expected recovery rates were excluded. Following former recommendations for the applied genotyping chips (Williams et al., 2010), the following CNVs were excluded: i) CNVs spanning less than 15 consecutive informative SNPs, ii) CNVs spanning known gaps of at least 200 kb in the SNP array by more than half of their lenght, and iii) CNVs in known segmental duplications present in the March 2006 human reference genome according to the Segmental Duplication DataBase (http://humanparalogy.gs.washington.edu/build36/build36.htm.). It is often observed that large CNVs are splitted into several CNV calls in application of CNV calling algorithms. Thus, iv) two individual adjacent CNV calls with size $>$ 200 kb and distance smaller than half of their entire length were merged into one CNV after appropriate visual evaluation of SNP-wise values.

Association testing

For association analyses, only rare CNV calls were investigated, that is only CNV calls with sample frequency $\leq 1\%$ for at least half of their length spanned regions. At each CNV locus, the hypothesis of an increased CNV frequency in ADHD cases compared with the CNV frequency in control subjects was tested by use of one-sided Fisher's exact tests. Additionally, the stratified hypotheses according to CNV type, that is an over-representation of deletions or duplications in cases versus controls, were tested. In order to avoid redundancies, association testing was limited to the set of unique CNV's start and end sites. Thus, overlapping CNVs were merged into several CNV containing regions (CNVRs), and each CNVR was devided into several sub-CNVRs, which are defined by the start and end positions of the involved individual CNVs (see Figure 6.7 for details). Genome-wide significance of each tested site was assessed via the permutation approach. That is, the genome-wide null hypothesis of no association of any CNV with ADHD was simulated by repeatedly (n = 100 000) permuting the individual's affection status and simultaneously preserving

the CNV's correlation structure. All statistical testing analyses were performed by use of the PLINK software (Purcell et al., 2007).

CNV validation and replication analyses

For the *PARK2* locus, for which CNVs were observed to be significantly associated with ADHD, qPCR experiments (TaqMan CNV assy HS03615859_cn at chr 6: 162 696 897 ± 50 bp, HCBI36/hg 18) to validate individual CNV states were performed by (i) the Department of Child and Adolescent Psychiatry of the University of Duisburg-Essen (GWAS cases) and (ii) the Department of Genomics, Life & Brain of the University of Bonn (HNR controls). Copy number states were determined for each individual for whom DNA was available, that is for GWAS cases and for a relevant subset of HNR controls (see Jarick et al. (2012) for details). Subsequently, genome-wide association testing was repeated on the basis of qPCR validated CNVs.

Moreover, the CNVR that showed statistically significant association in the GWAS discovery sample (at the *PARK2* locus), underwent follow-up analyses in the replication sample. Compared to the GWAS discovery sample, similar pre-calling and identical post-calling QC criteria were applied to all recruited ADHD patients (n = 461) and to all healthy control samples (n = 834) of the replication sample. In more detail, only individuals with i) a SNP call rate $< 95\%$ and ii) no disrepant sex status according to X-chromosomeal heterozygosity rates were considered in the CNV calling step (cases: n = 421, controls = 814). Based on PennCNV's CNV calling, individuals with i) sd(LRR) > 0.3 or ii) more than 90 CNV calls were excluded from the association testing step (cases: n = 35, controls = 33). Association testing in the replication sample was performed analogously to the GWAS discovery sample at sites, which were defined by CNV calls of individuals from the replication sample.

3.7.3 Results

In the 489 ADHD patients and 1 285 control subjects of the GWAS discovery sample, a total of 2 432 rare CNVs of high quality were identified (cases: n = 592, controls: n = 1 840). On average, each ADHD patient showed 1.2 rare CNVs with an average size of 226.3 kb (range: $9.3 - 2\,830.8$ kb) and each control subject was estimated to carry 1.4 rare CNVs with an average size of 186.4 kb (range: $5.6 - 4\,479.6$ kb).

There were no sex specific differences in genome-wide CNV rates between cases and controls (data not shown). Moreover, there was no evidence for a genome-wide burden of rare CNVs in ADHD patients compared to control subjects (for details see Jarick et al. (2012)).

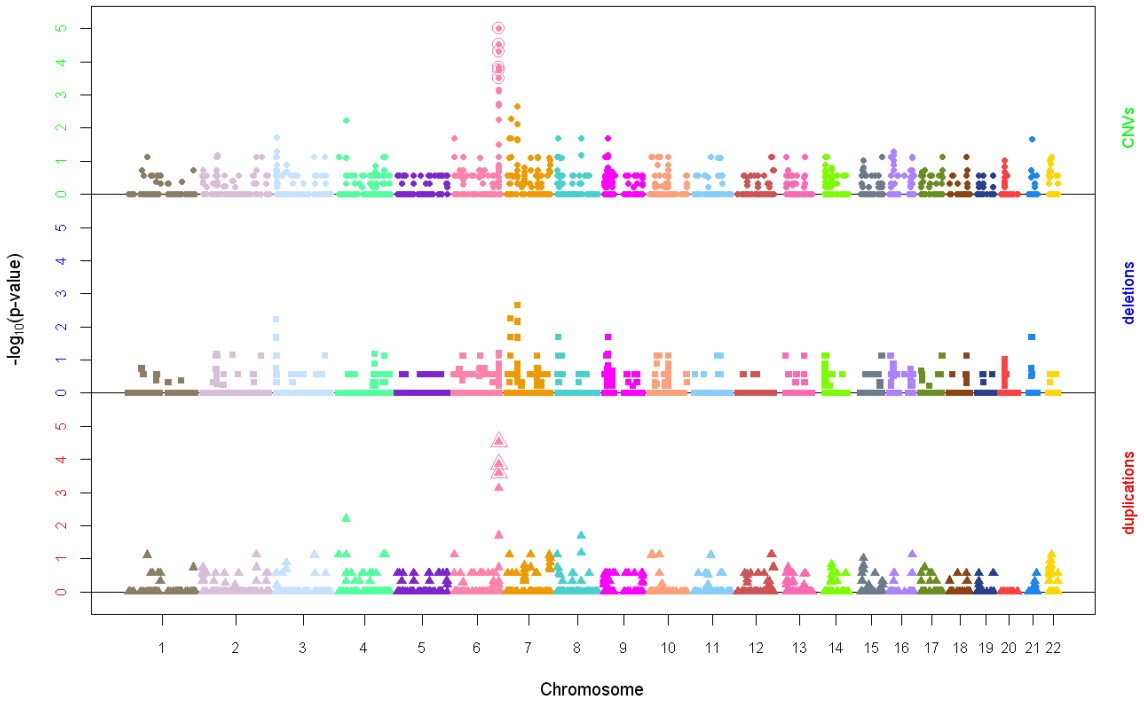

Figure 3.7: Manhattan plot for the genome-wide analysis of rare CNVs in 489 ADHD patients and 1 285 control subjects. Nominal one-sided p-values of Fisher's exact tests are presented. Genome-wide statistically significantly associated sites are circled.

Table 3.7: CNVs at the *PARK2* locus in the ADHD GWAS discovery sample.

Proband's ID	**Proband's Status**	**called CNV**	**CNV call's state**	**qPCR CNV state**
control_1	control subject	chr 6: 162 379 561 – 162 903 833	cn = 1	cn = 1
case_1	ADHD patient	chr 6: 162 477 709 – 162 724 935	cn = 1	cn = 1
control_2	control subject	chr 6: 162 554 327 – 162 659 755	cn = 1	DNA not available
control_3	control subject	chr 6: 162 594 083 – 163 001 802	cn = 3	DNA not available
case_2	ADHD patient	chr 6: 162 629 938 – 162 935 269	cn = 3	cn = 3
case_3	ADHD patient	chr 6: 162 637 688 – 162 809 965	cn = 3	cn = 3
case_4	ADHD patient	chr 6: 162 644 237 – 162 834 976	cn = 3	cn = 2
case_5	ADHD patient	chr 6: 162 644 237 – 162 834 976	cn = 3	cn = 3
case_6	ADHD patient	chr 6: 162 644 237 – 162 829 925	cn = 3	cn = 3
case_7	ADHD patient	chr 6: 162 644 237 – 162 829 925	cn = 3	cn = 3
case_8	ADHD patient	chr 6: 162 644 237 – 162 834 976	cn = 3	cn = 3
case_9	ADHD patient	chr 6: 162 644 237 – 162 829 925	cn = 3	cn = 3
case_10	ADHD patient	chr 6: 162 674 596 – 162 834 976	cn = 3	cn = 3
case_11	ADHD patient	chr 6: 162 687 672 – 162 896 029	cn = 1	cn = 1
case_12	ADHD patient	chr 6: 162 687 672 – 162 789 187	cn = 1	cn = 1
control_4	control subject	chr 6: 162 719 107 – 162 965 453	cn = 3	DNA not available
control_5	control subject	chr 6: 162 740 072 – 162 805 539	cn = 1	DNA not available
control_6	control subject	chr 6: 162 767 020 – 162 903 833	cn = 1	DNA not available

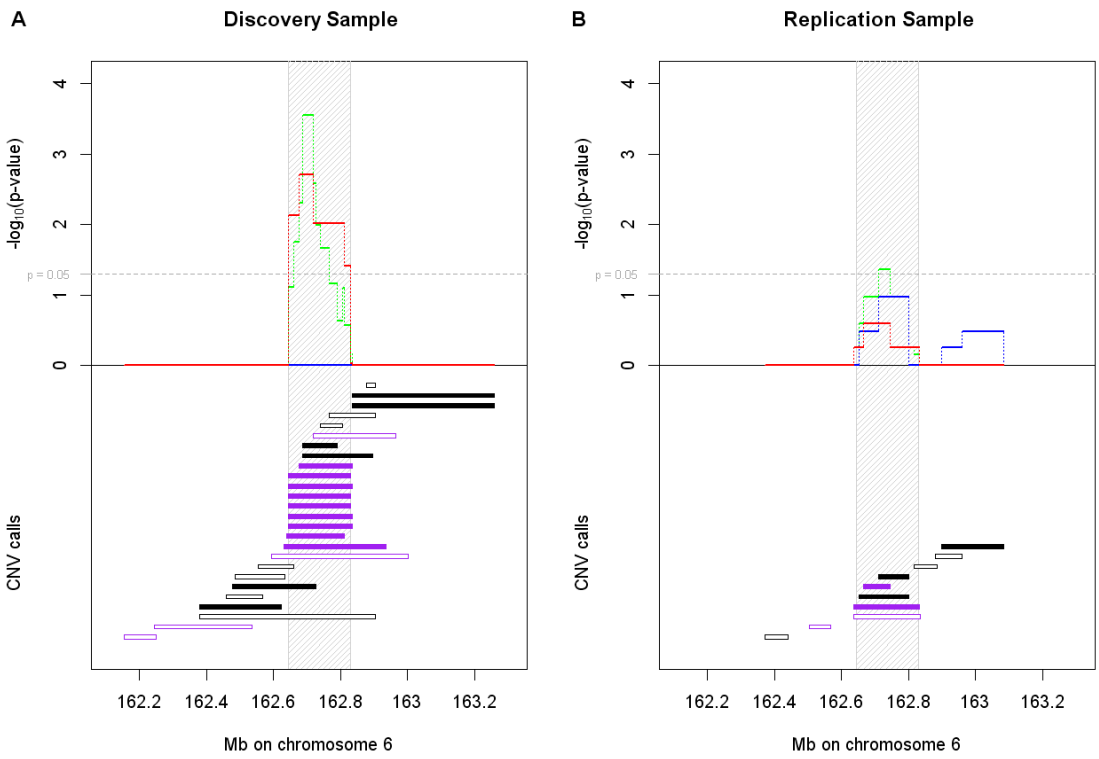

Figure 3.8: Association test results at the *PARK2* locus in the ADHD GWAS discovery and replication sample. Individual CNV calls are presented in the lower parts of panel A and B. black: heterozygous deletions, purple: hemizygous duplications, filled: in ADHD patients, unfilled: in control subjects. One-sided p-values of Fisher's exact tests are depicted in the upper parts of panel A and B. green: CNVs, blue: deletions, red: duplications. A) Genome-wide corrected empirical p-values are shown. B) Nominal p-values are shown. The shaded region highlights the genomic region with genome-wide statistically significant association test results (i.e. $p < 0.05$) in the discovery sample.

Table 3.8: CNVs at the *PARK2* locus in the ADHD replication sample.

Proband's ID	**Proband's Status**	**called CNV**	**CNV call's state**	**qPCR CNV state**
control_a	control subject	chr 6: 162 636 295 – 162 834 267	cn = 3	DNA not available
case_a	ADHD patient	chr 6: 162 636 295 – 162 832 504	cn = 3	DNA not available
case_b	ADHD patient	chr 6: 162 651 333 – 162 800 484	cn = 1	DNA not available
case_c	ADHD patient	chr 6: 162 665 597 – 162 745 876	cn = 3	DNA not available
case_d	ADHD patient	chr 6: 162 710 468 – 162 800 484	cn = 1	DNA not available
control_b	control subject	chr 6: 162 816 996 – 162 884 672	cn = 1	DNA not available

Table 3.9: Association test results at the *PARK2* locus in the ADHD GWAS discovery and replication sample. (red: $p < 0.05$)

	Discovery Sample								**Replication Sample**				
sub-CNVR on chr 6	cases DELs // DUPs[a]	controls DELs // DUPs[a]	EMP1 CNVs[b]	EMP2 CNVs[c]	EMP1 DELs[b]	EMP2 DELs[c]	EMP1 DUPs[b]	EMP2 DUPs[c]	cases DELs // DUPs[a]	controls DELs // DUPs[a]	EMP1 CNVs[b]	EMP1 DELs[b]	EMP1 DUPs[b]
162 636 295 – 162 644 236	1 // 2	2 // 1	0.2107	1.0000	0.6200	1.0000	0.1862	1.0000	0 // 1	0 // 1	0.5527	1.0000	0.5545
162 644 237 – 162 651 332	1 // 8	2 // 1	7.40E-04	0.0781	0.6200	1.0000	1.40E-04	0.0075	0 // 1	0 // 1	0.5527	1.0000	0.5545
162 651 333 – 162 659 755	1 // 8	2 // 1	7.40E-04	0.0781	0.6200	1.0000	1.40E-04	0.0075	1 // 1	0 // 1	0.2545	0.3300	0.5545
162 659 756 – 162 665 596	1 // 8	1 // 1	1.80E-04	0.0179	0.4737	1.0000	1.40E-04	0.0075	1 // 1	0 // 1	0.2545	0.3300	0.5545
162 665 597 – 162 674 595	1 // 8	1 // 1	1.80E-04	0.0179	0.4737	1.0000	1.40E-04	0.0075	1 // 2	0 // 1	0.1078	0.3300	0.2577
162 674 596 – 162 687 671	1 // 9	1 // 1	5.00E-05	0.0050	0.4737	1.0000	3.00E-05	0.0019	1 // 2	0 // 1	0.1078	0.3300	0.2577
162 687 672 – 162 710 467	3 // 9	1 // 1	1.00E-05	2.80E-04	0.0647	1.0000	3.00E-05	0.0019	1 // 2	0 // 1	0.1078	0.3300	0.2577
162 710 468 – 162 719 106	3 // 9	1 // 1	1.00E-05	2.80E-04	0.0647	1.0000	3.00E-05	0.0019	2 // 2	0 // 1	0.0431	0.1084	0.2577
162 719 107 – 162 724 935	3 // 9	1 // 2	3.00E-05	0.0026	0.0647	1.0000	2.60E-04	0.0098	2 // 2	0 // 1	0.0431	0.1084	0.2577
162 724 936 – 162 740 071	2 // 9	1 // 2	1.50E-04	0.0103	0.1841	1.0000	2.60E-04	0.0098	2 // 2	0 // 1	0.0431	0.1084	0.2577
162 740 072 – 162 745 876	2 // 9	2 // 2	3.00E-04	0.0219	0.3033	1.0000	2.60E-04	0.0098	2 // 2	0 // 1	0.0431	0.1084	0.2577
162 745 877 – 162 767 019	2 // 9	2 // 2	3.00E-04	0.0219	0.3033	1.0000	2.60E-04	0.0098	2 // 1	0 // 1	0.1088	0.1084	0.5545
162 767 020 – 162 789 187	2 // 9	3 // 2	6.80E-04	0.0697	0.4187	1.0000	2.60E-04	0.0098	2 // 1	0 // 1	0.1088	0.1084	0.5545
162 789 188 – 162 800 484	1 // 9	3 // 2	1.79E-03	0.2351	1.0000	1.0000	2.60E-04	0.0098	2 // 1	0 // 1	0.1088	0.1084	0.5545
162 800 485 – 162 805 539	1 // 9	3 // 2	1.79E-03	0.2351	1.0000	1.0000	2.60E-04	0.0098	0 // 1	0 // 1	0.5527	1.0000	0.5545
162 805 540 – 162 809 965	1 // 9	2 // 2	8.00E-04	0.0804	0.6200	1.0000	2.60E-04	0.0098	0 // 1	0 // 1	0.5527	1.0000	0.5545
162 809 966 – 162 816 995	1 // 8	2 // 2	0.0021	0.2682	0.6200	1.0000	7.40E-04	0.0388	0 // 1	0 // 1	0.5527	1.0000	0.5545
162 816 996 – 162 829 925	1 // 8	2 // 2	2.13E-03	0.2682	0.6200	1.0000	7.40E-04	0.0388	0 // 1	1 // 1	0.7003	1.0000	0.5545
162 829 926 – 162 832 503	1 // 5	2 // 2	0.0311	0.9983	0.6200	1.0000	0.0199	0.9375	0 // 1	1 // 1	0.7003	1.0000	0.5545

[a] DELs: deletions, DUPs: duplications.
[b] EMP1: nominal one-sided p-value of Fisher's exact test, based on comparing the CNV frequency in cases versus controls.
[c] EMP1: nominal one-sided p-value of Fisher's exact test, based on comparing the frequency of CNVs, deletions or duplications in cases versus controls.
[d] EMP2: genome-wide corrected empirical (based on 100 000 permutations) one-sided p-value of Fisher's exact test, based on comparing the frequency of CNVs, deletions or duplications in cases versus controls.

Association tests were performed at 3 964 genomic CNV sites, which represented 1 083 non-overlapping CNVRs (Figure 3.7). Only one CNVR, at chr 6: 162 659 756 − 162 829 925 within the *PARK2* gene (called *PARK2* locus in the following), showed genome-wide statistically significant association test results (Figure 3.8, Table 3.9).

On the one hand, there was a statistically significant over-representation of CNVs, including deletions and duplications, at the *PARK2* locus in ADHD patients compared to controls (minimal genome-wide corrected empirical one-sided p-value = 2.8×10^{-4}).

On the other hand, association tests stratified by CNV type revealed that this association effect was mainly driven by an over-representation of duplications in ADHD cases versus controls (minimal genome-wide corrected empirical one-sided p-value, deletions: p = 1, duplications: p = 0.0019). In total, this locus was covered by twelve CNVs of ADHD patients (three deletions and nine duplications) and six CNVs of control subjects (four deletions and two duplications) (Table 3.7).

At the *PARK2* locus, individual copy number states of the GWAS discovery sample were re-evaluated by use of qPCR experiments. With the exception of one ADHD case's duplication, each array-based CNV call could be technically validated. Subsequent genome-wide re-analysis did not meaningfully change association test results (for details see Jarick et al. (2012)).

For follow-up analyses, an independent sample of 386 ADHD patients and 781 healthy control subjects was examined. The finding of an association of CNVs at the *PARK2* locus with the trait ADHD was confirmed in the replication analyses with statistical significance (minimal nominal one-sided p-value = 0.043) (Figure 3.8, Table 3.9). At the *PARK2* locus, a total of four CNVs (two deletions and two duplications) were observed in ADHD patients compared with two CNVs (one deletion and one duplication) in control subjects of the replication sample (Table 3.8).

3.7.4 Discussion

A statistically genome-wide significant association of CNVs, including both deletions and duplications, at the *PARK2* locus could be detected in application of strategy S2 for the genome-wide analysis of rare CNVs to an ADHD case-control sample. This finding was technically validated by use of qPCR and it was additionally confirmed by statistically significant association test results in replication analyses of an independent ADHD case-control sample. QPCR validation experiments demonstrated that array-derived rare CNV calls were estimated with low false negative and low false positive rates at the *PARK2* locus.

Due to the focus on rare CNVs, strategy S1, which does not provide any possibility to filter CNV signals based on their population frequency, is not applicable to the data example for the trait ADHD. The main emphasis of strategy S2 is to ensure validity of CNV calls. Concerning the ADHD data set, additional pre-processing steps were necessary prior to CNV calling due to the fact that differing genotyping chips were used for ADHD patients and control subjects. In this sense, the whole analysis was limited to the intersecting set of probe sets between both SNP arrays, and conservative quality control criteria were applied.

With additional regard to the previous application to the phenotype obesity, taken together, strategy S2 has proven to be a flexible and versatilely usable strategy for the genome-wide analysis of both, rare and common CNVs. Compared to strategy S1, the additional CNV calling step might on the one hand be challenging, depending on how comparable the initial raw data was ascertained across the study subjects, and with regard to a lack of CNV calling tools with perfect validity. However, based on results of the CNV detection step CNV association testing may on the other hand be performed more focused on the respective type of CNVs being most probably causal for the trait of interest.

At least in case of rare CNVs being called for ADHD patients and control subjects at the *PARK2* locus, it was shown that rare CNVs can be called with high validity based on SNP genotyping data. This is in concordance with previous reports that evaluated the performance of several software suites in the identification of CNVs with a special focus on potential differences between different CNV types according to their population frequency. Zhang et al. (2011) found that the recovery rates of CNVs consistently increases with decreasing underlying CNV frequency throughout all investigates software tools, although to differing degrees. Moreover, Zhang et al. (2011) observed that rare CNV calls were less likely to be affected by technical biases, such as plate effects, in comparison to common CNV calls. The latter study even concludes that common CNV calls being derived from applying any CNV calling software to array data are not suitable for association studies without independent experimental genotyping. In contrast, rare CNVs are of substantial better quality.

4 Two Proposed Strategies for Genome-Wide CNV Association Analyses

Genome-wide association analyses of CNVs, based on raw CNV intensity data derived from SNP array experiments, involve complex analyses strategies. As touched upon in the previous chapter, modification of any single analysis step may greatly influence association test results. In this chapter, two new strategies for genome-wide association analyses of raw CNV data are briefly presented. Both strategies, PS1 and PS2, are characterized in more detail in the following two chapters.

The key aspects of the two new strategies, PS1 and PS2, are graphically illustrated in Figure 4.1 and outlined in Table 4.1. In each of the two proposed strategies, one single analysis step is modified relative to the standard CNV analyses strategies S1 and S2 (Table 3.5), while the remaining steps are performed in an unmodified way.

Table 4.1: Two proposed strategies for genome-wide analyses of raw CNV data.

	Strategy PS1	**Strategy PS2**
Statistical testing is based on ...	individual raw copy number values of CNV probe sets being located in pre-determined CNV regions.	individual CNVs that are called by use of sophisticated reference models.
Statistical tests are performed at ...	CNV probe sets in pre-determined CNV regions.	any genomic region offering at least one CNV.
The number of tests performed equals the ...	number of CNV probe sets in pre-determined CNV regions.	number CNV regions.
The proposed strategy is different to ..	strategy S1, in the way that markers for statistical testing are selected.	strategy S2, in the way that CNVs are called.

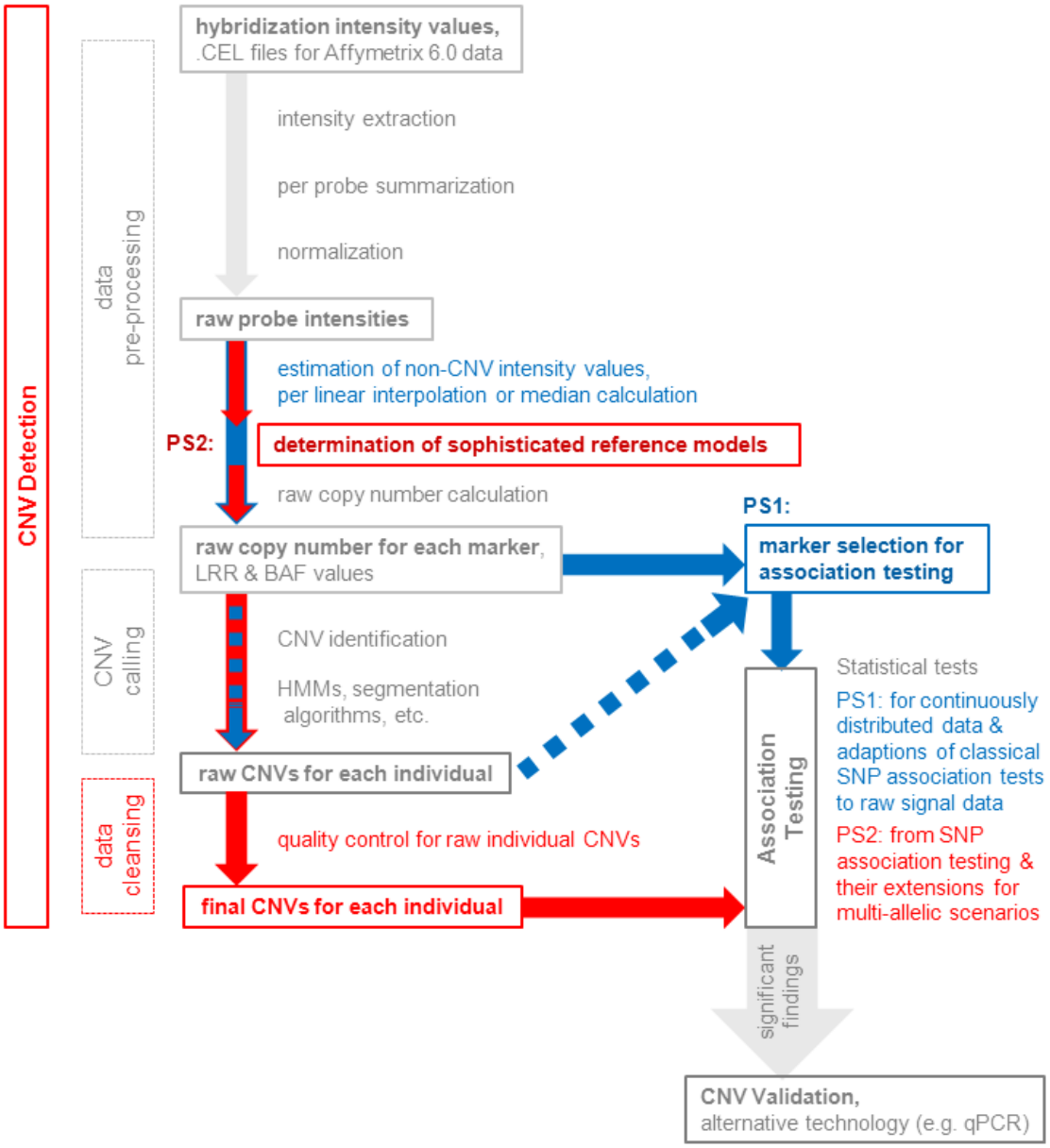

Figure 4.1: Schematic representation of the two proposed strategies, PS1 (blue) and PS2 (red), for genome-wide CNV analyses.

Compared to strategy S1, strategy PS1 incorporates a refinement on the selection of genetic markers to be tested for an association. In both strategies, S1 and PS1, association testing is performed directly on the basis of pre-processed intensity signals without classifying them into discrete copy number states. As illustrated in section 3.5, this procedure results in testing thousands of probe sets without biological diversity of copy number. This will not only artificially inflate the amount of multiple testing, but at the same time unnecessarily increase the probability for false positive

findings without biological plausibility. Instead, strategy PS1 restricts association testing on those probe sets that offer a minimal copy number variability with regard to estimates in the analysed sample or to public database entries. For the determination of relevant probe sets in copy number variable regions, a simplified CNV calling might be performed aiming to assess the potential range of probe-specific copy number variability. Otherwise, public databases such as the Database of Genomic Variants (DGV, http://projects.tcag.ca/variation/) or the Human Genome Structural Variation Project (http://humanparalogy.gs.washington.edu/structuralvariation/) may be helpful in the probe set selection step. The decision of CNVs to be or not to be called in the data set of interest should depend on the underlying hypothesis of interest, i.e. the 'rare disease - rare variant', the 'common disease - common variant' or the 'common disease - many rare variants' hypothesis (Mayo, 2007). Common CNVs with a population frequency above five percent will likely be recognized in public databases, which reflect CNV mapping results of several previous studies. However, CNVs with lower frequency that may even be specific to the phenotypic trait of interest might be missed without a sample-specific CNV identification.

Alternatively, strategy PS2 focuses on a modified procedure of the CNV genotype calling step. In more detail, sample-specific copy number neutral reference intensity values that are the basis of CNV calling are estimated in fitting a Gaussian mixture model to the observed hybridization intensity data at each available informative probe set. This approach is motivated by the observation that the widespread use of median probe-wise reference intensity values offers the risk of potential CNV misclassifications. It will be demonstrated that the burdensome procedure of determining sophisticated reference intensity values provides benefits in terms of stability, reproducibility, false positive and Mendelian inconsistency rates of CNV calls. An improved accurateness of CNV detection insures a higher reliability of subsequently performed association tests.

5 Strategy PS1: Association Testing Based on Copy Number Variation Signals

As outlined in the previous chapters, changing single parts of a CNV analysis strategy may remarkable influence association test results. The design of a whole genome-wide CNV association analysis may have an influence on both, the risk of false positive and false negative findings.

In this chapter, we will focus on how to select array probe sets for inclusion in association tests when genome-wide testing is based on pre-processed continuously distributed intensity values, which are an indirect measure of individual copy number states. First, the proposed analyses strategy PS1 will be introduced by giving a general overview of the new methodical aspects in comparison to strategy S1. Afterwards, key aspects concerning statistical association testing will be presented in detail for the case-control and the family-based analyses designs. Finally, the practical application of strategy PS1 on a real data example for the trait obesity is given. Exactly the same obesity data set that has previously been investigated by use of the two standard strategies, S1 and S2, will be re-analysed by applying strategy PS1 and potential benefits and disadvantages will be discussed based on this example.

5.1 Strategy PS1

When applying strategy S2 or PS2, one of the first steps of a genome-wide CNV analysis is the estimation of individual CNVs. Subsequently, genome-wide association testing is based on the obtained CNV calls. Thus, the number of tests to perform in a genome-wide analysis is given by the number of genomic regions that contain CNV calls. Contrarily, the CNV calling step is skipped when strategy S1 or PS1 is implemented. Instead, genome-wide association testing is based on pre-processed hybridization intensity data that have not been categorized into discrete

copy number classes, but are positively correlated with individual CNV states. Since in application of strategy S1 there is no CNV classification available at all for any of the CNV markers, each available CNV probe set is tested for an association with the phenotype of interest irrespective of whether it is biologically plausible, that is whether or not the probe set reflects an underlying CNV.

One disadvantage of strategy S1 is that significant association test results due to potential noisy data in genomic regions, which are in fact free of copy number variability, cannot be distinguished from significant findings in truly associated CNV

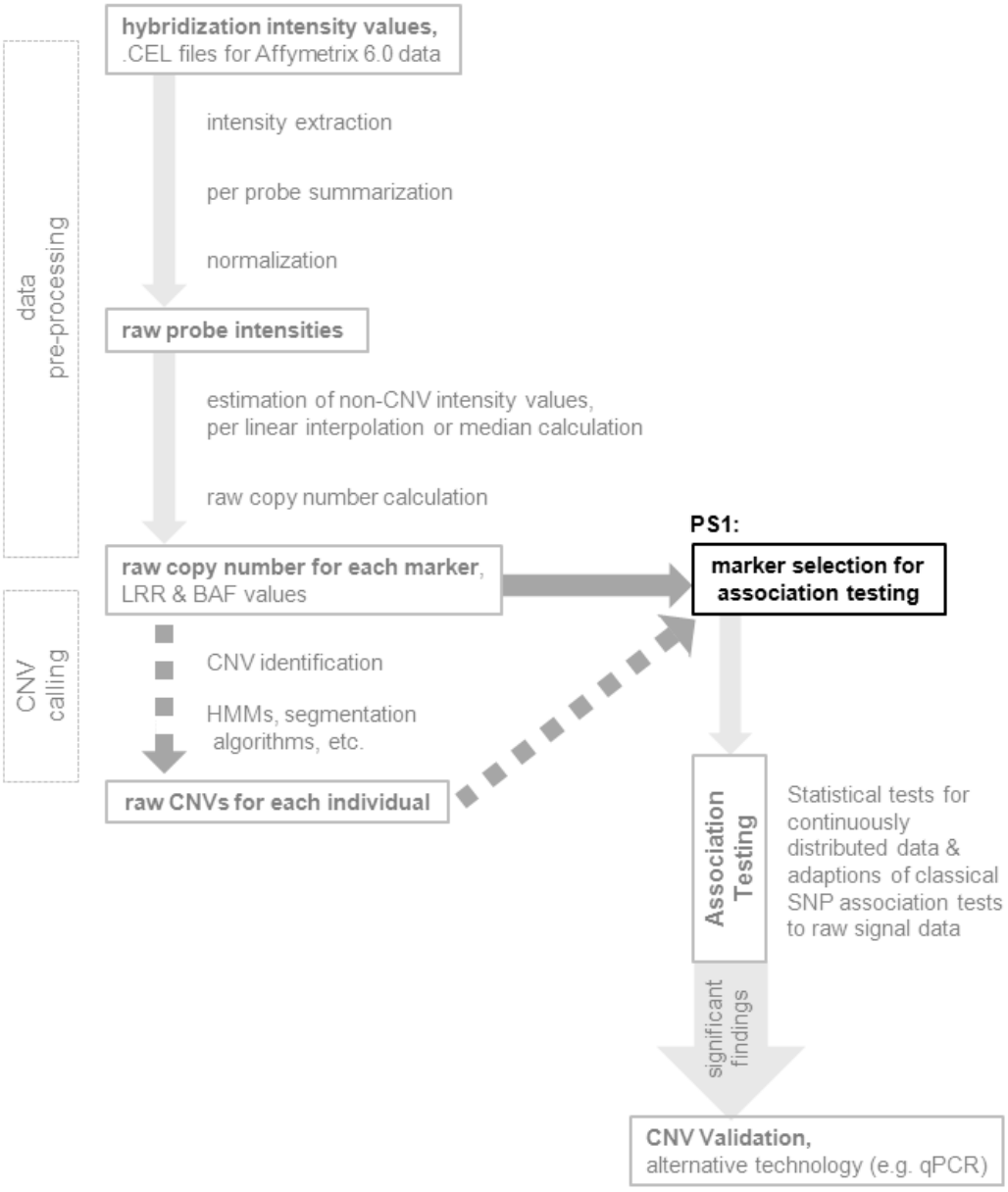

Figure 5.1: Schematic representation of the proposed CNV analyses strategy PS1.

regions. False positive association test results in genomic regions without CNV appearance can only be recognized as such in the final validation step. Additionally, several association tests are performed for markers in regions where the null hypothesis of no association can in fact not be rejected due to missing copy number variability of the region. This enrichment of null markers artificially exacerbates the multiple testing problem. Furthermore, performing the greatest possible number of tests involves many perfectly correlated tests for probe sets in the same CNV.

As shown in Figure 5.1, the proposed CNV analysis strategy PS1 includes one additional step in comparison to strategy S1. Prior to the association testing step, a subset of markers is selected for association testing. In this inserted step, markers may on the one hand be selected based on information of public CNV catalogues, such as the DGV (Database of Genomic Variants, http://projects.tcag.ca/variation/). As illustrated in Figure 5.1, sample-specific CNV information that result from a minimalistic CNV calling step may alternatively be considered. In short, the marker selection step aims to identify most informative probe sets for association analyses.

Let information for a total of p probe sets be available for CNV analyses from array experiments in a sample that comprises n individuals. Let r denote the number of CNV containing genomic regions, where each region i $(i = 1, \ldots, r)$ involves v_i probe sets, and in total $w = \sum_{i=1}^{r} v_i$ probe sets reflect true underlying copy number variability. Thus, the subset of probe sets

$$\{s_1, \ldots, s_w\} \subset \{1, \ldots, p\} \quad \text{with} \quad w = \sum_{i=1}^{r} v_i < p,$$

which are informative for association testing, is given by the condition that

$$\frac{|\{z \mid (z \leq n) \ \wedge \ (\text{ individual } z \text{ has a CNV at probe set } s_j)\}|}{n} > \epsilon, \tag{5.1}$$

for all s_j $(j = 1, \ldots, w)$, where $\epsilon > 0$ is a pre-specified threshold for the assumed minimal copy number variability of each CNV region. To check which of the p array probe sets satisfy condition (5.1), either a sample-wide CNV calling may be applied or alternatively ϵ may be estimated from publicly available databases.

Finally, the global null hypothesis

H_0 : no association between any $\{s_j \mid s_j$ satisfying condition (5.1) $\}$ with disease

is tested against the alternative hypothesis

H_A : at least one $\{s_j \mid s_j$ satisfying condition (5.1) $\}$ is associated with disease,

where the form of the single test statistic depends on the specific sample design. Hence, for rejecting any single null hypothesis at the probe level multiple testing procedures, which control the family-wise error rate (FWER) with respect to the set of null hypotheses determined above, have to be considered.

5.1.1 Case-Control Designs

For CNV analyses, we assume to have pre-processed hybridization signal intensity values available for a total of p probe sets in a sample comprising n individuals. In a case-control design, intensity signals of n_a affected cases and n_u healthy controls are compared at each probe set in order to detect genomic regions that are associated with disease. Let

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) \quad \text{with} \quad i = 1, \ldots, n$$

denote the $n = n_a + n_u$ vectors of observed individual pre-processed probe hybridization intensity signals.

5.1.1.1 The Logistic Regression Model for Raw CNV Data

For a case-control CNV association study, the objective is to compare the presence of certain CNVs between cases and controls in order to find association between the respective CNV and the disease or trait of interest. When association testing is based on pre-processed hybridization signal intensity values, counting and comparing the different types of CNVs between both groups is not feasible. Instead, the relationship between the indirect CNV measurement of hybridization intensities and individual disease affection status is most frequently explored in the context of a logistic regression model.

Suppose that two mutually exclusive disease groups are defined and let the development of the disease during a defined accession period be described by a Bernoulli random variable with values of 0 (= unaffected) and 1 (= affected). Suppose that the probability of developing the disease or trait of interest is specific to each individual and largely dependent on the individual's genetic information. This is in concordance with a case-control sample that was especially designed to describe the relationship between the binary affection status and explanatory genetic variants. In this special genetic case-control design all recruited individuals are either healthy

or sick, and in particular they are all genetically unrelated, that is any relatives are excluded in order to explore the influence of genetic variants on the disease status.

Concerning raw CNV data, before grouping CNV probe sets in the CNV calling step into subsets, which make up CNVs or CNV regions, each available array probe set is initially assumed to independently contribute to the individuals' phenotypes. Consequently, a separate logistic regression model is developed for each available CN probe set. Hence, for each probe set $k = 1, \ldots, p$, the disease affection status of the n independent individuals of the above described case-control sample can formally be described as a vector of n independent Bernoulli random variables

$$\mathbf{Y} = (Y_1, \ldots, Y_n) \quad \text{with} \quad Y_i \sim \mathcal{B}_{(1,\pi_{ik})} \quad \text{for} \quad i = 1, \ldots, n.$$

That is, each random variable Y_i can either have values of 1 or 0, which stands for the presence or absence of the disease, with probabilities of π_{ik} and $1 - \pi_{ik}$, respectively.

In a prospective study in which initially healthy individuals are followed throughout the accession period to observe disease incidence, the probability for the i-th individual of having developed the trait of interest with respect to the underlying intensity information for the k-th probe set would equal $\mathbb{P}(Y_i = 1|x_{ik}) = \pi_{ik}$. For each available probe set $k = 1, \ldots, p$, the expected value of the individual affection status, $\mathbb{E}(Y_i|x_{ik}) = \pi_{ik}$, would then be assumed to depend on the observed intensity signal x_{ik} in the following form

$$\mathbb{E}(Y_i|x_{ik}) = F(\beta_{0k} + \beta_{1k}x_{ik}) \quad \text{with} \quad \beta_{0k}, \beta_{1k} \in \mathbb{R}, \tag{5.2}$$

where $F(.)$ denotes the logistic distribution function of the presumed Logit model

$$F(z) = \frac{e^z}{1 + e^z}.$$

In other words, a generalized linear model (GLM) given by equation (5.2) with logit link function $F^{-1}(z) = \text{logit}(z) = \ln(\frac{z}{1-z})$ is assumed. Then, the odds of being affected with respect to observed intensities at probe set $k = 1, \ldots, p$ can equivalently be expressed as

$$\frac{\mathbb{P}(Y_i = 1|x_{ik})}{1 - \mathbb{P}(Y_i = 1|x_{ik})} = e^{\beta_{0k} + \beta_{1k}x_{ik}}.$$

Consequently,

$$\frac{\mathbb{P}(Y_i = 1|x_{ik})}{\mathbb{P}(Y_i = 0|x_{ik})} \Big/ \frac{\mathbb{P}(Y_i = 1|x_{ik_0})}{\mathbb{P}(Y_i = 0|x_{ik_0})} = e^{\beta_{1k}(x_{ik} - x_{ik_0})}. \tag{5.3}$$

equals the odds ratio for being affected ($Y_i = 1$) for an individual with genetic characteristic x_{ik}, relative to that for an individual with some standard genetic regression variable x_{ik_0}.

Contrarily, a case-control study involves direct sampling from $\mathbb{P}(x_{ik}|Y_i)$. Since $\mathbb{P}(x_{ik}|Y_i)$ does not completely determine $\mathbb{P}(Y_i|x_{ik})$, the full prospective model cannot be estimated from case-control data alone. However, under the assumption that the selection of cases and controls is independent of covariate values and with regard to the Bayes' theorem

$$\mathbb{P}(Y_i|x_{ik}) = \frac{\mathbb{P}(x_{ik}|Y_i)\,\mathbb{P}(Y_i)}{\mathbb{P}(x_{ik})},$$

the odds ratios (5.3) can equivalently be written as

$$e^{\beta_{1k}(x_{ik}-x_{ik_0})} = \frac{\mathbb{P}(x_{ik}|Y_i=1)}{\mathbb{P}(x_{ik_0}|Y_i=1)} \Big/ \frac{\mathbb{P}(x_{ik}|Y_i=0)}{\mathbb{P}(x_{ik_0}|Y_i=0)}.$$

It follows that the odds ratio (5.3) can be estimated from case-control data. Prentice and Pyke (1979) have shown that through assuming a prospective logistic model (5.2), maximum likelihood estimates from all regression coefficients except for the constant term can be obtained by ignoring the case-control scheme, i.e. the case-control problem can be treated as a prospective one. Thus, although logistic modeling is likewise applicable to retrospective case-control studies, there is one important limitation. For case-control studies, the fitted logistic model cannot be used to predict risk for an individual with specified independent variables.

For a case-control CNV association study based on raw hybridization signal intensity values, the logistic regression coefficients β_{1k} are separately estimated for each probe set $k = 1, \ldots, p$. Subsequently, the hypotheses

$$H_0 : \beta_{1k} = 0 \quad \text{versus} \quad H_A : \beta_{1k} \neq 0 \quad \text{for} \quad k = 1, \ldots, p$$

are tested. Prentice and Pyke (1979) have shown that maximum likelihood estimates (MLEs) obtained by pretending that the case-control data resulted from a prospective study have the usual properties associated with MLEs. Specifically, they are asymptotically normally distributed for large sample sizes, and thus the Wald test is frequently applied in the testing step. Alternatively, two asymptotically equivalent tests can be applied to the total model: the likelihood ratio test or the score test, which is presented for the family-based design in chapter 5.1.2. However, the latter likelihood ratio techniques are not technically correct in the underlying situation since the likelihood function is based on an incorrect model, i.e. a prospective model for an retrospective sampling scheme.

5.1.1.2 Multi-Marker Logistic Regression CNV Models

For the univariate model presented above, the complex logistic regression framework would not necessarily be needed. The comparison of one continuously distributed explanatory variable between the two differently exposed groups could be performed in a more simple way by use of the unpaired t-test given that normality holds or otherwise by use of the non-parametric Mann-Whitney test. However, the logistic regression model offers the advantage that modifications and extensions can flexibly be implemented. For example, the impact of several additional explanatory covariates like age or sex, denoted by $\mathbf{a} = (a_i, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, can easily be incorporated in the following way

$$\ln\left(\frac{\mathbb{P}(Y_i = 1|(x_{ik}, a_i, b_i))}{1 - \mathbb{P}(Y_i = 1|(x_{ik}, a_i, b_i))}\right) = \beta_{0k} + \beta_{1k}x_{ik} + \beta_{2k}a_i + \beta_{3k}b_i \quad \text{for} \quad k = 1, \ldots, p.$$

Alternatively, multivariate logistic regression can be used to fit multi-marker models in which several adjacent probe sets, $\mathbf{c}_j = \{c_{j1}, \ldots, c_{jl_j}\} \subset \{1, \ldots, p\}$ with $j < p$, are assumed to collectively have an influence on the probability for an individual to be affected with the disease or trait of interest

$$\ln\left(\frac{\mathbb{P}(Y_i = 1|(x_{ic_{j1}}, \ldots, x_{ic_{jl_j}}))}{1 - \mathbb{P}(Y_i = 1|(x_{ic_{j1}}, \ldots, x_{ic_{jl_j}}))}\right) = \beta_{j0} + \beta_{j1}x_{ic_{j1}} + \cdots + \beta_{jl_j}x_{ic_{jl_j}} \quad \text{with} \quad j < p.$$

Finally, each mentioned approach for association testing in case-control data can be traced back to a comparison of the two distributions derived from intensity observations in cases and controls. As such, multiple confounding factors such as batch effects and noisy data as well as the known confounders of genotyping array data such as the effective amount of DNA hybridized, background fluorescence and hybridization quality can bias association test results (Chai et al., 2010). Furthermore, different genotyping procedures or platforms can lead to group differences that are impossible to be distinguished from association effects without pre-processing of the data. The risk of confounding bias, which may result from the described variation between arrays of non-biological origin, can for instance be reduced by a careful data normalization procedure prior to any association testing (Bolstad et al., 2003).

5.1.1.3 Marker Selection for Case-Control Association Testing

In a case-control sample, only those probe sets

$$\{s_1, \ldots, s_w\} \subset \{1, \ldots, p\} \quad \text{with} \quad w < p$$

with minimal copy number variability of $\epsilon > 0$ referred to the controls group,

$$\frac{|\{u \mid (u \leq n_u) \wedge (\text{ control } u \text{ has a CNV at probe set } s_j)\}|}{n_u} > \epsilon \quad \text{for} \quad j = 1, \ldots, w,$$

are tested for an association with disease. The estimated frequency of CNVs in the controls is not influenced by the trait of interest and should therefore represent an unbiased estimate of the underlying CNV frequency.

5.1.2 Family-Based Trio Designs

For simplicity and with regard to the following application to data, we will limit all considerations for the family-based design to parent-offspring trio samples. That is, the focus will be on samples, which exclusively consist of n_t nuclear families, each comprising exactly one affected child and both biological parents. However, all concepts can directly be adapted to larger families comprising more than one affected or healthy child.

5.1.2.1 FBAT for Trio Designs and Raw CNV Data

We assume to have pre-processed hybridization signal intensity data on p probe sets for n_t independent trios, each comprising one affected child and both biological parents. Let

$$\mathbf{X}_{ji} = (X_{ji1}, \ldots, X_{jip}) \quad \text{with} \quad j \in \{o, m, f\} \quad \text{and} \quad i = 1, \ldots, n_t$$

denote the vector of random variables for probe set intensities of either the offspring ($j = o$), the mother ($j = m$) or the father ($j = f$) of the i-th trio, respectively. Let furthermore $\mathbf{x}_{ji}$ denote the vector of observations for the corresponding random variable X_{ji}.

As proposed by Ionita-Laza et al. (2008), to test each marker probe set $k = 1, \ldots, p$ for a CNV - phenotype association the following generalized family-based association test (FBAT) scores can be used

$$S_k = \sum_{i=1}^{n_t} x_{oik} - \mathbb{E}_0(X_{oik} | x_{mik}, x_{fik}) \quad \text{for} \quad k = 1, \ldots, p,$$

where $\mathbb{E}_0(X_{oik} | x_{mik}, x_{fik})$ denotes the expected value of the offspring's hybridization intensity given the two observed parental intensities and computed under the null hypothesis of no association. We assume that both parental intensity data are

observed and available for association testing analyses for each of the p available probe sets.

In order to specify the conditional null distribution of the offspring's marker intensity data, the exact null hypothesis needs to be specified. In general, genetic family-based association tests have composite null and alternative hypotheses that account for the marker's state of (1): linkage with a hypothetical disease-susceptibility locus (DSL) and (2): association with the disease causing mutant allele of the DSL or direct association with the respective phenotypic trait in case that the tested marker is the true causal variant. Thus, tested null hypotheses may either be (Laird and Lange, 2006): 'no association and no linkage', 'no association in the presence of linkage' or 'no linkage in the presence of association'. In follow-up studies, in which statistically significant results from previous case-control association studies are re-analysed in further samples, the null hypothesis of 'no linkage in the presence of association' is tested. Furthermore, the null hypothesis of 'no association in the presence of linkage' is used for follow-up studies that account for statistically significant results of previous linkage studies when analysing additional samples. Contrarily, for genome-wide association studies the null hypothesis of

$$H_0\text{: 'no association and no linkage'}$$

is tested against the only testable alternative hypothesis (Ott, 1989) of

$$H_A\text{: 'linkage and association' between the marker and a DSL}$$

at each available marker probe set. Since both linkage and in particular association between the marker and a DSL affecting the trait have to be present in order to reject the null hypothesis in family-based association testing (Ott, 1989), the FBAT can especially be seen as a test for association. The precise null hypothesis of the FBAT is of particular interest when data of more than one offspring per family is incorporated in the test statistic. In more detail, transmissions from the same parents to multiple offspring are correlated in the presence of linkage (Laird and Lange, 2006).

According to Mendel's laws, each parent is equally likely to transmit his or her genotypes and parental transmissions are independent under the null hypothesis of neither association nor linkage. Consequently, conditional on the sufficient statistic of the parental within-family information, the offspring's expected intensity can under the null hypothesis be calculated as (Rabinowitz and Laird, 2000)

$$\mathbb{E}_0(X_{oik}|x_{mik}, x_{fik}) = \frac{1}{2}(x_{mik} + x_{fik}) \quad \text{for} \quad k = 1, \ldots, p.$$

Thus, the offspring's intensities are treated as being random, whereas parental intensities are assumed to be fixed. Then, by conditioning on the sufficient statistic, $T_{ik} = (x_{mik}, x_{fik})$, for the true unobserved x_{oik} for all trios $i = 1, \ldots, n_t$, the distribution of S_k is the same in all null hypotheses for each genetic model, each sampling plan and potential population admixture. Hence, the computations of p-values for S_k conditionally given the sufficient test statistics are the same for all models in the null hypothesis. That is, any p-value will result in rejecting the null hypothesis with a correct type 1 error rate, irrespective of which model is true in the null hypothesis.

Finally, S_k is standardized to a large sample normal or χ^2 approximation. By construction, since $\mathbb{E}_0(X_{oik}|x_{mik}, x_{fik})$ is centred around the true x_{oik} and under the assumption that the n_t trios are stochastically independent, S_k has an expected value of 0 under the null hypothesis for all tested marker $k = 1, \ldots, p$. Due to technological limitations, such as potential noisy data from a variety of genotyping platforms, Ionita-Laza et al. (2008) proposed to use the empirical variance estimator of X_{oik}, conditional on the parental intensity information, for standardization instead of the theoretical variance under the null hypothesis. Thus, the final FBAT statistic is obtained as

$$\text{FBAT}_k = \frac{\{\sum_{i=1}^{n_t} x_{oik} - \mathbb{E}_0(X_{oik}|x_{mik}, x_{fik})\}^2}{\sum_{i=1}^{n_t}[x_{oik} - \mathbb{E}_0(X_{oik}|x_{mik}, x_{fik})]^2} \sim \chi^2_1, \tag{5.4}$$

which is under the null hypothesis asymptotically (for $n_t \to \infty$) distributed according to a chi-squared distribution with one degree of freedom.

Of note, the FBAT approach presented in equation (5.4) is invariant under linear transformation of the hybridization intensity signals. Consequently, neither a marker specific normalization procedure (see chapter 3.2) nor a transformation of intensity measurements into raw copy number measurements, as described in chapter 6.1.1, have an impact on probe-wise FBAT results. As a consequence, the FBAT can likewise be applied to raw hybridization intensity values as well as to raw copy number values and moreover to any outcome of each intermediate data pre-processing step that is based on the idea of linear intensity transformations.

More generally, the FBAT statistic presented in equation (5.4) can also be derived as Rao's score test statistic of a prospective GLM for the offspring's phenotype (Lunetta et al., 2000), which is defined by

$$\mu_{ik} = \mathbb{E}(Y_i|x_{oik}) = g^{-1}(\beta_{0k} + \beta_{1k}x_{oik}) \quad \text{with} \quad k = 1, \ldots, p \text{ and } \beta_{0k}, \beta_{1k} \in \mathbb{R},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_{n_t})$ denotes the vector of random variables for the children's disease status with some appropriate link function $g(.)$ and where the distribution of Y_i $(i = 1, \ldots, n_t)$ is assumed to be a member of the canonical exponential family. Thus, the joint probability density function (p.d.f.) for Y at probe set $k = 1, \ldots, p$ is of the form

$$f(Y_1 = y_1, \ldots, Y_{n_t} = y_{n_t}, \theta_1, \ldots, \theta_{n_t}) = \exp\left[\sum_{i=1}^{n_t} \frac{y_i\theta_{ik} + \eta(\theta_{ik})}{h(\phi)} + \xi(y_i, \phi)\right],$$

with parameters β_{0k}, β_{1k} such that $g(\mu_{ik}) = \theta_{ik} = \beta_{0k} + \beta_{1k}x_{oik}$. Moreover, for each canonical GLM it is known that the first two moments of Y_i are $\mathbb{E}(Y_i) = \eta'(\theta_{ik})$ and $\mathrm{Var}(Y_i) = \eta''(\theta_{ik})h(\phi)$. For a dichotomous phenotype, $f(y_i, \theta_{ik})$ is from the Bernoulli distribution and $g(\cdot)$ is the natural logit link function as presented for the case-control design in chapter 5.1.1.1. Contrarily, $f(y_i, \theta_{ik})$ is normal and $g(\cdot)$ is the identity for continuous phenotypes in form of a quantitative trait. With $\mathcal{L}(y, \beta_{1k})$ denoting the log likelihood of $\mathbf{Y}$, the score S for β_{1k} of probe set $k = 1, \ldots, p$,

$$S(\beta_{1k}) = \frac{\partial\mathcal{L}(y, \beta_{1k})}{\partial\beta_{1k}} = \frac{\partial}{\partial\beta_{1k}}\left[\sum_{i=1}^{n_t} \frac{y_i\theta_{ik}(\beta_{1k}) - \eta(\theta_{ik}(\beta_{1k}))}{h(\phi)}\right] = \sum_{i=1}^{n_t} \frac{[y_i - \mathbb{E}(Y_i)]\, x_{oik}}{h(\phi)},$$

and the Fisher information $\mathcal{I}$ for β_{1k},

$$\mathcal{I}(\beta_{1k}) = \mathbb{E}\left[\frac{\partial^2\mathcal{L}(y, \beta_{1k})}{\partial\beta_{1k}^2}\right] = \sum_{i=1}^{n_t} \frac{\mathbb{E}[\mathrm{Var}(Y_i)\, x_{oik}^2]}{h(\phi)^2} = \frac{\sum[y_i - \mathbb{E}(Y_i)]^2\, x_{oik}^2}{h(\phi)^2},$$

the score test statistic to test the null hypothesis of no association is given by

$$R = \frac{S(\beta_{1k} = 0)^2}{\mathcal{I}(\beta_{1k} = 0)} = \frac{[\sum_{i=1}^{n_t}(y_i - \mu_k)\, x_{oik}]^2}{\sum_{i=1}^{n_t}(y_i - \mu_k)^2\, x_{oik}^2}.$$

In case of a trio design only affected offspring are considered, which implies that $y_i = 1$ for all families $i = 1, \ldots, n_t$. Thus, the term $(y_i - \mu_k) = (1 - \mu_k)$ vanishes from the score statistic R since it acts as a multiplicative constant for the trio design. The score statistic R then equals the FBAT statistic of equation (5.4) after application of one additional assumption: The offspring's genotypic intensities x_{oik} are assumed to consist of two orthogonal components

$$x_{oik} = b_{ik} + w_{ik} \quad \text{for} \quad k = 1, \ldots, p,$$

where b_{ik} denotes the between-family component and w_{ik} the within-family component, which is not sensitive to population structures and is statistically significant only in the presence of linkage (Lange et al., 2002). Here, b_{ik} represents the average within-family genotypic intensity level and is set to $b_{ik} = \mathbb{E}_0(X_{oik}|x_{mik}, x_{fik})$ (Lange et al., 2002). Hence, testing the null hypothesis $H_0 : \beta_{wk} = 0$ in the derived model

$$\mathbb{E}(Y_i|x_{oik}) = g^{-1}(\beta_{0k} + \beta_{bk} b_{ik} + \beta_{wk} w_{ik}) \quad \text{for} \quad k = 1, \ldots, p$$

yields the FBAT statistic presented in equation (5.4).

Although the trio design is no prospective study design with respect to the individual's affection status but instead involves sampling from $\mathbb{P}(x_{oik}|Y_i)$, the FBAT statistic was above developed as a score test statistic by modeling $\mathbb{P}(Y_i|x_{oik})$. However, as stated by Lunetta et al. (2000), these two approaches are equivalent and, in case of the trio design, result in the same tests. Advantages of the derivation of FBAT_k as a score statistic is that additional covariates, such as other known risk factors, can easily be incorporated and that effect size estimates are indirectly given.

5.1.2.2 Multi-Marker CNV FBATs

CNVs typically span more than one array probe set. Referring to this, Ionita-Laza et al. (2008) additionally proposed a multivariate extension of the above FBAT, which incorporates information on multiple adjacent probe sets. As the single-marker FBAT, the multi-marker FBAT is a conditional score test that conditions upon the within-family information of parental intensity data. The score for the z-variate score test is given by

$$S_{[z]} = \sum_{i=1}^{n_t} \begin{pmatrix} x_{oi1} - \mathbb{E}_0(X_{oik}|x_{mi1}, x_{fi1}) \\ \vdots \\ x_{oiz} - \mathbb{E}_0(X_{oik}|x_{miz}, x_{fiz}) \end{pmatrix} \quad \text{with} \quad z < p.$$

The z-variate FBAT statistic, including information on z adjacent marker, has the form

$$\text{FBAT}_{[z]} = S_{[z]}^t \hat{V}^{-1} S_{[z]} \sim \chi_z^2,$$

where $\hat{V}$ denotes the empirical variance-covariance matrix with rank z.

5.1.2.3 Marker Selection for Family-Based Association Testing

In order to guarantee at least minimal variability in copy number, say $\epsilon > 0$, for the tested probe sets, only those probe sets

$$\{s_1, \ldots, s_w\} \subset \{1, \ldots, p\} \quad \text{with} \quad w < p$$

for which

$$\frac{|\{p \mid (p \leq 2n_t) \wedge (\text{ parent } p \text{ has a CNV at probe set } s_j)\}|}{2\,n_t} > \epsilon \quad \text{for} \quad j = 1, \ldots, w$$

are tested for an association with disease. The lower limit for copy number variability of each tested probe set is referred to the parents only, because a sample-based approach might overestimate the true CNV-variability in regions where CNVs are preferentially transmitted to the affected offspring.

5.2 Application of Strategy PS1 to the Phenotype Obesity

Application of strategy S1 to genome-wide raw CNV data of a family-based obesity sample comprising 424 obesity trios revealed no evidence for any association between CNVs and the trait obesity. In this chapter, results of applying the proposed analysis strategy PS1 to exactly the same data set will be presented. Parts of the genetic results of this chapter have been published in Jarick et al. (2011).

5.2.1 Data Set

The family-based obesity sample comprised 424 nuclear families with one extremely obese child or adolescent and both biological parents (for details see chapter 3.5.1 and Jarick et al. (2011), Supplementary Table S1). For statistically significant findings from the family-based obesity sample, follow-up analyses were performed in a case-control sample of 453 extremely obese children and adolescents (cases) and 435 normal-weight or lean adult control subjects (controls) (for details see chapter 3.5.1 and Jarick et al. (2011), Supplementary Table S1). Finally, additional follow-up analyses for most promising findings were performed in an independent sample of 365 obesity trios, which was recruited similarly as the family-based discovery sample (for details see Jarick et al. (2011), Supplementary Table S1).

5.2.2 Methods

Data pre-processing.

For each individual of the family-based as well as of the case-control sample genotyping was performed on the Affymetrix 6.0 chip by the ATLAS Biolabs GmbH (Berlin, Germany). Afterwards, raw hybridization intensity signals of the $\sim 890\,000$ autosomal CN probe sets were extracted from the individual '.CEL' files by use of the R-package 'affxparser' (Bengtsson et al., 2008a). Prior to association testing, the raw intensity signals of the case-control sample were quantile normalized (Bolstad et al., 2003) to account for potential plate effects, whereas, raw fluorescence intensity signals of the family-based samples were directly incorporated into statistical tests since the family-based design and the family-wise assignment to genotyping plates allows for a control of the inter-individual variability.

CNV calling.

As secondary analyses in both samples, CNVs were estimated at a genome-wide level by use of the Affymetrix Genotyping Console (GTC) 3.0 (Korn et al., 2008; McCarroll et al., 2008). As described in chapter 3.5.2, the HMM algorithm of the GTC software is based on comparing individual signal intensity levels against those of a reference sample. The reference sample size was limited to 106 parental pairs due to computational constraints. As explained previously, two differently composed reference samples were used for CNV frequency estimation to address potential reference group effects on CNV calls. In more detail, one randomly designed set and a second set of those 106 parental pairs with minimal mean BMI standard deviation scores were considered as reference groups (for details see chapter 3.5.2 and Jarick et al. (2011), Supplementary Table S1). Only CNV calls that were consistently assigned via both approaches were investigated subsequently.

Marker selection for association testing.

The first essential step of strategy PS1, the selection of probe sets for association testing, was performed on the basis of a lower frequency threshold for estimated CNVs per probe set. In more detail, only CN probe sets in estimated CNV regions (CNVRs) with at least five percent copy number variability were tested for an association with obesity. Being defined as a region of overlapping CNVs, a CNVR was claimed to consist of at least three consecutive CN probe sets. Moreover, CNV frequencies were separately considered in offspring and parents of the family-based

discovery sample as well as in cases and controls of the case-control sample, and the lower CNV frequency limit of five percent was applied to each of the four sub-groups separately.

Association testing.

Afterwards, the CNV FBAT approach was initially applied in the family-based GWAS discovery sample of 424 obesity trios to each previously selected CN probe set. Subsequently, significant probe sets were identified with regard to the lfdr method by use of the central matching estimation method implemented in the R-package 'locfdr' (Efron et al., 2011), which was described in detail in chapter 3.5.2.

Follow-up analyses were performed in the case-control sample for each probe set within CNVRs with lfdr significant CNV FBAT results by applying logistic regression with predictors normalized intensities, sex and age. In the follow-up analyses, significance of probe sets was again determined by use of the lfdr method. The number of probe sets that was analysed for follow-up in the case-control sample was relative low. Consequently, the 'locfdr' function of the R-package 'locfdr' was applied with non-default parameters 'bre', 'df' and 'type', which were selected to minimize the differences between maximum likelihood and central matching estimation.

For fine-scale analyses, multi-marker FBATs as described in chapter 5.1.2.1 were applied in the family-based discovery sample to each CNVR with lfdr significant results in the family-based discovery sample as well as in the case-control follow-up sample. For each such selected CNVR, any subset of consecutive CN probe sets was incorporated into multi-marker FBATs. In order to allow a comparison to single marker FBAT results, multi-marker z-values were derived in application of an inverse standard normal transformation on p-values.

Moreover, linkage disequilibrium (LD) in form of the squared pairwise Pearson's correlation coefficients of parental intensities from the family-based discovery sample was calculated for all probe sets in CNVRs with lfdr significance in the family-based discovery sample as well as in the case-control follow-up sample. Finally, recombination rates are reported with respect to data of the 1000 Genomes project (www.1000genomes.org).

CNV validation and replication.

In order to ensure reliability of chip-based association test results and to additionally specify precise individual copy number states, the most promising newly identified CNV region at chr 11q11 with evidence for an association with the binary trait obe-

sity was on the one hand technically validated in the family-based discovery sample by use of the qPCR technology. On the other hand, this region was furthermore followed-up in a second family-based obesity sample of 365 independent obesity trios, which was likewise analysed by use of qPCR. For both trio samples, qPCR derived copy number number states were tested for an association with obesity in application of the FBAT approach by assuming an additive genetic effect model. More precisely, the coding for the three observed CN marker genotypes was specified as 0, 1, 2 in concordance with the determined total number of DNA copies.

5.2.3 Results

Genome-wide CNV analyses

A total of 244 autosomal CNVRs comprising 8 051 CN probe sets were detected and tested for an association with obesity. The majority of CNVRs (n = 240) was listed in the DGV (http://projects.tcag.ca/variation). Details on location, size and marker distribution of the CNVRs can be found in Jarick et al. (2011) (Table 1).

Genome-wide association testing of the 8 051 selected CN probe sets in the family-based GWAS discovery sample of 424 obesity trios revealed eight probe sets with lfdr below 0.20, which is a sensible threshold as proposed by Efron (2004) (Figure 5.2 and Figure 5.3). The eight lfdr significant probe sets are located in seven CNVRs (Table 5.1). The only lfdr significant CNVR that contained more than one lfdr significant probe set, at chr 11q11, also harboured the probe set with minimal FBAT p-value (CN_063559, p-value = 0.0074).

A total of 291 probe sets in those seven CNVRs with lfdr significant FBAT results were analysed for follow-up in an independent case-control sample. Logistic regression analyses of the 291 probe sets yielded eight significant probe sets with lfdr values below 0.2 in a total of four CNVRs (Figure 5.4, Table 5.2). Only for one of these four CNVRs, at chr 11q11, the lfdr significant negative effect direction observed in the trio sample was consistently re-observed with lfdr significance in the case-control sample. The remaining three CNVRs showed contradictory lfdr significant effects in the two analysed samples.

For a low number of simultaneously performed tests, the lfdr method has only limited applicability. In order to address this limitation, follow-up analyses were repeated in application of a more relaxed lfdr discovery threshold, which resulted in a higher number of follow-up probe sets for the re-analyses in the case-control sample. For a more conservative threshold of lfdr < 0.3, 26 probe sets in 14 CNVRs showed significant FBAT results in the trio sample (Figure 5.5 and Table 5.3).

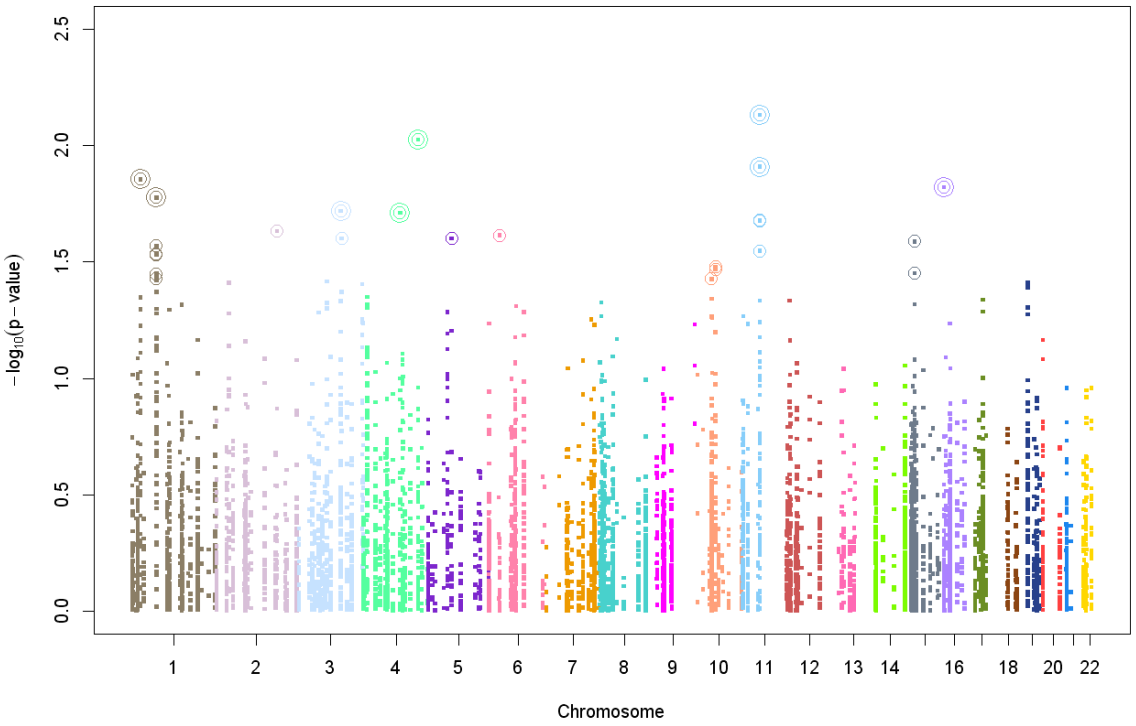

Figure 5.2: Manhattan plot for the genome-wide CNV analysis of 424 obesity trios accounting for 8 051 CN probe sets in 244 CNVRs. For all probe sets, the $-\log_{10}$ transformed CNV FBAT p-values are shown relative to their chromosomal position. 26 (eight) FBAT results with lfdr < 0.3 (< 0.2) are circled (twice).

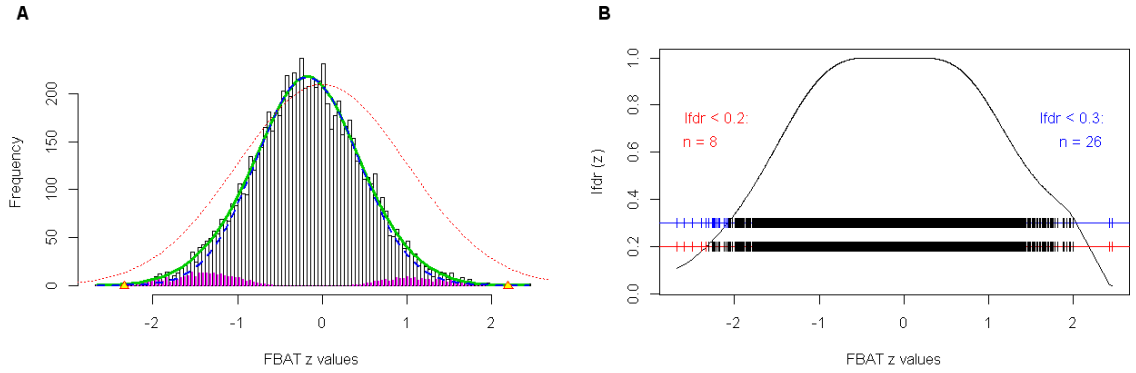

Figure 5.3: Histogram and lfdr curve of CNV FBAT z-values for the genome-wide analysis of 424 obesity trios at 8 051 CN probe sets in 244 CNVRs. Panel A: Histogram. The red dashed curve depicts the standard normal distribution, the dashed blue line is $\hat{p}_0\hat{f}_0$, the empirical null density, $\mathcal{N}(-0.172, 0.602^2)$, and the green line is the empirically estimated mixture density. The small pink bars represent estimated non-null counts. Panel B: Lfdr curve, derived from empirical estimates of f_0, f and p_0 (Panel A). Observed CNV FBAT z-values are illustrated as ticks on the horizontal lines, those with lfdr < 0.2 (< 0.3) are printed in red (blue).

Table 5.1: Results for 26 (8) probe sets with lfdr < 0.3 (0.2) across 8 051 CNV FBATs in 424 obesity trios. (in chromosomal order)

CN probe set ID	Chr: Position (hg18)	Cytoband [hg18]	FBAT z-value	FBAT p-value	lfdr	in known CNV (DGV)	in CNVR [chr: bp, hg18]	# probe sets in CNVR
CN_020771	1 : 25 484 027..25 484 052	p36.11	2.458	0.0140	0.034	YES	1 : 25 468 522 – 25 534 812	29
CN_517829	1 : 72 543 719..72 543 744	p31.1	−2.212	0.0269	0.245	YES	1 : 72 541 074 – 72 583 749	47
CN_517834	1 : 72 547 710..72 547 735	p31.1	−2.101	0.0356	0.290	YES	1 : 72 541 074 – 72 583 749	47
CN_517839	1 : 72 551 683..72 551 708	p31.1	−2.392	0.0168	0.181	YES	1 : 72 541 074 – 72 583 749	47
CN_519935	1 : 72 581 295..72 581 320	p31.1	−2.179	0.0293	0.258	YES	1 : 72 541 074 – 72 583 749	47
CN_519936	1 : 72 581 948..72 581 973	p31.1	−2.082	0.0373	0.299	YES	1 : 72 541 074 – 72 583 749	47
CN_519938	1 : 72 582 418..72 582 443	p31.1	−2.101	0.0356	0.290	YES	1 : 72 541 074 – 72 583 749	47
CN_519939	1 : 72 583 514..72 583 539	p31.1	−2.177	0.0295	0.259	YES	1 : 72 541 074 – 72 583 749	47
CN_818148	2 : 184 502 773..184 502 798	q32.1	−2.268	0.0234	0.224	YES	2 : 184 502 747 – 184 510 699	16
CN_978208	3 : 131 275 089..131 275 114	q21.3	−2.343	0.0191	0.197	YES	3 : 131 245 537 – 131 290 979	56
CN_993322	3 : 133 475 914..133 475 939	q22.1	−2.240	0.0251	0.234	YES	3 : 133 475 451 – 133 478 387	3
CN_1034271	4 : 108 291 065..108 291 090	q25	−2.336	0.0195	0.200	YES	4 : 108 285 188 – 108 293 270	25
CN_1063738	4 : 161 286 278..161 286 303	q32.1	−2.595	0.0095	0.121	YES	4 : 161 282 532 – 161 289 730	13
CN_1139749	5 : 70 290 649..70 290 674	q13.2	−2.240	0.0251	0.234	YES	5 : 68 903 038 – 70 343 313	22
CN_1175510	6 : 32 638 110..32 638 135	p21.32	−2.251	0.0244	0.230	YES	6 : 32 560 895 – 32 638 289	25
CN_524300	10 : 46 478 786..46 478 811	q11.22	−2.080	0.0375	0.300	YES	10 : 46 338 178 – 46 812 351	161
CN_548211	10 : 58 195 507..58 195 532	q21.1	−2.119	0.0341	0.283	YES	10 : 58 186 369 – 58 196 856	40
CN_548214	10 : 58 195 736..58 195 761	q21.1	−2.128	0.0334	0.279	YES	10 : 58 186 369 – 58 196 856	40
CN_587558	11 : 55 130 612..55 130 637	q11	−2.502	0.0123	0.147	YES	11 : 55 130 596 – 55 210 165	58
CN_587579	11 : 55 153 205..55 153 230	q11	−2.307	0.0211	0.210	YES	11 : 55 130 596 – 55 210 165	58
CN_589638	11 : 55 196 554..55 196 579	q11	−2.308	0.0210	0.209	YES	11 : 55 130 596 – 55 210 165	58
CN_589644	11 : 55 203 896..55 203 921	q11	−2.191	0.0284	0.253	YES	11 : 55 130 596 – 55 210 165	58
CN_063559	11 : 55 204 029..55 204 054	q11	−2.679	0.0074	0.106	YES	11 : 55 130 596 – 55 210 165	58
CN_685264	15 : 28 339 425..28 339 450	q13.2	−2.228	0.0259	0.239	YES	15 : 28 280 641 – 28 609 063	103
CN_119211	15 : 28 595 222..28 595 247	q13.2	−2.103	0.0355	0.290	YES	15 : 28 280 641 – 28 609 063	103
CN_721778	16 : 14 897 328..14 897 353	p13.11	2.430	0.0151	0.037	YES	16 : 14 796 084 – 14 987 969	63

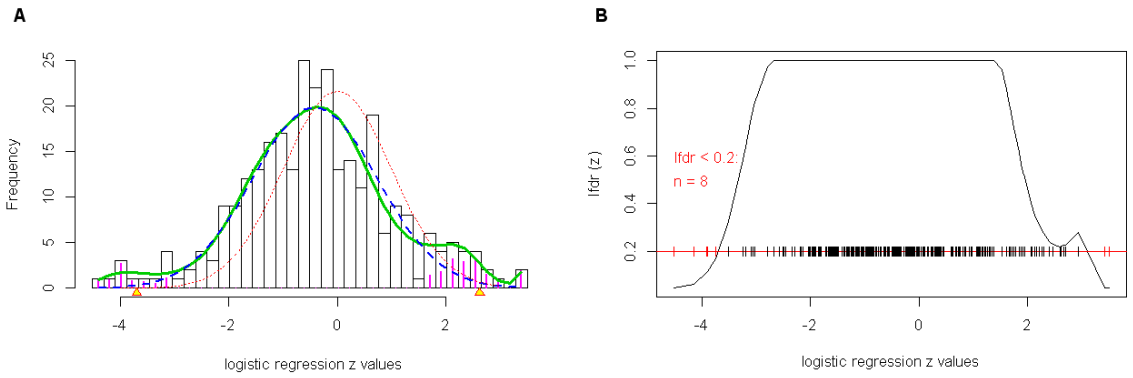

Figure 5.4: Histogram and lfdr curve for logistic regression z-values of 291 CN probe sets at seven CNVRs in a case-control follow-up sample. Panel A: Histogram. The red dashed curve depicts the standard normal distribution, the dashed blue line is $\hat{p}_0\hat{f}_0$, the empirical null density, $\mathcal{N}(-0.409, 1.147^2)$, and the green line is the empirically estimated mixture density. The small pink bars represent estimated non-null counts. Panel B: Lfdr curve, derived from empirical estimates of f_0, f and p_0 (Panel A). Observed logistic regression z-values are illustrated as ticks on the horizontal lines, those with lfdr < 0.2 are printed in red.

Table 5.2: Results for eight probe sets with lfdr < 0.2 across 291 logistic regression tests in 453 obesity cases and 435 lean controls. (in chromosomal order) Probe sets with directionally consistent, lfdr significant effect in the family-based sample are highlighted in red. Each probe represents 25 bp.

CN probe set ID	Chr: Position [hg18]	FBAT z-value	FBAT p-value	lfdr	in CNVR [chr: bp, hg18]
CN_484327	$1:25\,500\,952^{+}$	-3.750	1.77×10^{-4}	0.168	$1:25\,468\,522-25\,534\,812$
CN_980258	$3:131\,275\,502^{+}$	-3.915	9.03×10^{-5}	0.110	$3:131\,245\,537-131\,290\,979$
CN_980259	$3:131\,276\,124^{+}$	-3.896	9.78×10^{-5}	0.116	$3:131\,245\,537-131\,290\,979$
CN_980277	$3:131\,289\,676^{+}$	3.506	5.55×10^{-4}	0.046	$3:131\,245\,537-131\,290\,979$
CN_587564	$11:55\,132\,844^{+}$	-4.162	3.16×10^{-5}	0.064	$11:55\,130\,596-55\,210\,165$
CN_721771	$16:14\,844\,813^{+}$	-3.898	9.69×10^{-5}	0.115	$16:14\,796\,084-14\,987\,969$
CN_721784	$16:14\,956\,349^{+}$	3.425	6.16×10^{-4}	0.046	$16:14\,796\,084-14\,987\,969$
CN_723869	$16:14\,968\,128^{+}$	-4.518	6.24×10^{-6}	0.046	$16:14\,796\,084-14\,987\,969$

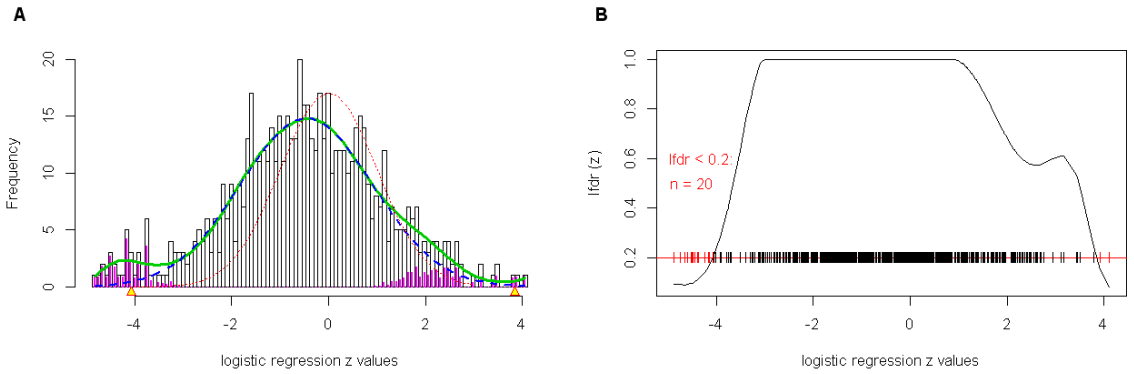

Figure 5.5: Histogram and lfdr curve for logistic regression z-values of 661 CN probe sets at 14 CNVRs in a case-control follow-up sample. Panel A: Histogram. See Figure 5.4 for a detailed description. The empirical null density is $\mathcal{N}(-0.458, 1.369^2)$. Panel B: Lfdr curve, derived from empirical estimates of f_0, f and p_0 (Panel A). Observed logistic regression z-values are illustrated as ticks on the horizontal lines, those with lfdr < 0.2 are printed in red.

Table 5.3: Results for 20 probe sets with lfdr < 0.2 across 661 logistic regression tests in 453 obesity cases and 435 lean controls. (in chromosomal order) Probe sets with directionally consistent, lfdr significant effect in the family-based sample are highlighted in red. Each probe represents 25 bp.

CN probe set ID	Chr: Position [hg18]	FBAT z-value	FBAT p-value	lfdr	in CNVR [chr: bp, hg18]
CN_303769	$5:69\,269\,171^+$	-4.243	$2.20\times10{-}5$	0.138	$5:68\,903\,038-70\,343\,313$
CN_558684	$10:46\,376\,766^+$	-4.487	$7.21\times10{-}6$	0.097	$10:46\,338\,178-46\,812\,351$
CN_524300	$10:46\,478\,786^+$	-4.498	$6.87\times10{-}6$	0.096	$10:46\,338\,178-46\,812\,351$
CN_558758	$10:46\,478\,929^+$	-4.135	$3.54\times10{-}5$	0.173	$10:46\,338\,178-46\,812\,351$
CN_558759	$10:46\,478\,952^+$	-4.437	$9.12\times10{-}6$	0.103	$10:46\,338\,178-46\,812\,351$
CN_524302	$10:46\,479\,180^+$	-4.754	$2.00\times10{-}6$	0.091	$10:46\,338\,178-46\,812\,351$
CN_524305	$10:46\,479\,195^+$	-4.658	$3.20\times10{-}6$	0.090	$10:46\,338\,178-46\,812\,351$
CN_524306	$10:46\,479\,220^+$	-4.382	$1.18\times10{-}5$	0.110	$10:46\,338\,178-46\,812\,351$
CN_524301	$10:46\,479\,382^+$	-4.585	$4.54\times10{-}6$	0.091	$10:46\,338\,178-46\,812\,351$
CN_524303	$10:46\,479\,399^+$	-4.145	$3.40\times10{-}5$	0.169	$10:46\,338\,178-46\,812\,351$
CN_524304	$10:46\,479\,414^+$	-4.651	$3.31\times10{-}6$	0.090	$10:46\,338\,178-46\,812\,351$
CN_524310	$10:46\,479\,759^+$	-4.363	$1.28\times10{-}5$	0.113	$10:46\,338\,178-46\,812\,351$
CN_524312	$10:46\,479\,800^+$	-4.073	$4.65\times10{-}5$	0.198	$10:46\,338\,178-46\,812\,351$
CN_524320	$10:46\,487\,406^+$	-4.135	$3.54\times10{-}5$	0.173	$10:46\,338\,178-46\,812\,351$
CN_524343	$10:46\,522\,903^+$	-4.141	$3.46\times10{-}5$	0.171	$10:46\,338\,178-46\,812\,351$
CN_587564	$11:55\,132\,844^+$	-4.162	$3.16\times10{-}5$	0.163	$11:55\,130\,596-55\,210\,165$
CN_685275	$15:28\,377\,334^+$	3.914	$9.09\times10{-}5$	0.152	$15:28\,280\,641-28\,609\,063$
CN_685278	$15:28\,383\,727^+$	-4.870	$1.12\times10{-}6$	0.093	$15:28\,280\,641-28\,609\,063$
CN_685298	$15:28\,439\,410^+$	4.095	$4.22\times10{-}5$	0.082	$15:28\,280\,641-28\,609\,063$
CN_723869	$16:14\,968\,128^+$	-4.518	$6.24\times10{-}6$	0.095	$16:14\,796\,084-14\,987\,969$

Compared to a more stringent lfdr discovery threshold of 0.3, probe sets in seven additional CNVRs where incorporated into repeated case-control follow-up analyses, which resulted in a total of 661 probe sets that were again tested in a logistic regression context. Concerning these follow-up investigations, application of the lfdr method identified 20 probe sets in five CNVRs to be significant in the follow-up case-control sample. For three CNVRs, at chr 5q13.2, chr 10q11.22 and chr 11q11, significant follow-up effects are directionally consistent with negative discovery effects previously observed the family-based sample.

Results of the CNVR at chr 11q11, which was the only CNVR that was initially detected to be associated with obesity in the family-based discovery sample and directionally consistent re-found in follow-up analyses of the case-control sample in application of a stringent lfdr discovery threshold of 0.2, remained stable after inclusion of a larger set of follow-up probe sets when applying a more relaxed lfdr discovery threshold of 0.3 (Table 5.3).

Additionally, follow-up analyses in the case-control sample accounting for a relaxed lfdr discovery threshold of 0.3 yielded two more CNVRs (at chr 5q13.2 and at chr 10q11.22) with lfdr significant follow-up results that reflect directionally consistent effects relative to those observed in the family-based discovery sample (Table 5.3). One of these two additional CNVRs, at chr 10q11.22, was previously reported to be the only CNV that was associated with BMI in a genome-wide CNV association study for a sample of 597 elderly Chinese Han subjects (Sha et al., 2009).

Exploration of statistically significant CNVRs

Details on association test results as well as on the correlation structure of a total of those seven CNVRs with lfdr significant probe sets in the family-based GWAS discovery sample and with additional lfdr significant follow-up results with respect to a lfdr discovery threshold of either 0.2 or 0.3 are presented in Figures 5.6 - 5.12.

There is an enrichment of low p-values with negative effect direction near the most significant follow-up probe set at CNVR chr 10q11.22 (Figure 5.9). This enrichment is regionally limited by two recombination peaks. Multi-marker FBAT results in this LD-block are considerably more significant than single marker tests with effects that are likewise of negative direction. In sum, these results suggest that deletions in this region at CNVR chr 10q11.22 seem to be associated with obesity, which is in concordance with results of a previous report (Sha et al., 2009).

The second CNVR that showed multi-marker FBAT results with remarkably lower p-values compared to single-marker tests is the CNVR at chr 11q11 (Figure 5.10). Of note, multi-marker and single-marker FBATs as well as significant logistic regression

tests have almost exclusively negative effect directions in this region. Thus, deletions at CNVR chr 11q11 are suggested to be associated with obesity.

Similarly, results indicated that deletions at CNVR chr 5q13.2 might be associated with obesity (Figure 5.8). Although, this conclusion has to be constrained by large marker gaps of up to ten kb and a low correlation structure across probe sets in the respective region.

For the remaining four CNVRs, at chr 1p36.11, 3q21.3, 15q13.2 and 16p13.11, there is neither a regional enrichment of low p-values nor are multi-marker FBAT p-values lower than single-marker FBAT p-values. In addition, there is no concordance in the effect direction of probe sets with significant test results between the discovery and the follow-up analyses (Figures 5.6, 5.7, 5.11 and 5.12). Thus, an association of CNVs at these four regions with obesity does not seem to be plausible.

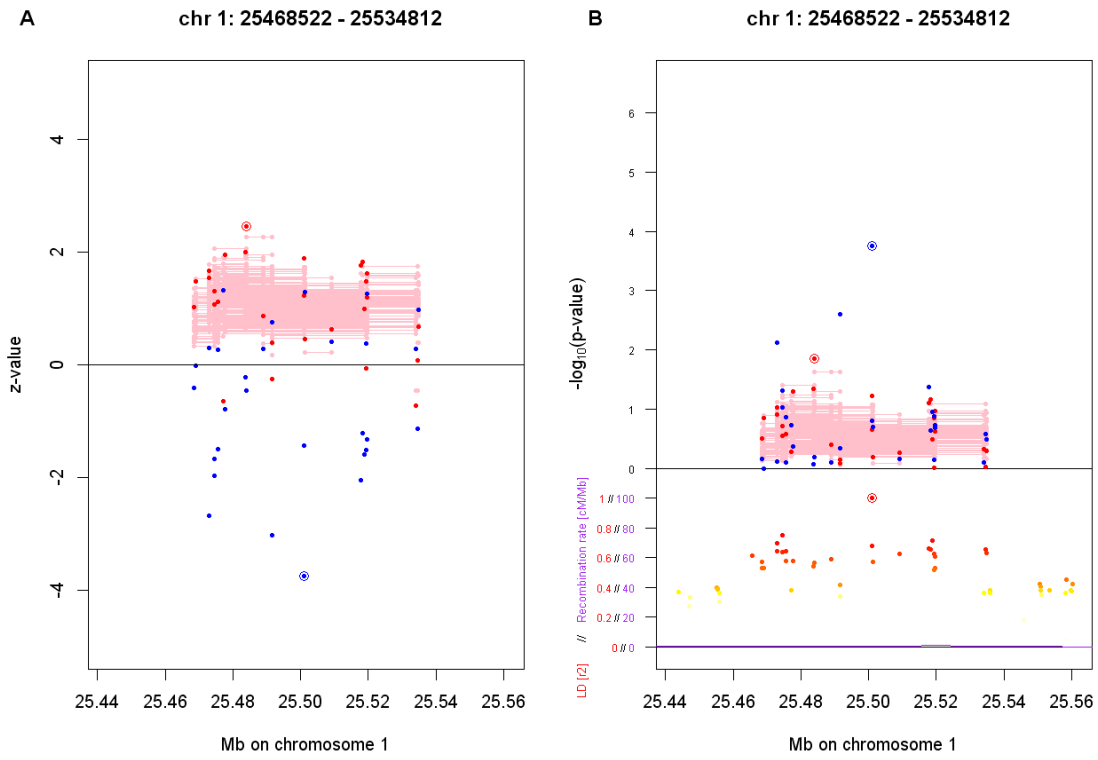

Figure 5.6: Association test results for CNVR chr 1p36.11. Panel A: Z-values of CNV FBATs (logistic regression tests) are depicted in red (blue). Lfdr significant results are circled. Panel B: P-values of CNV FBATs (logistic regression tests) are shown in red (blue) in the upper part of the panel. Lfdr significant results are circled. Recombination rates and pairwise linkage disequilibrium values relative to the probe set with minimal p-value are presented in the lower part of the panel.

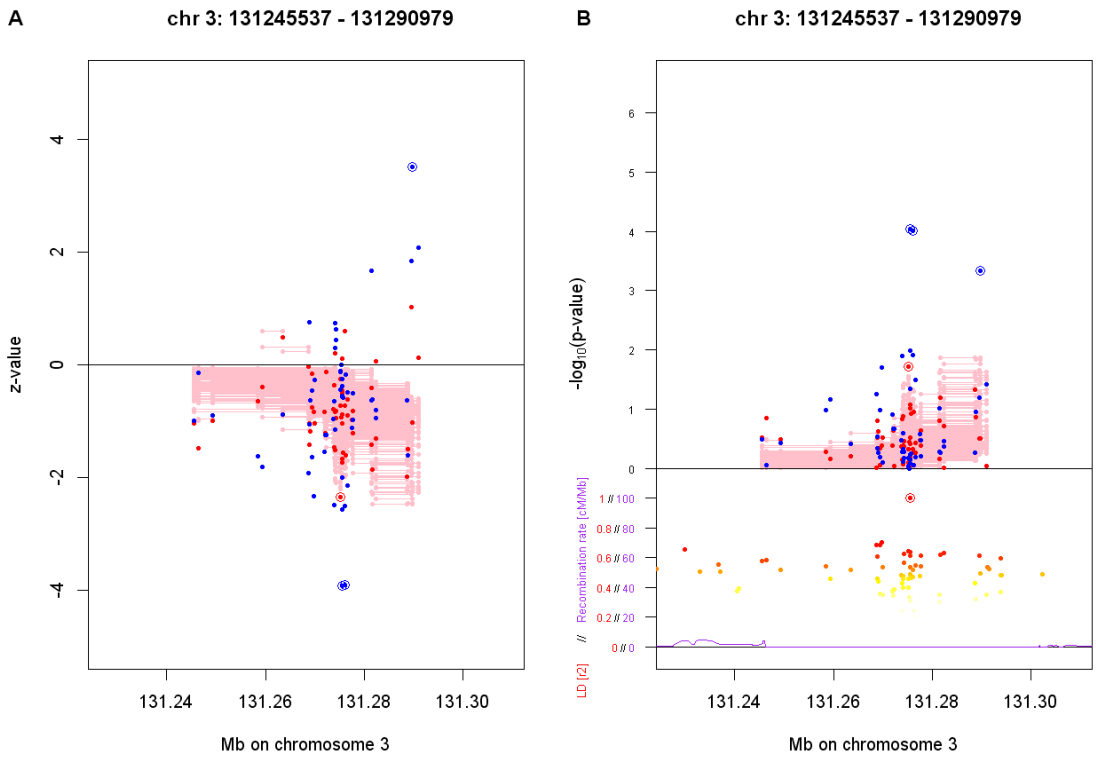

Figure 5.7: Association test results for CNVR chr 3q21.3. (details at Figure 5.6)

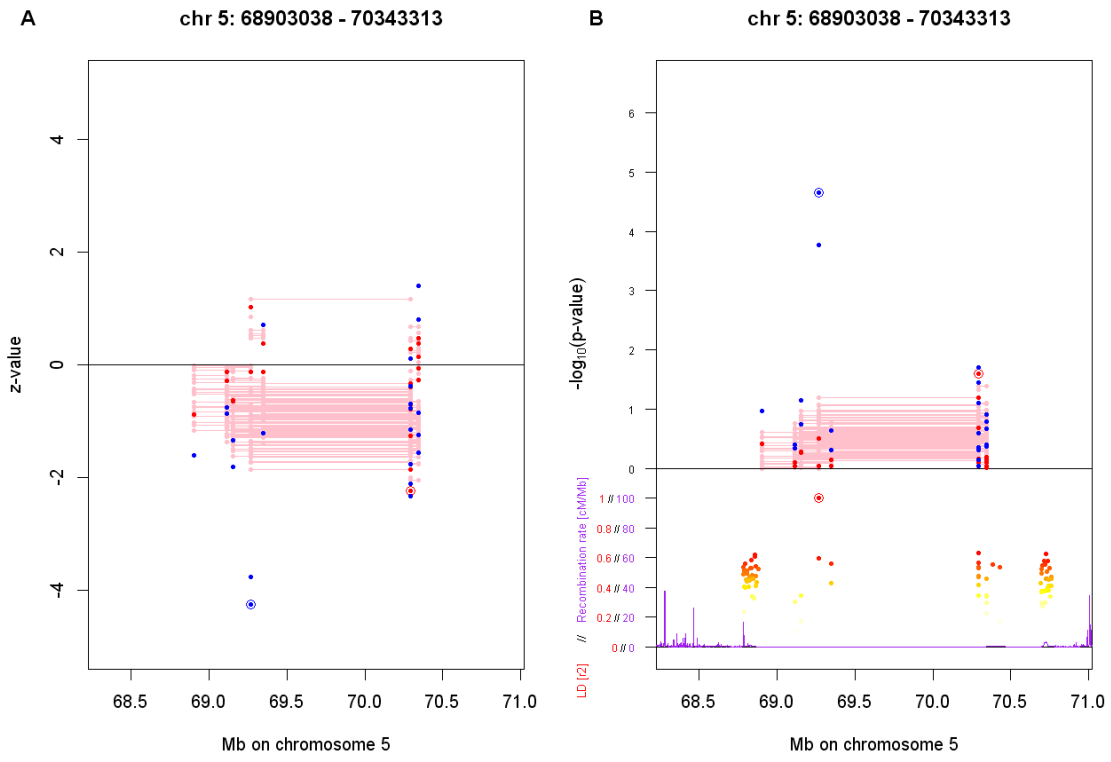

Figure 5.8: Association test results for CNVR chr 5q13.2. (details at Figure 5.6)

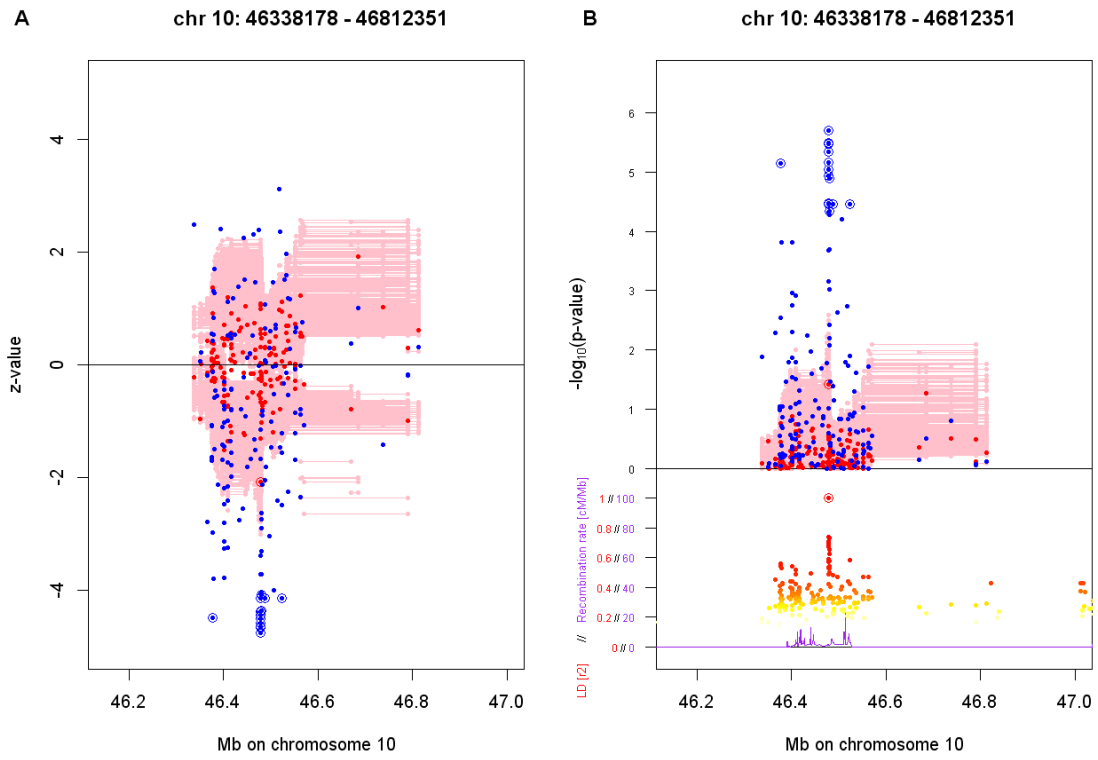

Figure 5.9: Association test results for CNVR chr 10q11.22. (details at Figure 5.6)

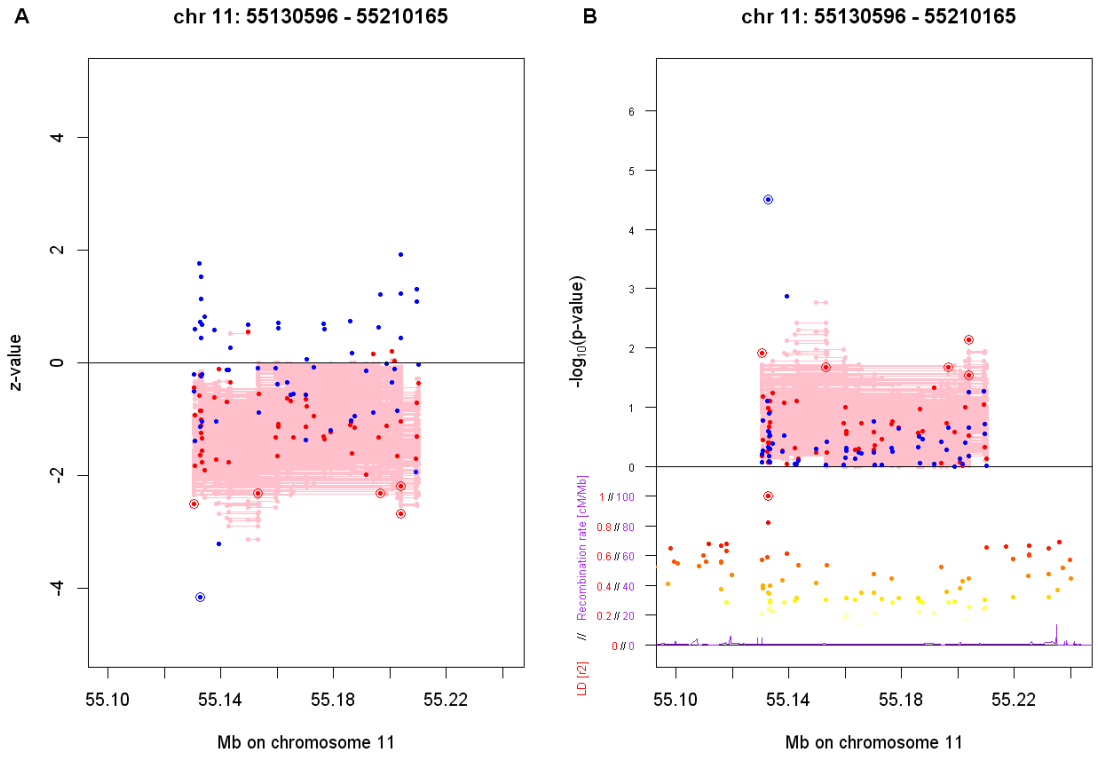

Figure 5.10: Association test results for CNVR chr 11q11. (details at Figure 5.6)

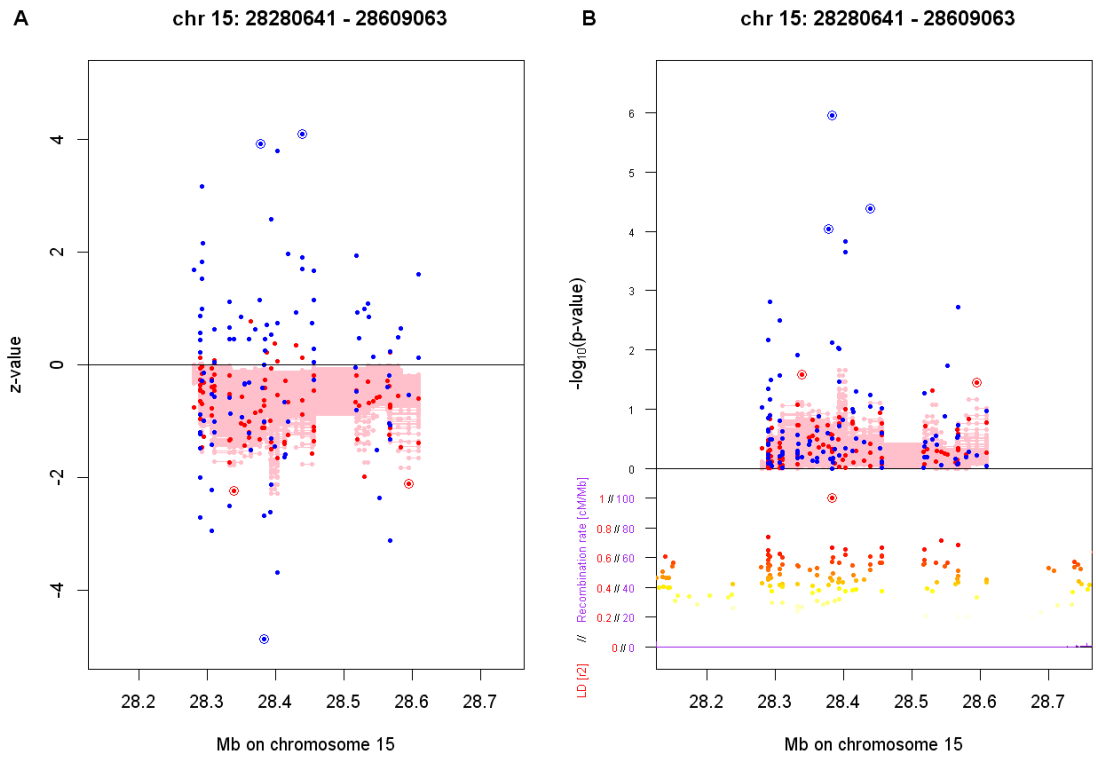

Figure 5.11: Association test results for CNVR chr 15q13.2. (details at Figure 5.6)

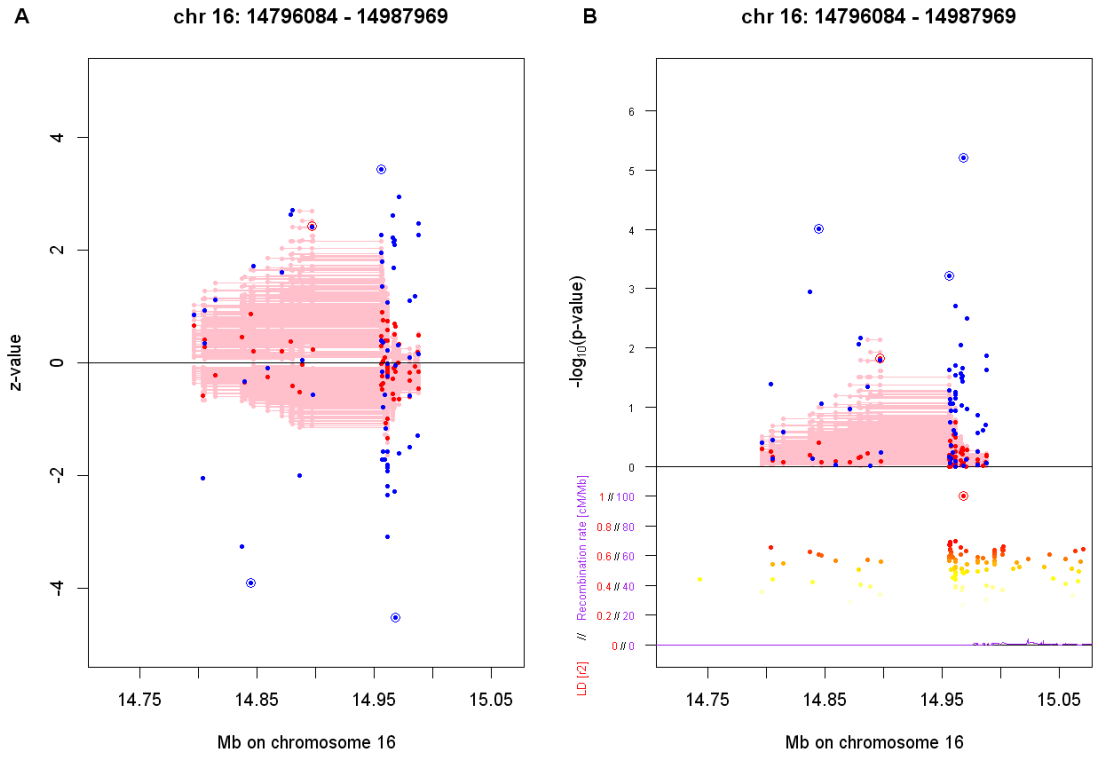

Figure 5.12: Association test results for CNVR chr 16p13.11. (details at Figure 5.6)

Validation and replication of CNVR chr 11q11

Among those seven CNVRs with statistically significant results in the family-based discovery as well as in the case-control follow-up sample, the CNVR at chr 11q11 showed the lowest p-value ($p = 0.0074$) in the family-based discovery sample. Thus, this region represents the most promising newly identified CNVR with association to obesity. Consequently, this region at chr 11q11 was further analysed by use of the qPCR technique.

The observed effect direction of deletions at CNVR chr 11q11 being associated with obesity could be validated in the family-based discovery sample by qPCR analyses. With regard to qPCR findings, this region could be identified as a bi-allelic deletion region. In more detail, 7.71% (9.72%) homozygotes and 40.35% (40.03%) heterozygotes for the deletion were observed among the parents (offspring) of the 424 discovery trios. Here, the slight increase of homozygous deletions in children compared to parents is consistent with the initially observed transmission disequilibrium. Association testing of the qPCR derived copy number states indicated a trend towards a preferable transmission of the deletion allele to the obese children (OR = 1.171, 95% confidence interval = $0.947 - 1.448$, one-sided p-value = 0.066).

Similarly, qPCR-based analyses in the family-based follow-up sample of 365 independent obesity trios revealed a directionally consistent trend (OR = 1.214, 95% confidence interval = $0.959 - 1.537$, one-sided p-value = 0.066). Finally, all 789 obesity trios (discovery and follow-up sample), for which qPCR-derived CNV states were available, were jointly analysed to increase the precision of the effect size estimator, resulting in an OR of 1.190 for the deletion allele (95% confidence interval = $1.016 - 1.394$, one-sided p-value = 0.015).

5.2.4 Discussion

In application of the proposed analyses strategy PS1, a genome-wide CNV association study for the phenotype obesity was performed in a family-based discovery sample and a case-control follow-up sample. Association tests were restrictedly applied to 8 051 pre-selected probe sets in 244 estimated CNVRs. Significance was assessed via the lfdr method by Ionita-Laza et al. (2008). Deletions in two CNVRs, at chr 10q11.22 and at chr 11q11, were identified to be associated with obesity. One of these two findings, the CNVR at chr 10q11.22, was concordantly reported with an association effect on BMI in a previous genome-wide CNV study of 597 elderly Chinese Han subjects (Sha et al., 2009).

The application of strategy S1 resulted in no statistically significant finding of any CNV - obesity association at all (see chapter 3.5). Contrarily, the only previous

report concerning associations between common CNVs and obesity on a CNV at chr 10q11.22 to be associated with BMI (Sha et al., 2009) could be strengthened in use of the proposed strategy PS1. Moreover, the results allow additional support for a newly identified association of deletions at another region, at chr 11q11, with obesity.

Consequently, the proposed analyses strategy PS1 yielded more useful results and is therefore clearly to be favored over strategy S1 in the underlying situation. However, it has to be kept in mind that such real data applications do not allow any general conclusions regarding power or type 1 error levels.

Being aware of testing CNV signals exclusively in known CNV regions may be an useful approach to explicitly alleviate the multiple testing issue (Ionita-Laza et al., 2008), to our knowledge, this has not been investigated systematically previously. Here, at least a data-driven investigation of this approach is provided. In particular, the previous straightforwad application of the well established genome-wide CNV analysis strategy S1 (Ionita-Laza et al., 2008) to the phenotype obesity (see chapter 3.5) allows a comparison of both strategies, S1 and PS1.

Besides relaxing the multiple testing issue, a further advantage of strategy PS1 over strategy S1 is that CNV probe sets can be assigned to different CNVs or CNVRs based on estimated CNV population frequency estimates, which are available from strategy PS1's marker selection step. Thus, follow-up analyses have no longer to be restricted to exactly the same probe sets, which were initially discovered to be statistically significantly associated with the trait of interest. Instead, each probe set that is covered by a CNV with detected significance in the discovery sample at any involved probe set was incorporated in follow-up analyses of the case-control sample. However, due to this procedure special care is needed with regard to the precise correlation structure of the respective CNVs. In order to ensure that statistically significant discovery and follow-up signals come from exactly the same CNV loci, recombination rates were considered across the entire CNV regions.

In the presented exemplary analysis, strategy PS1 involved considerably more expensive computing efforts in identifying those probe sets that represent common CNVs. Although the actual individual CNV genotype does not need to be collected since these are not involved in association testing, at the very least, sample-specific realistic CNV frequencies have to be provided for the marker selection step. Here, this was realized in calling CNVs in the very same sample that was later on tested for CNV-trait associations. Thus, the leading motivation for the specific design of strategy S1, which is to overcome the inaccuracy in CNV detection from SNP genotyping array data, is to some extend returned back into the proposed strategy PS1.

With regard to current suggestions (Winchester et al., 2009) and aiming to minimize the risk of false CNV classifications, CNV calling was repeatedly implemented based on two differently composed reference groups, whereas only the overlap of both calling results was considered to be informative enough to reflect appropriate CNV frequencies. However, it is imaginable that this rudimentary CNV calling step may be completely skipped in future, when the quality and information content of public databases has become increasingly better and when these catalogs truly reflect the human CNV map stratified by racial and ethnic origin as well as by the individual state of health.

6 Strategy PS2: Estimation of CNVs by Use of Sophisticated Reference Models

The second proposed CNV analysis strategy, which is closely related to the most widespread genome-wide CNV analysis strategy S2, will be introduced and examined in detail in this chapter. After having motivated the importance of reference models in array-based CNV detection, one particular approach that fits probe-wise intensity signals to a Gaussian mixture model will be suggested. This aspect of estimating probe-specific copy number neutral hybridization intensity values on array-derived CNV data prior to the application of standard CNV detection pipelines forms the major modification of strategy PS2 relative to strategy S2. Consequently, the impact of alternative reference values compared to commonly used median reference intensity values is investigated in detail by examining stability, reproducibility and reliability of CNV calls in publicly available Affymetrix 6.0 data of the HapMap sample (http://hapmap.ncbi.nlm.nih.gov/) as well as on replicate data being previously analysed in several CNV mapping studies (e.g. Tuzun et al. (2005); Redon et al. (2006)). Finally, application of the whole genome-wide CNV analysis strategy PS2 is examplarily presented for the phenotype obesity. It will be demonstrated that applying strategy PS2 instead of strategy S2 will dramatically change genetic association results.

6.1 Strategy PS2

The outline of our proposed CNV analysis strategy PS2 is diagrammed in Figure 6.1. Compared to the standard CNV analysis strategy S2, which is described in detail in chapters 3.3.1 ff., the use of a set of sophisticated global reference values for CNV detection instead of the most widely used probe-wise median reference intensities forms the major extension in strategy PS2. In more detail, we propose to separately fit a finite Gaussian mixture model to the samples' intensity data of each

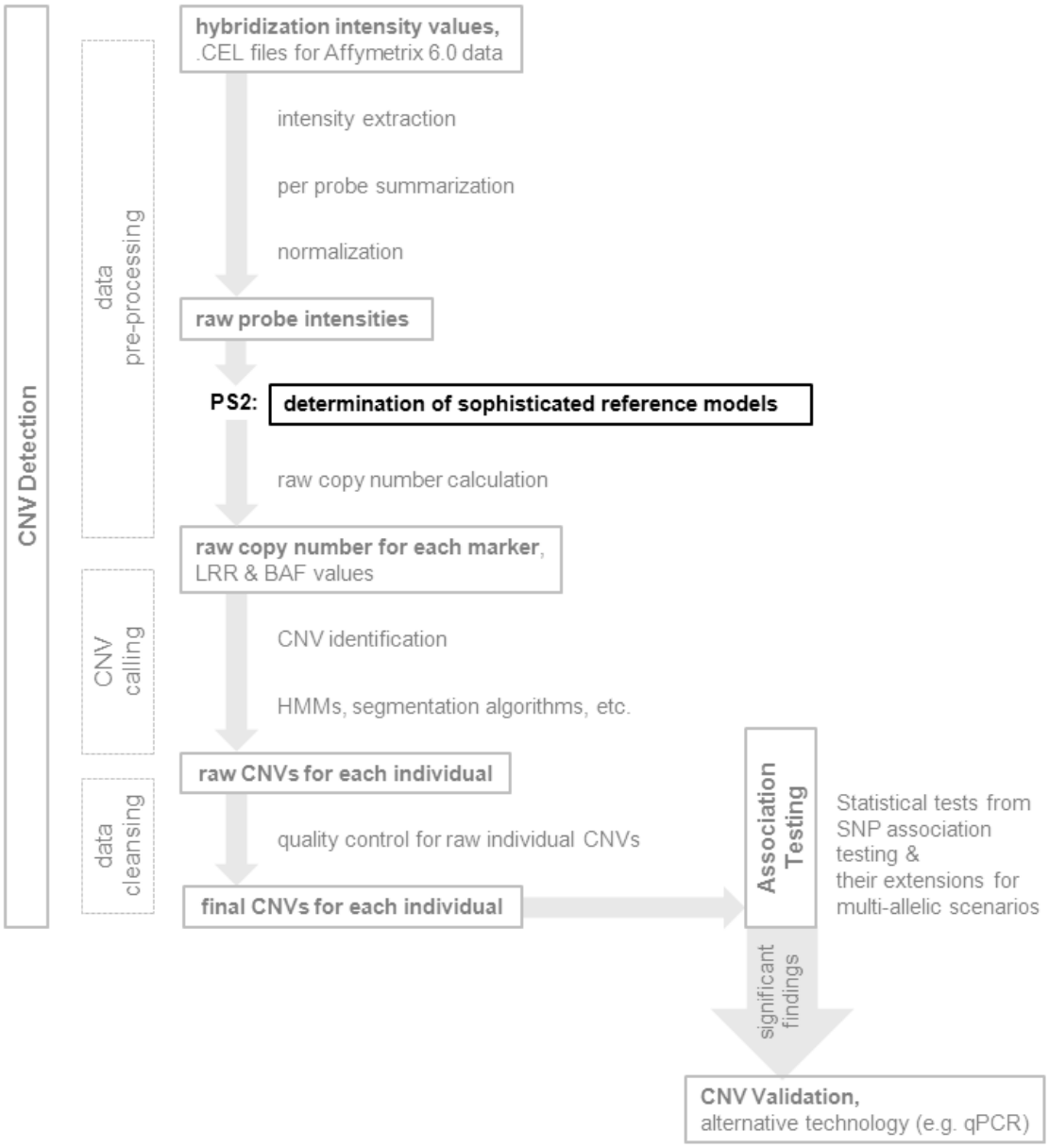

Figure 6.1: Schematic representation of the proposed CNV analyses strategy PS2.

available array probe set, and to subsequently chose the mean of each probe set's copy number neutral cluster for the application as the probe-wise reference intensity value in the following $\log_2$ intensity ratio calculation.

Several authors presented CNV genotype calling algorithms (CNVtools: Barnes et al. (2008), CNVassoc: Subirana et al. (2011), CNVmix: Marioni et al. (2007), etc.) that are based on modelling probe-wise sample-wide $\log_2$ ratios using a finite Gaussian mixture model. However, these models suffer from the risk of misclassified cluster locations due to imprecise reference intensity estimation. Consequently, we

propose to improve the precision in the estimation of the global reference valuess prior to the application of one of the various available CNV calling software tools.

6.1.1 Motivation for a Sophisticated Reference Model

Estimating raw copy number states from hybridization intensities

The general approach for the estimation of copy number states is derived from the definition of a CNV as "a segment of DNA that is one kb or larger and is present at a variable copy number in comparison with a reference genome" (Levy et al. (2007), cf. chapter 2). Thus, the individual copy number state C_{ind} of a certain DNA segment is estimated relative to the same DNA segment's copy number C_{ref} of a single reference sample

$$C_{ind} = \tau \; C_{ref},$$

where τ denotes the relative difference between the two integer-valued copy number states. Usually, the reference sample is selected to be of copy number neutral, so called normal, state and thus harbours an absolute number of two copies of the respective DNA segment, one on each of the two homologous chromosomes. With this assumption of $C_{ref} = 2$, values of $\tau = \frac{C_{ind}}{C_{ref}}$ below or above one can intuitively be interpreted as copy number losses or gains, respectively.

Fluorescence intensities from the hybridization step of microarray experiments are assumed to be linearly indicative for the amount of DNA transcripts labeled to the probes on the array. Consequently, for an array probe set that represents the respective DNA segment, τ can equivalently be expressed in terms of hybridization intensities, X_{ind} and X_{ref}, of the test and the reference sample

$$C_{ind} = 2 \; \frac{X_{ind}}{X_{ref}}.$$

Although the presented intensity ratios $\frac{X_{ind}}{X_{ref}}$ provide an intuitive measure of copy number changes, they have the disadvantage of treating losses and gains differently. Duplications with a relatively doubled absolute copy number of four have an intensity ratio of two, whereas hemizygous deletions with half of the copies from the reference sample have an intensity ratio of $\frac{1}{2}$. To induce symmetry and to facilitate interpretation, it is convenient to apply $\log_2$ transformations on the microarray intensity data

$$\log_2 \left(\frac{C_{ind}}{2} \right) = \log_2 \left(\frac{X_{ind}}{X_{ref}} \right). \tag{6.1}$$

With this most widely used transformation, the $\log_2$ intensity ratio on the right side of equation (6.1) is symmetric around zero, which means that reciprocal copy number changes have symmetric $\log_2$ intensity ratios: $\log_2(2) = 1$ and $\log_2\left(\frac{1}{2}\right) = -1$ and so on. Additional advantages of the $\log_2$ transformation are that linearity, additivity and normality, which is of special interest for statistical analyses, are achieved. Biological plausibility of the identity (6.1) has empirically been shown in X-chromosome dosage response experiments including several replicates of samples with one to five copies of the X chromosome (Huang et al., 2004).

Estimating copy number states for Affymetrix 6.0 data

Similar to most genotyping arrays, the Affymetrix 6.0 technology is characterized by the use of several short oligonucleotide (25-mers) probe sets to characterize the structure of genomic DNA regions. That is, CNVs that typically make up at least one kb of DNA cannot be represented by hybridization intensities of only one single probe set. Instead, a set of intensity signals from several adjacent SNP and CN probe sets covering the CNV is needed for individual copy number estimation. Exemplarily, probe-wise intensity ratios, $\log_2$ intensity ratios and corresponding raw copy number values are depicted in Figure 6.2 for a hemizygous deletion and a duplication on chromosome 6 of a hypothetical individual. Several algorithms have been proposed to detect individual CNVs on the basis of individual raw copy number values. All these software tools rely on the fundamental characteristic that a CNV is a relatively long DNA sequence with a constant number of copies in each individual human being (cf. chapter 3.2).

As already stated in chapter 3.1.1, the SNP array technology, such as the Affymetrix 6.0 genotyping arrays, does not use a specific control or reference sample. Alternatively, reference values need to be artificially provided for each probe set separately on the basis of intensity information for a group of collectively processed individuals. The most widely used assumption for the determination of global reference intensity values is that the majority of randomly selected individuals is free of specific CNVs. Consequently, in most currently available software tools for CNV discovery, such as Birdsuite, CNAT, CNVPartition 1.2.1, dChip SNP, GADA, ITALICS, PennCNV or QuantiSNP (cf. chapter 3.2), probe-wise reference intensities are defined to equal the sample mean or a robust sample average. For instance, the sample median or the trimmed sample mean are robust sample averages in the sense that they are only marginally influenced by outliers.

The basic assumption for the statistical estimation of individual copy number states from intensities of high density oligonucleotide SNP arrays is that intensity

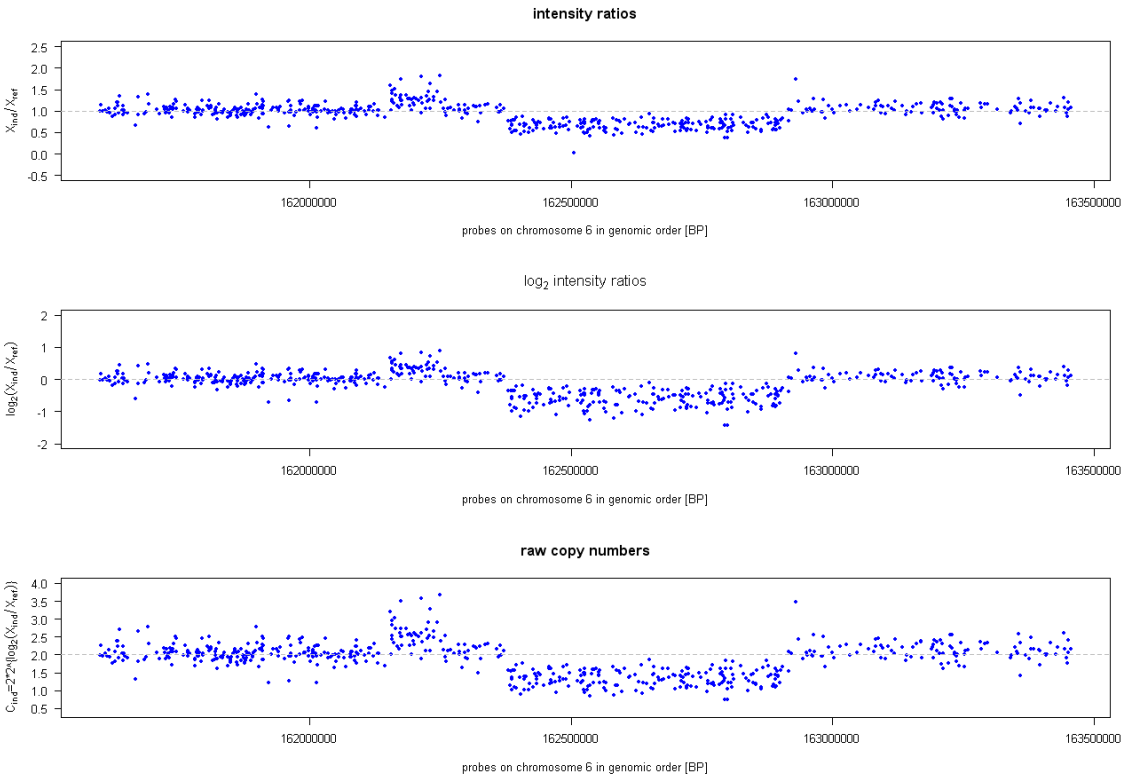

Figure 6.2: Exemplary probe-wise intensity ratios, $\log_2$ intensity ratios and raw copy numbers of an individual (hypothetical) are presented for each Affymetrix 6.0 probe set covering the depicted region on chromosome 6. Intensity ratios, $\frac{X_{ind}}{X_{ref}}$, and $\log_2$ intensity ratios, $\log_2\left(\frac{X_{ind}}{X_{ref}}\right)$, are displayed in the middle and upper part of the panel. Transformed individual raw copy numbers $C_{ind} = 2 \cdot 2^{\log_2\left(\frac{X_{ind}}{X_{ref}}\right)}$, referred to the upper $\log_2$ intensity ratios, are shown in the lower part of the panel.

levels across individuals are comparable to each other. On the one hand, signal intensities can be influences by experimental noise, which can be modeled as an additive zero-mean Gaussian error term component. On the other hand, hybridization intensity levels can be biased between samples due to low-quality measurements and plate- or batch-effects, which may arise when laboratory conditions have changed in the meantime of processing different plates or batches of plates. To separate difference of biological origin from those of non-biological origin and in order to allow meaningful biological comparisons, a transformation at the sample level, referred to as normalization, is typically applied to SNP array data before any statistical estimation is done (Quackenbush, 2002). Several methods were proposed to give

each array the same distribution, among which quantile normalization has proven to perform favorably (Bolstad et al., 2003).

Median reference intensity values

As already noted by several authors, such as Komura et al. (2006), Bengtsson et al. (2008b), Pique-Regi et al. (2009) or Zhang et al. (2011), the application of median values as reference intensities is appropriate for most CNVs but can be problematic under specific conditions. Median reference values may especially be inappropriate for probe sets in genomic regions that contain common CNVs. In genomic regions with common CNVs, the sample's major CNV state does no longer equal the assumed normal CNV state of two copies. By definition, a biased estimation of reference hybridization intensity values potentially implicates a misclassification of raw continuous copy numbers at the single locus level. Slightly erroneous continuous raw copy numbers may nonetheless result in a correct classification into discrete raw copy numbers at the single locus level. However, accumulated misclassified raw copy numbers for adjacent single probe sets may on the one hand result in completely missing to detect specific CNVs. On the other hand, especially complex CNVs may be wrongly identified in the course of the CNV discovery step, when one of the currently available CNV genotype calling tools is directly applied to the raw copy number measurements.

The above described concern about median reference values is illustrated in Figure 6.3, which displays the Affymetrix 6.0 median based $\log_2$ intensity ratio profile of replicate 4 on sample NA15510 (see section 6.2.1 for details), who was shown to harbour a small deletion on chromosome 1 (Korbel et al., 2007). Indeed, the depicted chromosomal region was later on shown to harbour two deletion alleles, one ten kb deletion with low population frequency at chr $1 : 72\,528 - 72\,536$ kb and another 45 kb deletion with considerably greater population frequency at chr $1 : 72\,540 - 72\,585$ kb (Willer et al., 2009; McCarroll et al., 2008; McCarroll, 2010). Of note, exclusively the occurrence of deletions and hence the absence of duplications was validated for the 45 kb deletion region on chromosome 1 by use of independent validation technologies, such as sequencing methods (McCarroll, 2010).

In the upper part of Figure 6.3, the intensity histograms of two Affymetrix 6.0 CN probe sets, CN_517821 and CN_517842, are shown. Each probe set is selected to exemplarily represent one of the two above described CNVs and thus reflects the true underlying CNVs' frequency in a sample comprising approximately 300 individuals of European origin (270 HapMap samples and 25 replicate data sets, for details see section 6.2.1). As depicted in the lower part of Figure 6.3, median reference values

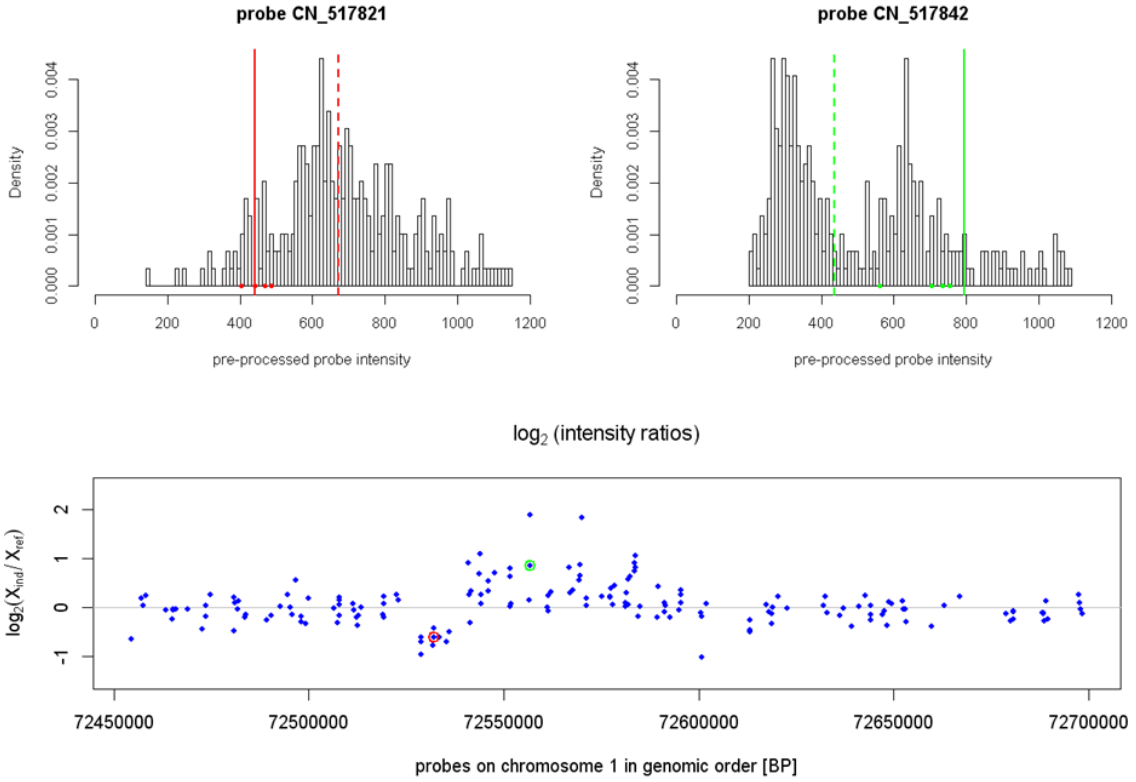

Figure 6.3: Median based $\log_2$ intensity ratios for replicate 4 of sample NA15510, carrying a small deletion at chr $1 : 72\,528\,701 - 72\,535\,958$, are presented in the lower part of the panel. Histograms of pre-processed hybridization intensity values of a sample comprising $\sim$ 300 individuals are depicted in the upper part of the panel for two selected probe sets. In both histograms, sample intensity medians are represented by a dashed colored line, the intensity levels of replicate 4 on sample NA15510 are presented by a solid colored line and intensities of replicates 1,2,3,5 of sample NA15510 are shown as colored points. Corresponding $\log_2$ intensity ratios of NA15510's replicate 4 for the two probe sets CN_517821 and CN_517842 are highlighted in red and green in the lower part of the panel, respectively.

are useful to identify the small infrequent deletion of sample NA15510. Contrarily, the application of median reference intensities would falsely induce to conclude that the same individual additionally carries a larger duplication upstream of the small deletion. As represented in the intensity histogram for probe set CN_517842, the sample median intensity is clearly shifted towards an intensity level below those intensity value, which represents an individual exhibiting two copies of the respective DNA segment. Thus, since almost half of the presented sample harbours a deletion at probe set CN_517842, and even though the intensity of sample NA15510 can

visually be unambiguously assigned to those intensity cluster representing two DNA copies, the $\log_2$ intensity ratio value of 0.863 hints to a duplication. This wrong conclusion about the presence of a duplication at chr $1 : 72\,543\,731 - 72\,583\,736$ in the sample NA15510 is strengthened by several adjacent probe sets with sample intensity profiles that are similar to this of CN_517842.

Alternative reference intensity values

Oldridge et al. (2010) presented three correction methods to account for the inaccuracy of median reference models for common CNVs. The first method is based on the observation that misclassified reference values lead to an unusual high number of CNV carriers in the analysed sample. Thus, CNVs with predicted sample frequency above a certain threshold (e.g. $\geq 66\%$) were supposed to be discarded. In the second approach, a trimmed sample intensity median, accounting for only mid-valued intensities, is proposed for the use as reference. Thirdly, one and the same prespecified single reference individual was suggested to be used for $\log_2$ intensity ratio calculation at each single probe set. After the determination of $\log_2$ intensity ratios by any of the three correction methods, three commonly used segmentation algorithms (GLAD, DNAcopy, APT) were applied to call CNV genotypes. With regard to false positive and false negative rates, being with respect to CNVs determined by use of the aCGH technology, the third correction method clearly outperformed the two other presented methods.

Taking into account the above described results of Oldridge et al. (2010), we suggest the application of a sophisticated reference model for raw copy number calculation prior to CNV genotype calling. Empirically, it was demonstrated by Oldridge et al. (2010) that a biologically plausible global reference system, such as reference intensities that are genome-wide taken from one and the same single sample, perform favorably in comparison to standard median-based reference intensity determination approaches. The single reference sample design however still implicates erroneous CNV genotype calling at CNV regions in which the reference individual itself harbours a deletion or a duplication. For example, a deletion in the reference individual would imply overestimated sample-wide $\log_2$ ratios and would thereby result in the detection of duplications in biologically copy number neutral individuals. Optimally, a global reference model would be piece-wise composed of intensity values from several single samples that are all free of any CNV in the respective chromosomal region. That is, intensity values of probe sets in chromosomal order are taken from one single reference sample as long as this individual is not deleted or duplicated at the respective probe set, and otherwise intensities of another

reference individual that is free of any CNV at the respective DNA region are used. However, since CNV states for none of the analysed samples are a priori known, this procedure for the estimation of an optimal reference model is impossibly applicable to Affymetrix 6.0 data. Instead, we propose to predict the theoretically optimal reference model by estimating the expected intensity values for the presence of 2 DNA copies by separately fitting a finite Gaussian mixture model at each available probe set.

6.1.2 Sophisticated Reference Models for CNV genotype Calling

A Gaussian mixture model for sample hybridization intensities

Assume to have pre-processed hybridization intensity data from SNP genotyping microarrays, such as the Affymetrix 6.0 or the Illumina 1M platform. For a total of n genotyped individuals, let intensity data for p probe sets, spanning the whole genome, be available from chip experiments. For simplicity, all probes are assumed to be non-polymorphic probes, so called CN probes (see section 3.1.2 for details). Hybridization intensity signals for the A and the B allele of SNP probe sets may be additively summarized into one single intensity measure reflecting the total amount of DNA labeled to the respective probe set (Peiffer et al., 2006). With this procedure CN and SNP probe sets do not necessarily need to be distinguished throughout statistical analyses. However, for SNP probe sets it might be appropriate to calculate the reference intensities conditional on the predicted individual SNP genotypes (either AA, AB or BB) (Peiffer et al., 2006).

For each probe set, $k = 1, \ldots, p$, the vector of observed individual pre-processed hybridization intensity signals $\mathbf{x}_{[k]} = (x_{1k}, \ldots, x_{nk})$ is assumed to be a realization of a vector of independent and identically distributed (i.i.d.) random variables

$$\mathbf{X}_{[k]} = (X_{1k}, \ldots, X_{nk}) \quad \text{for} \quad k = 1, \ldots, p.$$

Each random variable X_{ik} is assumed to follow a mixture of several Gaussian distributions, which are called components. More precisely, for each $k = 1, \ldots, p$ and $i = 1, \ldots, n$ the random variable X_{ik} is assumed to have the following mixture probability density function (p.d.f.)

$$f_{\boldsymbol{\theta}_k}(x_{ik}) = \sum_{j=1}^{c} \lambda_{kj} \; \phi_{kj}(x_{ik}),$$

with parameters $\boldsymbol{\theta}_k = (\boldsymbol{\lambda}_k, \boldsymbol{\phi}_k) = (\lambda_{k1}, \ldots, \lambda_{kc}, \phi_{k1}, \ldots, \phi_{kc})$. In particular, the mixing proportion parameters are non-negative, $\lambda_{kj} \geq 0$, and sum to unity, $\sum_{j=1}^{c} \lambda_{kj} = 1$ for each $k = 1, \ldots, p$. The total number of components c equals the assumed underlying number of CNV states with biological plausibility, whereas in general $c = 5 = |\{0, 1, 2, 3, 4\}|$. Moreover, the functions ϕ_{kj} are assumed to be drawn from the family of univariate Gaussian densities, that is each ϕ_{kj} is the p.d.f. of some normal distribution $\mathcal{N}(\mu_{kj}, \sigma_{kj}^2)$ with parameters $\mu_{kj}, \sigma_{kj}^2 \in \mathbb{R}$,

$$\phi_{kj}(x_{ik}) = \frac{1}{\sigma_{kj}\sqrt{2\pi}} \, e^{-\frac{1}{2}\left(\frac{x_{ik}-\mu_{kj}}{\sigma_{kj}}\right)}.$$

Thus, model parameters $\boldsymbol{\theta}_k$ reduce to $\boldsymbol{\theta}_k = (\boldsymbol{\lambda}_k, (\mu_{k1}, \sigma_{k1}^2), \ldots, (\mu_{kc}, \sigma_{kc}^2))$.

As an example of probe-wise pre-processed hybridization intensity data from a mixture model, the sample distribution of intensities for probe set CN_517842 is

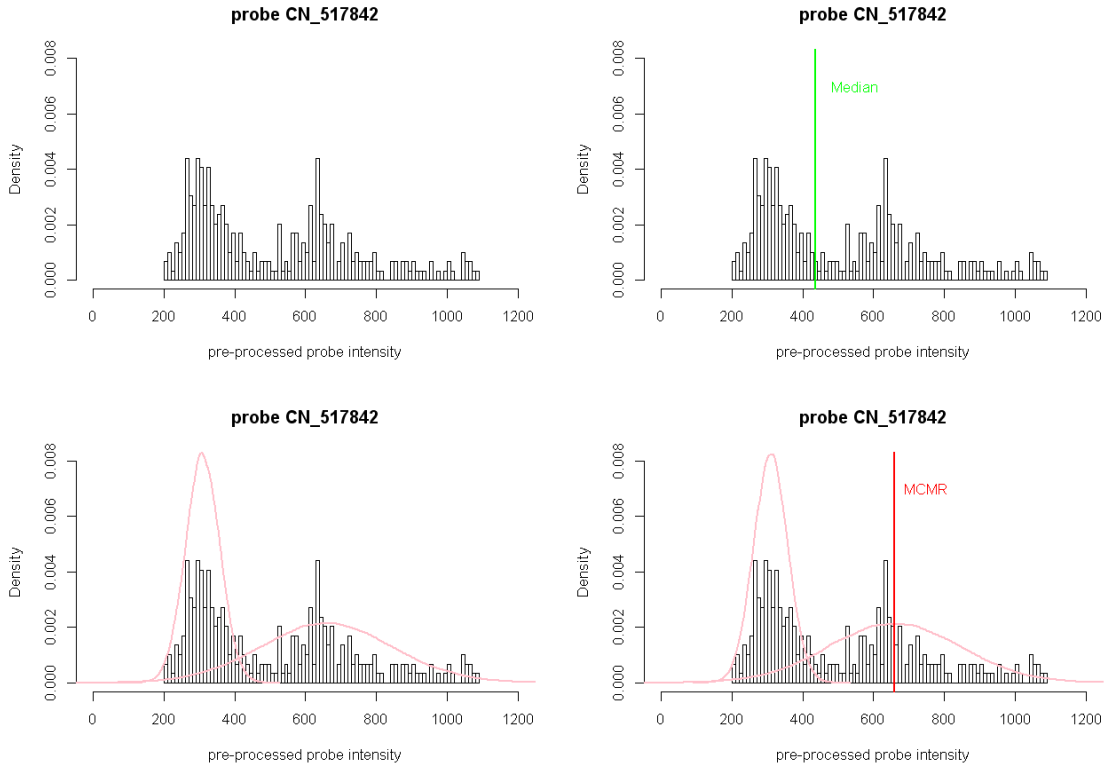

Figure 6.4: Intensity data of probe set CN_517842 for a sample comprising approximately 300 individuals with highlighted median and MCMR reference intensity values. In the lower row, densities for the two estimated mixture components are included in pink. Additionally, median and MCMR reference intensity values are highlighted in green and red in the right part of the panel, respectively.

depicted in Figure 6.4. For the presented intensities of a sample comprising approximately 300 individuals (270 HapMap samples and 25 replicate data sets, for details see section 6.2.1), a two-component mixture model is a reasonable with regard to the bi-modality of the histogram. The presented model component densities were estimated by use of the R-package 'mclust' (Fraley et al., 2012).

MCMR reference intensity values

The choice of a sophisticated reference intensity value, called MCMR, is illustrated for probe set CN_517842 in Figure 6.4. From a biological point of view, the most plausible intensity reference value might be the expected value of those component of the estimated Gaussian mixture model, which represents individuals with two copies of the respective DNA segment. In general, the component that each individual comes from is unobserved and for each individual it is modeled as a vector of Bernoulli random variables $\mathbf{D}_{ik} = (D_{ik1}, \ldots, D_{ikc})$ with $D_{ikj} \in \{0, 1\}$. Since for each $k = 1, \ldots, p$, each single observation X_{ik} comes from exactly one component, the random variables D_{ikj} sum to unity for each $i = 1, \ldots, n$: $\sum_{j=1}^{c} D_{ikj} = 1$. Additionally,

$$\mathbb{P}(D_{ikj} = 1) = \lambda_{kj} \quad \text{and} \quad (X_{ik}|D_{ikj}) \sim \phi_{kj},$$

for $j = 1, \ldots, c$ and $k = 1, \ldots, p$. Theoretically, the samples mean most probably follows the copy number neutral component. Consequently, sophisticated reference values are proposed to equal the mean of those component that the samples mean, $m_k = \bar{X}_{[k]} = \frac{1}{n} \sum_{i=1}^{n} X_{ik}$, is most probably underlying. This reference value will be called Mean Component Mean Reference, it is abbreviated as MCMR, and for each probe set $k = 1, \ldots, p$ it is formally defined as

$$\text{MCMR}_k = \mu_{kz}$$

with

$$\begin{aligned} z &= \underset{j \in \{1,\ldots,c\}}{\text{argmax}}\ \mathbb{P}(D_{ikj} = 1|m_k) \\ &= \underset{j \in \{1,\ldots,c\}}{\text{argmax}}\ \frac{\mathbb{P}(D_{ikj} = 1)\ \phi_{kj}(m_k)}{\sum_{\kappa=1}^{c} \mathbb{P}(D_{ik\kappa} = 1)\ \phi_{k\kappa}(m_k)} \\ &= \underset{j \in \{1,\ldots,c\}}{\text{argmax}}\ \frac{\lambda_{kj}\ \phi_{kj}(m_k)}{\sum_{\kappa=1}^{c} \lambda_{k\kappa}\ \phi_{k\kappa}(m_k)}. \end{aligned}$$

6.2 Comparison of Partially Applied Strategy S2 and strategy PS2 to HapMap and Replicate Data

In the following, the applicability and performance of alternative MCMR reference values will be presented by means of examples containing real data of alive individuals. CNVs will be analysed in two different data sets. The main focus will be on strengths and weaknesses of MCMR reference intensity values for CNV detection compared to median (Default) reference intensity values.

By using real data examples, there is no need to artificially simulate raw data of CNVs accounting for different models of CNV probe occurrences. In contrast to genetic phenomena that affect only one specific chromosomal base pair position (e.g. SNPs), a CNV requires a much more complicated simulation setting due to the variety of involved microarray probe sets. In such a simulation, apart from characteristics of the particular underlying genotyping technology, the number of CNV affected probe sets and their probe-wise intensity level should be considered in dependence of their distance and correlation structure as well as the CNV's type, chromosomal location and its population frequency. Real data examples provide the possibility of comparing different reference models for CNV detection without potential bias due to misclassified simulation models.

Publicly available Affymetrix 6.0 microarray data was analysed. On the one hand, this public data set comprises the so-called HapMap individuals and on the other hand it includes replicate microarrays for several individuals. The HapMap data set is of special interest as the very same individuals were precisely analysed in a variety of previous studies by application of several genotyping technologies. The latter fact offers the possibility of calculating estimates for false positive and false negative rates without the necessity of additionally collecting validation data. Moreover, CNV calling results of replicate experiment's data can be compared across individuals with regard to stability and reproducibility rates.

6.2.1 Data Sets

CNVs were called genome-wide in a publicly available data set (www.affymetrix.com/support/technical/sample_data/genomewide_snp6_data.affx) of Affymetrix 6.0 data for 270 samples from the International HapMap Project (www.hapmap.org) and for an additional collection of five replicates on five single samples. Thus, a total of 295 Affymetrix 6.0 microarray data sets were analysed.

The 270 HapMap samples are comprised of 30 Utah residents trios with ancestry from northern and western Europe (abbreviation: CEPH), 30 Yoruban trios from

Ibadan in Nigeria (abbreviation: YRI), 45 unrelated Han Chinese samples from Beijing in China (abbreviation: CHB) and 45 unrelated Japanese samples from Tokyo in Japan (abbreviation: JPT). CNVs of the full set of HapMap samples or of a subset of all 270 HapMap samples, respectively, have been extensively analysed by several previous studies, such as Conrad et al. (2006), Redon et al. (2006), Kidd et al. (2008), Perry et al. (2008), McCarroll et al. (2008), Shaikh et al. (2009), Conrad et al. (2010), Park et al. (2010).

Of special interest for the exploration of CNVs are the five sets of five Affymetrix 6.0 microarrays, each being processed for the same out of five individuals (NA10851, NA15510, NA04626, NA01416, NA06061). Each of the five single samples has a different number of copies of the X chromosome, varying from one to five. Sample NA10851 is a normal male with one copy of the X chromosome, sample NA15510 is a normal female with two copies and the other three samples have abnormal numbers of X chromosome copies of three, four, and five, respectively. The female sample has been extensively studied by fosmid paired end sequencing by Tuzun et al. (2005). Additionally, Korbel et al. (2007) applied paired-end mapping to map structural variations in the genomes of the male and the female sample. Moreover, Redon et al. (2006) used the male and the female sample to train threshold parameters for the analysis of the HapMap samples based on SNP genotyping arrays. Finally, Kidd et al. (2008) mapped and sequenced structural variation from eight human genomes with regard to NA15510 as a reference sample.

6.2.2 Methods

Reference models

The CNV detection performance of the proposed global MCMR reference model was compared to the most widespread median reference model. In the course of the data pre-processing step, the set of all 295 Affymetrix 6.0 microarray data sets of 295 '.CEL' files was quantile normalized and median polished by use of the Affymetrix Power Tools (APT) standard protocols. Subsequently, probe-wise median reference values were determined by use of the PennCNV software (Wang et al., 2007). For the calculation of global MCMR reference values, a Gaussian mixture model was first fitted to the probe-wise pre-processed sample-wide intensities by use of the R-package 'mclust' (Fraley et al., 2012). In use of the function 'mclust()', the parameters of the Gaussian mixture model are estimated via the EM algorithm and the optimal model is selected according to the Bayesian information criterion (BIC). The corresponding R-function that was used to specify probe-wise MCMR reference intensity values is given in the Appendix.

CNV detection

Detection of CNVs was performed in application of the PennCNV software (Wang et al., 2007) by use of both, median and MCMR reference intensity values, respectively. With regard to false positive, false negative and Mendelian inconsistency rates of CNV calling, called CNVs were subsequently compared to published CNVs of Tuzun et al. (2005), Korbel et al. (2007), Redon et al. (2006), Kidd et al. (2008), Conrad et al. (2006), Perry et al. (2008), McCarroll et al. (2008), Shaikh et al. (2009), Conrad et al. (2010) and Park et al. (2010), respectively. Additionally, CNVs of those 5 individuals with available replicate data sets were compared across replicates with respect to stability and reproducibility rates of CNV calls.

6.2.3 Results

6.2.3.1 Stability Rates of CNVs

To evaluate the CNV detection performance of median (Default) and MCMR reference values, the concordance of PennCNV's CNV calls was investigated across replicate sets of Affymetrix 6.0 data, each consisting of five chip experiments accounting for the same individual. Such replicate data were available for a total of five individuals. Thus, genome-wide CNV calls for 25 microarrays were compared at the segment level. Respective results are summarized in Figure 6.6 and Table 6.1.

Comparisons of CNV call's stability were performed pairwise at the segment level. In more detail, a total of 5×10 pairs of sets made-up of individual CNV calls were checked against each other for chromosomal segments with concordant CNV calling results. As shown in Figure 6.5, overlapping CNV calls across replicates of identical type, i.e. duplications or deletions, were summed up into one so-called CNV segment. In these considerations, only duplications and deletions were discriminated, whereas copy number differences within duplications or deletions were not taken into account.

Figure 6.5: Schematic representation of a CNV segment defined by three overlapping CNV calls from two replicate data sets for the same individual.

Copy number segments whose reciprocal overlap was above a certain threshold were regarded as concordant segments. For the calculation of individual pairwise stability rates, overlap thresholds of $> 0\%$, $> 50\%$, $> 80\%$ and 100% were applied.

Individual pairwise stability rates were defined to equal the pairwise concordance rates of CNV segments being detected in use of median (Default) and MCMR reference intensity values. Results for overlap threshold of $> 50\%$ and $> 80\%$ were similar. Consequently, in Table 6.1 mean and median values of individual stability rates are only presented for overlap thresholds of $> 0\%$, $> 50\%$ and 100%.

Moreover, summarized stability rates are given for different CNV detection thresholds, that is stratified by the minimal number of involved array probe sets (Figure 6.6, Table 6.1). For both, Default and MCMR reference intensity values, highest stability rates were observed for CNV calls with more than 20 probe sets and with regard to any pairwise overlap between replicate calls.

Individual CNV's stability rates were statistically significantly higher with MCMR reference intensity values compared to Default reference values, when incorporating CNV calls with $> 3, > 5$ or > 10 probe sets and accounting for any pairwise overlap between replicates. Additionally, throughout almost any overlap threshold as well as for almost any CNV detection threshold, stability rates were at least slightly higher with MCMR than with Default intensity reference values.

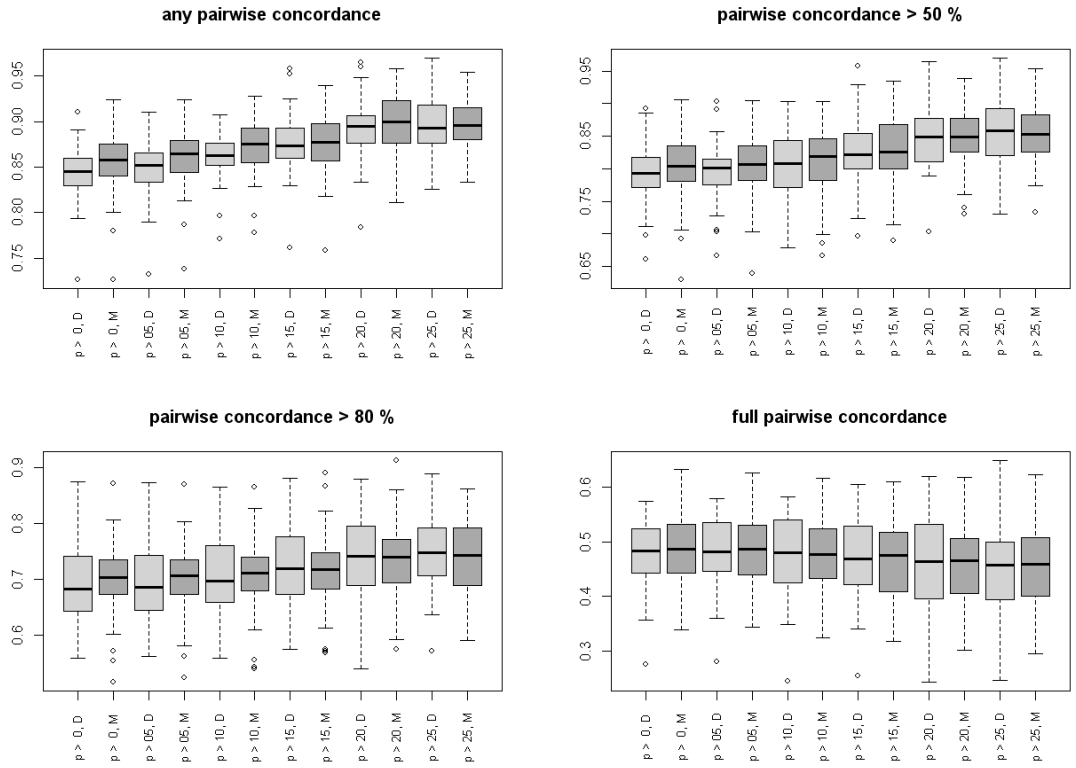

Figure 6.6: Stability rate (= pairwise concordance rates) of CNV calls using replicate's data sets on same individuals (n = 5). p: probe sets per CNV call, D: Default calling, M: MCMR calling

Table 6.1: Stability rates (= pairwise concordance rates) of CNV calls using five replicate's data sets on same individuals (n = 5).

claimed # of probe sets per CNV call	median number of CNV calls		minimal pairwise concordance (overlap)	median pairwise concordance rates		mean pairwise concordance rates		sd* of pairwise concordance rates		p-value, two-sided paired t-test
	Default Calling	MCMR Calling		Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	
			any	84.50%	85.83%	84.28%	85.55%	3.00%	3.40%	4.17×10^{-4}
> 3	62	61	$> 50\%$	79.30%	80.35%	79.11%	80.19%	4.20%	5.04%	0.0145
			100%	48.24%	48.65%	47.81%	48.21%	6.31%	6.74%	0.5200
			any	85.15%	86.48%	84.91%	86.00%	3.16%	3.18%	0.0041
> 5	61	60	$> 50\%$	80.15%	80.70%	79.73%	80.59%	4.33%	4.83%	0.0564
			100%	48.12%	48.61%	47.88%	48.25%	6.17%	6.53%	0.5498
			any	86.27%	87.50%	86.28%	87.26%	2.68%	3.19%	0.0100
> 10	53	53	$> 50\%$	80.77%	81.84%	80.72%	81.42%	4.67%	5.15%	0.1467
			100%	48.04%	47.67%	47.53%	47.23%	7.17%	7.04%	0.6759
			any	87.36%	87.70%	87.68%	87.79%	3.10%	3.52%	0.7428
> 15	47	46	$> 50\%$	82.14%	82.61%	82.75%	82.93%	4.75%	5.34%	0.7471
			100%	46.79%	47.46%	47.06%	46.52%	7.06%	7.38%	0.4821
			any	89.51%	89.97%	89.44%	89.84%	3.08%	3.11%	0.2350
> 20	40	38	$> 50\%$	84.93%	84.93%	84.41%	84.79%	4.81%	4.61%	0.4652
			100%	46.33%	46.48%	46.23%	45.82%	8.43%	7.50%	0.5873
			any	89.26%	89.57%	89.55%	89.51%	3.24%	3.01%	0.9116
> 25	33	32	$> 50\%$	85.90%	85.29%	85.69%	85.37%	4.81%	4.67%	0.5975
			100%	45.66%	45.90%	45.70%	45.03%	7.73%	7.70%	0.4145

* sd = standard deviation.

6.2.3.2 Reproducibility Rates of CNVs

CNV calls of the above described replicate data for five individuals have also been investigated with regard to their reproducibility across five replicate experiments. Individual reproducibility rates were determined for each genomic interval with identical replicate-wise CNV configuration (Figure 6.8, Table 6.2).

For this purpose, overlapping CNVs of each individual's five replicates were at first merged into one CNV region (CNVR). The breakpoints of the maximum interval covered by any overlapping CNV were chosen as the CNVR's boundaries. Afterwards, individual complex CNVRs including replicate CNVs with discordant estimated boundaries but overlapping segments, were subdivided into several sub-CNVRs as depicted in Figure 6.7. Thus, each sub-CNVR was defined to contain only one specific replicate CNV and not to harbour two different copy number states per replicate. Consequently, the sub-CNVR's boundaries were exactly given by the set of CNV's breakpoints. Briefly, a CNVR represents a union of overlapping CNVs and the sub-CNVRs precisely describe the exact structure of the CNVR.

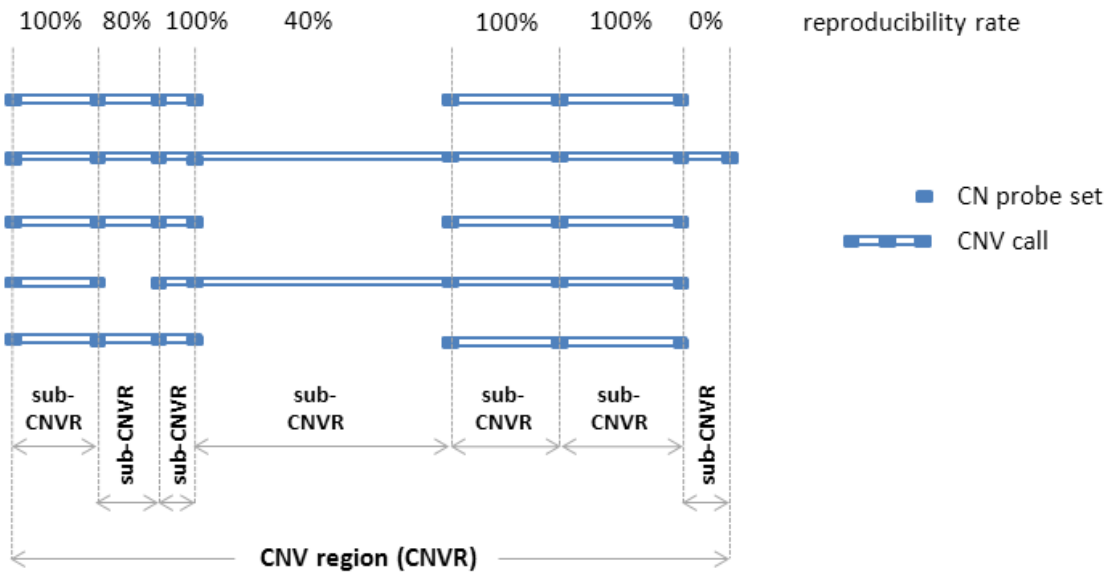

Figure 6.7: Schematic representation of a CNV region (CNVR) containing six sub-CNVRs, being defined by overlapping CNV calls of five replicates for the same individual, with reproducibility rates ranging from 0 to 100%.

With regard to the number of available replicate experiment data sets per individual (n = 5), reproducibility rates were calculated with respect to 2/5 %, 3/5 %, 4/5 %, and 100 % concordance across CNV calls at the sub-CNVR level. In more detail, the x% reproducibility rate was defined to equal the proportion of individual sub-CNVRs with consistent CNV calling results in at least x% of the replicates.

Neither the stability rate nor the reproducibility rate directly predict the CNV calling performance of Default or MCMR reference intensity values; however, they both indicate the stability and reproducibility of CNV detection performance. Table 6.2 shows that CNV calls based on MCMR reference intensity values have higher average estimated reproducibility compared with Default reference intensity values. The difference in reproducibility rates decreases when the minimal number of informative array probe sets per CNV call increases, that is with an increased CNV detection threshold. Reproducibility rates are statistically significantly higher in application of MCMR reference intensity values in comparison to Default reference intensity values, when CNV calls where claimed to harbour more than three, five or ten probe sets and when then a complete reproducibility of 100% is considered.

Similarly to the stability rates, reproducibility rates were at the highest level for a CNV detection threshold of 20 probe sets per CNV call, and this was consistently observed across each reproducibility value as well as for both, MCMR and Default reference intensity values. Thus, this threshold for CNV detection of > 20 probe sets per CNV call will be applied throughout the following considerations if not stated otherwise.

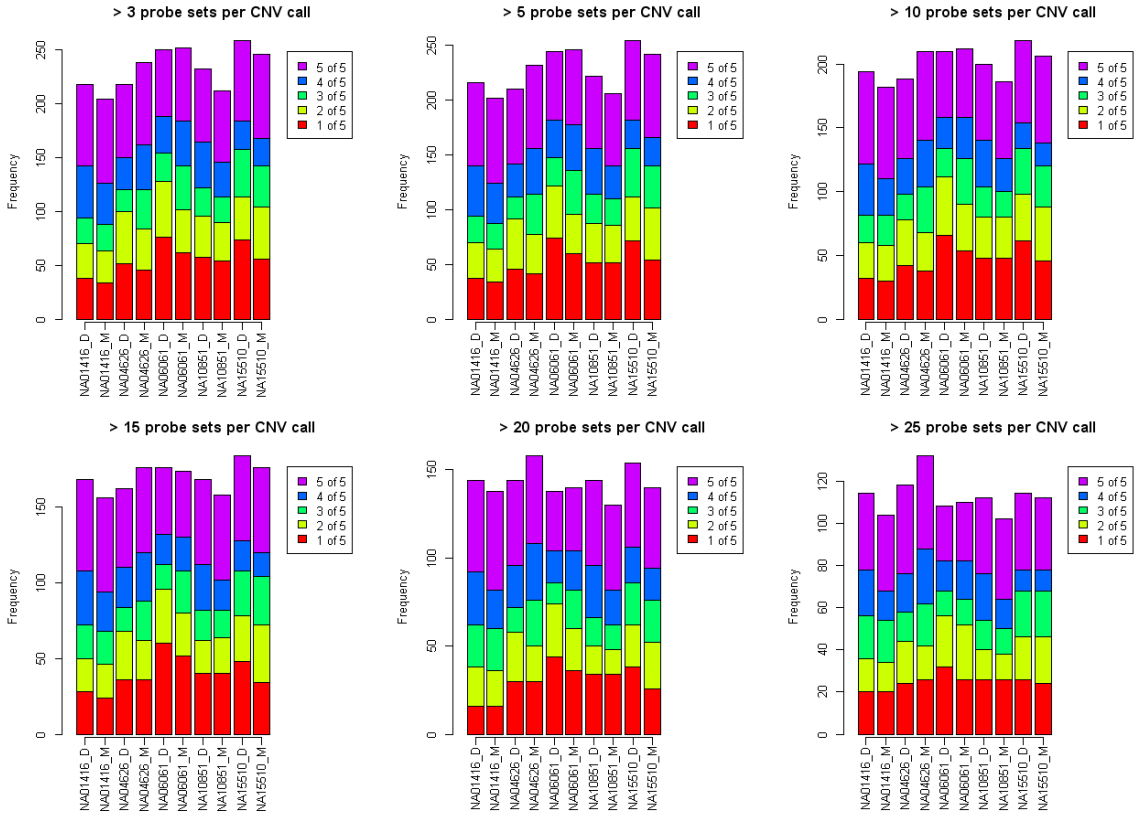

Figure 6.8: Reproducibility of CNV calls from five replicate's data sets for a total of five individuals. D: Default calling, M: MCMR calling

Table 6.2: Reproducibility rates of CNV calls using five replicate's data sets on same individuals (n = 5).

claimed # of probe sets per CNV call	mean number of sub-CNVRs with Default [MCMR] reference	repro-ducibi-lity	mean of reproducibility rates[a]		sd[b] of reproducibility rates[a]		p-value, two-sided paired t-test
			Default Calling	MCMR Calling	Default Calling	MCMR Calling	
> 3	235.2 [230.4]	≥ 40%	74.93%	78.23%	5.03%	3.70%	0.0675
		≥ 60%	57.05%	61.63%	7.04%	4.87%	0.1441
		≥ 80%	45.28%	47.72%	7.57%	5.82%	0.1304
		100%	29.77%	32.00%	3.68%	4.03%	0.0088
> 5	229.2 [225.6]	≥ 40%	75.68%	78.62%	5.10%	3.75%	0.1265
		≥ 60%	58.01%	62.36%	6.50%	4.76%	0.1742
		≥ 80%	45.94%	48.15%	7.36%	5.62%	0.1888
		100%	30.21%	32.49%	3.75%	3.95%	0.0122
> 10	202.0 [199.2]	≥ 40%	75.46%	78.36%	5.76%	4.23%	0.1463
		≥ 60%	57.86%	61.51%	8.12%	5.81%	0.2524
		≥ 80%	45.67%	46.79%	8.61%	6.01%	0.4715
		100%	30.84%	32.73%	4.58%	5.00%	0.0357
> 15	171.6 [168.0]	≥ 40%	75.42%	77.93%	6.35%	5.62%	0.1482
		≥ 60%	58.88%	61.58%	9.07%	6.28%	0.2909
		≥ 80%	46.83%	46.67%	8.17%	7.38%	0.8683
		100%	31.32%	32.82%	4.02%	5.33%	0.1167
> 20	144.8 [141.2]	≥ 40%	77.58%	79.80%	7.52%	6.00%	0.2720
		≥ 60%	60.95%	65.07%	9.93%	6.34%	0.1677
		≥ 80%	48.59%	49.57%	7.77%	5.97%	0.3683
		100%	31.72%	33.54%	4.33%	5.62%	0.1634
> 25	113.2 [112.0]	≥ 40%	77.29%	78.10%	4.48%	2.65%	0.6109
		≥ 60%	60.64%	61.98%	7.67%	6.37%	0.4285
		≥ 80%	46.18%	46.64%	6.94%	5.90%	0.7506
		100%	30.99%	32.20%	4.21%	4.52%	0.4212

[a] The x% reproducibility rate equals the proportion of individual sub-CNVRs with consistent CNV calling results in at least x% of the individual's replicates.
[b] sd = standard deviation.

6.2.3.3 False Positives and Negatives for CNVs of NA15510

Individual NA15510 has previously been investigated using various technical and algorithmic approaches, as for instance by Tuzun et al. (2005), Korbel et al. (2007), Redon et al. (2006) and Kidd et al. (2008). In all these four reports, CNV results were followed up by at least one alternative experimental method. CNVs that could be validated in this way were considered in the following investigations of false positive and false negative CNV calls of NA15510's five replicates.

A comparison of previous results is summarized in Table 6.3 and in Figure 6.9. Apart from results of Tuzun et al. (2005) and Kidd et al. (2008), the overlap of CNVs across the studies is in general low. Out of a total of 681 CNV segments, only four segments were consistently reported in all four publications. The majority of CNVs (67.40%) was exclusively detected in one study. Due to this inconsistency in CNV results across various designs, quality assessment for CNV detection algorithms is complicated. Here, false positve and false negative rates for replicates of NA15510 are likewise given with respect to each previous study.

Table 6.3: Overlap between CNVs for NA15510 reported by four publications.

All reported CNV calls

overlap in %*	Kidd	Korbel	Redon	Tuzun
Kidd ($n = 248$)	-	22.18%	5.65%	77.82%
Korbel ($n = 340$)	16.18%	-	8.24%	14.71%
Redon ($n = 160$)	8.75%	17.50%	-	7.50%
Tuzun ($n = 218$)	88.53%	22.94%	5.50%	-

Validated reported CNV calls

overlap in %*	Kidd	Korbel	Redon	Tuzun
Kidd ($n = 198$)	-	17.17%	7.07%	41.41%
Korbel ($n = 114$)	28.95%	-	5.26%	13.16%
Redon ($n = 125$)	11.20%	4.80%	-	4.80%
Tuzun ($n = 95$)	86.32%	15.79%	6.32%	-

* CNV overlap has been calculated at the CNV segment level (see Figure 6.5), that is several overlapping CNV calls of the same type were combined into one CNV segment.

Concerning previous reports, there seems to be a trend of longer CNVs being more concordantly detected by use of different technical and methodical approaches (Figure 6.9). However, this trend is not statistically significant (Kruskal-Wallis rank sum test; all CNVs: p-value = 0.32, validated CNVs only: p-value = 0.17).

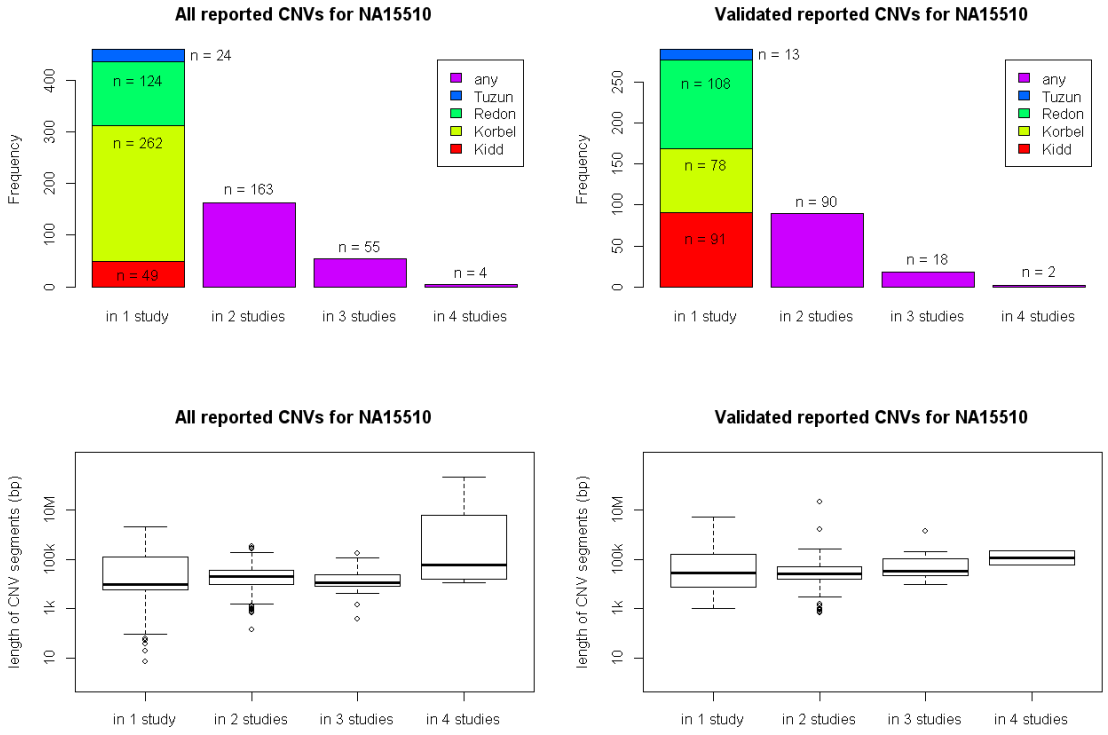

Figure 6.9: Similarity of CNVs for NA15510 reported by four publications.

A variety of false negative CNV reports in several studies is potentially caused by technological limitations coupled with restricted genomic resolution. As shown in Figure 6.10, the distributions of CNV lengths across studies are statistically significantly different depending on the applied experimental method (Kruskal-Wallis rank sum test p-value $< 2.2 \times 10^{-16}$).

Although gains and losses < 1 kb were in fact primarily not regarded as a CNV, this definition increasingly broke down with the ability of sequencing technologies to detect much smaller variants with varying copy number in a population. Thus, currently listed CNVs in the Database of Genomic variants (DGV) additionally encompass such small deletions and insertions of less than 1 kb. Consequently, all gain or loss data were considered as CNVs here. In contrast to the application of sequencing techniques (Tuzun et al. (2005), Korbel et al. (2007) and Kidd et al. (2008)), the use of former SNP arrays limited CNV detection to the identification of larger variants due to their sparse probe density on the genome (Redon et al. (2006)). However, since the genomic probe coverage of currently available SNP genotyping arrays, such as the Affymetrix 6.0 array, is doubled relative to those SNP arrays used by Redon et al. (2006), CNVs of almost the complete size spectrum became detectable without sequencing approaches (see results for PennCNV in Figure 6.10).

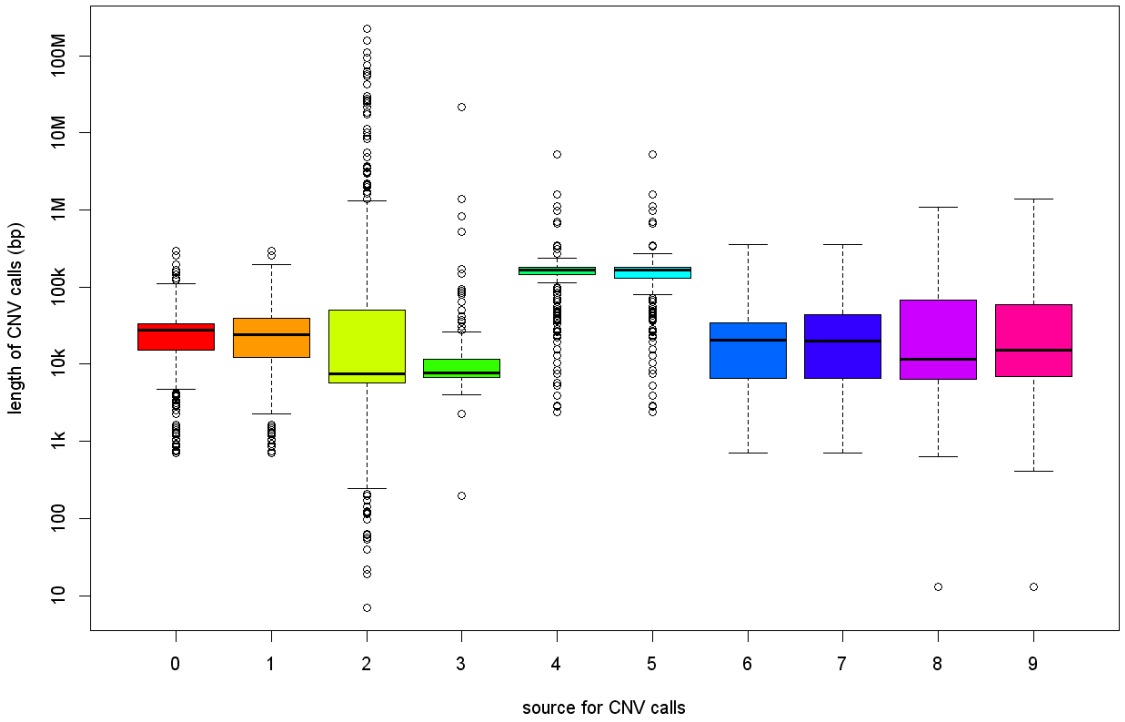

Figure 6.10: Distribution of CNV length for NA15510 in four publications and in application of PennCNV with Default and MCMR reference values. Sources for CNV calls: 0 = Kidd et al. (2008), 1 = only validated CNVs from Kidd et al. (2008), 2 = Korbel et al. (2007), 3 = only validated CNVs from Korbel et al. (2007), 4 = Redon et al. (2006), 5 = only validated CNVs from Redon et al. (2006), 6 = Tuzun et al. (2005), 7 = only validated CNVs from Tuzun et al. (2005), 8 = PennCNV with MCMR reference intensity values, 9 = PennCNV with Default reference intensity values.

CNV calls for the sample NA15510 were estimated in application of the PennCNV software in consideration of median (Default) and MCMR reference intensity values, respectively. The percentage of false positive and false negative findings was evaluated with respect to results of the four previous publications (Table 6.4). In more detail, results of each single study, the respective validated CNV results as well as the combined set of all reported or all validated CNVs were considered as gold standard for sensitivity and specificity analyses. Thus, a total of ten different gold standard test settings were investigated in order to address the above mentioned difficulty in finding an appropriate global CNV reference set. Apart from false negative estimates, results for Redon et al. (2006) were identical to those for only validated CNVs of Redon et al. (2006). As expected, false negative estimates were lower with respect to validated CNVs. However, this was consistently observed for both, De-

fault and MCMR reference values, whereas the respective specificity estimates were identical. Thus, false positive and false negative estimates are only presented once for Redon et al. (2006) in Table 6.4.

Both, false positive and false negative estimates were calculated at the CNV segment level (see Figure 6.5). For each replicate, the percentage of false positive CNV calls was defined to equal the number of called but not confirmed CNV segments among the total number of called CNV segments. Contrarily, the percentage of false negative CNV calls was calculated as the percentage of non-called CNVs among the set of gold standard CNVs. Estimates for false positive and negative rates were determined in application of the CNV detection thresholds of more than $3, 5, 10, 15, 20$ and 25 probe sets per CNV call and for confirmation thresholds of more than $0\%, 50\%, 80\%$ and exactly 100% overlap between PennCNV's and gold standard CNV calls. In order to allow a fair comparison, CNV detection thresholds were likewise applied to gold standard CNVs.

Differences for Default and MCMR based CNV calls showed comparable directional trends across CNV detection and confirmation thresholds. Since the inclusion of CNVs that involved more than 20 array probe sets yielded the highest stability and reproducibility rates (see Table 6.1 and Table 6.2), only results for this CNV detection threshold of > 20 array probe sets per CNV are presented in Table 6.4.

Moreover, results for confirmation thresholds of more than 80% and exactly 100% overlap were similar, whereas false positive and false negative estimates were lowest when any overlap between PennCNV's CNV calls and gold standard CNVs was considered as a confirmation of CNV calls. Consequently, results are only given with respect to a confirmation threshold of more than 0% overlap in Table 6.4.

On the one hand, the number of called CNVs is lower in application of MCMR reference intensity values (n=43) compared to Default reference intensity values (n=49). On the other hand, the percentage of false positive CNV calls was for almost all gold standards, with the exception of one (Korbel et al. (2007)), reduced by application of alternative MCMR reference intensity values. For most gold standards, the false positive rate was even statistically significantly reduced by on average 2.1% and especially by up to 4% with respect to the set of validated reference CNVs from all four publications. Moreover, the percentage of false negative CNVs was overall comparable between Default and MCMR based CNV calls with an average increase of 0.6% false negatives for alternative MCMR reference intensity values. However, there was no difference at all for half of the considered gold standard CNV sets, and with respect to the set of all validated reported CNVs an increase of 0.3% false negatives was observed for MCMR based CNVs realative to Default CNV calls.

Table 6.4: Genome-wide false positive and false negative estimates for CNV calls of five replicates for NA15510.

Gold Standard	replicate	confirmed[a]		non-confirmed[b]		total		% false positives[c]		% false negatives[d]	
		Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling
Korbel	replicate 1	25	23	17	16	42	39	40.48%	41.03%	66.22%	68.49%
	replicate 2	21	20	18	17	39	37	46.15%	45.95%	71.62%	72.60%
	replicate 3	22	21	16	15	38	36	42.11%	41.67%	70.27%	71.23%
	replicate 4	22	20	15	14	37	34	40.54%	41.18%	70.27%	72.60%
	replicate 5	22	20	15	15	37	35	40.54%	42.86%	70.27%	72.60%
	average [total, unique]	22.4 [28]	20.8 [25]	16.2 [21]	15.4 [18]	38.6 [49]	36.2 [43]	41.96%	42.54%	69.73%	71.50%
	p-value[e], two-sided paired t-test	-		-		-		0.3011		0.0057	
Korbel - validated	replicate 1	12	11	30	28	42	39	71.43%	71.79%	55.56%	59.26%
	replicate 2	10	10	29	27	39	37	74.36%	72.97%	62.96%	62.96%
	replicate 3	10	10	28	26	38	36	73.68%	72.22%	62.96%	62.96%
	replicate 4	10	9	27	25	37	34	72.97%	73.53%	62.96%	66.67%
	replicate 5	11	10	26	25	37	35	70.27%	71.43%	59.26%	62.96%
	average [total, unique]	10.6 [14]	10 [12]	28 [35]	26.2 [31]	38.6 [49]	36.2 [43]	72.54%	72.39%	60.74%	62.96%
	p-value[e], two-sided paired t-test	-		-		-		0.3693		0.0705	

Table 6.4: continued.

Gold Standard	replicate	confirmed[a]		non-confirmed[b]		total		% false positives[c]		% false negatives[d]	
		Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling
Kidd	replicate 1	22	21	20	18	42	39	47.62%	45.15%	86.00%	86.67%
	replicate 2	21	21	18	16	39	37	46.15%	43.24%	86.67%	86.67%
	replicate 3	20	20	18	16	38	36	47.37%	44.44%	88.00%	88.00%
	replicate 4	19	19	18	15	37	34	48.65%	44.12%	88.00%	88.00%
	replicate 5	18	19	19	16	37	35	51.35%	45.71%	88.00%	87.33%
	average [total, unique]	20 [24]	20 [23]	18.6 [26]	16.2 [21]	38.6 [49]	36.2 [43]	48.23%	44.73%	87.33%	87.33%
	p-value[e], two-sided paired t-test	-		-		-		0.0084		1.0000	
Kidd - validated	replicate 1	20	19	22	20	42	39	52.38%	51.28%	83.19%	84.03%
	replicate 2	19	19	20	18	39	37	51.28%	48.65%	84.03%	84.03%
	replicate 3	18	18	20	18	38	36	52.63%	50.00%	85.71%	85.71%
	replicate 4	17	17	20	17	37	34	54.05%	50.00%	85.71%	85.71%
	replicate 5	17	18	20	17	37	35	54.05%	48.57%	85.71%	84.87%
	average [total, unique]	18.2 [22]	18.2 [21]	20.4 [27]	18 [22]	38.6 [49]	36.2 [43]	52.88%	49.70%	84.87%	84.87%
	p-value[e], two-sided paired t-test	-		-		-		0.0128		1.0000	

Table 6.4: continued.

Gold Standard	replicate	confirmed[a]		non-confirmed[b]		total		% false positives[c]		% false negatives[d]	
		Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling
Tuzun	replicate 1	19	18	23	21	42	39	54.76%	53.85%	85.27%	86.05%
	replicate 2	18	18	21	19	39	37	53.85%	51.85%	86.05%	86.05%
	replicate 3	17	17	21	19	38	36	55.26%	52.78%	87.60%	87.60%
	replicate 4	16	16	21	18	37	34	56.76%	52.94%	87.60%	87.60%
	replicate 5	16	17	21	18	37	35	56.76%	51.43%	87.60%	86.82%
	average [total, unique]	17.2 [21]	17.2 [20]	21.2 [28]	19 [23]	38.6 [49]	36.2 [43]	55.48%	54.47%	86.82%	86.82%
	p-value[e], two-sided paired t-test	-		-		-		0.0154		1.0000	
Tuzun - validated	replicate 1	8	7	34	32	42	39	80.95%	82.05%	87.10%	88.71%
	replicate 2	7	7	32	30	39	37	82.05%	81.08%	88.71%	88.71%
	replicate 3	8	8	30	28	38	36	78.95%	77.78%	88.71%	88.71%
	replicate 4	7	7	30	27	37	34	81.08%	79.41%	88.71%	88.71%
	replicate 5	7	7	30	28	37	35	81.08%	80.00%	88.71%	88.71%
	average [total, unique]	7.4 [9]	7.2 [8]	31.2 [40]	29 [35]	38.6 [49]	36.2 [43]	80.82%	80.06%	88.39%	88.71%
	p-value[e], two-sided paired t-test	-		-		-		0.1892		0.3739	

Table 6.4: continued.

Gold Standard	replicate	confirmed[a]		non-confirmed[b]		total		% false positives[c]		% false negatives[d]	
		Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling
Redon	replicate 1	20	20	22	19	42	39	52.38%	48.72%	86.67%	86.67%
	replicate 2	18	18	21	19	39	37	53.85%	51.35%	88.00%	88.00%
	replicate 3	18	18	20	18	38	36	52.63%	50.00%	88.00%	88.00%
	replicate 4	18	18	19	16	37	34	51.35%	47.06%	88.00%	88.00%
	replicate 5	18	18	19	17	37	35	51.35%	48.57%	88.00%	88.00%
	average [total, unique]	18.4 [20]	18.4 [20]	20.2 [29]	17.8 [23]	38.6 [49]	36.2 [43]	52.31%	49.14%	87.73%	87.73%
	p-value[e], two-sided paired t-test	-		-		-		$7.82E-04$		1.0000	
Korbel, Kidd, Tuzun & Redon	replicate 1	38	35	4	4	42	39	9.52%	10.26%	86.71%	87.72%
	replicate 2	34	33	5	4	39	37	12.82%	10.81%	88.11%	88.42%
	replicate 3	34	33	4	3	38	36	10.53%	8.33%	88.11%	88.42%
	replicate 4	33	31	4	3	37	34	10.81%	8.82%	88.46%	89.12%
	replicate 5	34	33	3	2	37	35	8.11%	5.71%	88.11%	88.42%
	average [total, unique]	34.6 [42]	33 [38]	4 [7]	3.2 [5]	38.6 [49]	36.2 [43]	10.36%	8.79%	87.90%	88.42%
	p-value[e], two-sided paired t-test	-		-		-		0.0542		0.0206	

Table 6.4: continued.

Gold Standard	replicate	confirmed[a]		non-confirmed[b]		total		% false positives[c]		% false negatives[d]	
		Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling	Default Calling	MCMR Calling
Korbel, Kidd, Tuzun & Redon - validated	replicate 1	36	34	6	5	42	39	14.29%	12.82%	84.62%	85.47%
	replicate 2	32	32	7	5	39	37	17.95%	13.51%	86.32%	86.32%
	replicate 3	32	32	6	4	38	36	15.79%	11.11%	86.75%	86.75%
	replicate 4	31	30	6	4	37	34	16.22%	11.76%	86.75%	87.18%
	replicate 5	32	32	5	3	37	35	13.51%	8.57%	86.32%	86.32%
	average [total, unique]	32.6 [40]	32 [37]	6 [9]	4.2 [6]	38.6 [49]	36.2 [43]	15.55%	11.55%	86.15%	86.41%
	p-value[e], two-sided paired t-test	-		-		-		0.0033		0.2071	

Both, PennCNV's CNV calls (Default and MCMR) as well as gold standard CNVs were claimed to harbour more than 20 informative Affymetrix 6.0 probe sets.

[a] Confirmed CNV calls are those that were likewise reported in the gold standard study.

[b] Non-confirmed CNV calls are those that were called but not reported otherwise.

[c] The percentage of false positives was calculated as the cardinal number of the overlap between PennCNV's CNV calls and gold standard CNVs divided by the total number of PennCNV's CNV calls.

[d] The percentage of false negatives was calculated as the number of gold standard CNVs that were not called with PennCNV divided by the total number of gold standard CNVs.

[e] P-values < 0.05 with an effect direction indicating a superiority (inferiority) of MCMR reference intensity values over Default reference intensity values are printed in red (green).

6.2.3.4 False Positives and Negatives for HapMap CNVs

Results of eight previous publications, namely Kidd et al. (2008), Park et al. (2010), Perry et al. (2008), Conrad et al. (2006), Shaikh et al. (2009), Conrad et al. (2010), McCarroll et al. (2008) and Redon et al. (2006), for CNVs of up to 270 HapMap individuals were used to estimate the percentage of false positive and false negative CNV calls derived from PennCNV in application of Default and MCMR reference intensity values.

At first, previously reported CNV results were compared against each other with regard to an overall concordance. Since not all studies investigated CNVs on all 270 HapMap samples, only CNVs of those individuals that were analysed in both of any pair of two publications were checked for consistency at the CNV segment level (see Figure 6.5 for details). As presented in Table 6.5, the pairwise overall overlap between any two previous reports on HapMap CNVs is low, it ranges from 0.62% to 67.8% (mean = 24.03%, median = 16.73%).

Table 6.5: Pair-wise overall between HapMap CNVs reported by eight publications.

overlap in % [# of samples]*	Kidd	Park	Perry	Conrad (Mendel)	Shaikh	Conrad (Tiling)	McCarroll	Redon
Kidd	-	-	9.0 [2]	0.6 [3]	1.7 [3]	10.8 [8]	6.1 [8]	1.7 [8]
Park	-	-	16.8 [5]	-	2.7 [8]	45.9 [20]	14.4 [20]	4.3 [20]
Perry	10.0 [2]	21.1 [5]	-	1.2 [12]	3.1 [11]	18.0 [30]	12.3 [30]	10.4 [30]
Conrad (Mendel)	44.0 [3]	-	45.0 [12]	-	28.6 [27]	54.4 [60]	46.8 [60]	19.4 [60]
Shaikh	40.6 [3]	67.8 [8]	49.1 [11]	12.8 [27]	-	54.5 [111]	48.0 [111]	26.7 [111]
Conrad (Tiling)	16.6 [8]	46.7 [20]	13.1 [30]	1.1 [60]	2.3 [111]	-	13.1 [270]	3.1 [270]
McCarroll	43.1 [8]	62.9 [20]	40.6 [30]	4.2 [60]	8.9 [111]	58.3 [270]	-	13.1 [270]
Redon	25.1 [8]	39.1 [20]	66.1 [30]	3.2 [60]	9.3 [111]	26.4 [270]	25.1 [270]	-

* CNV overlap has been calculated at the CNV segment level (see Figure 6.5 for details). The presented percentaged overlap equals the proportion of CNVs reported by the publication in row that were also reported by the publication in column.

Furthermore, the concordance of results from those three studies (Conrad et al., 2010; McCarroll et al., 2008; Redon et al., 2006) that investigated CNVs on all 270 HapMap samples is depicted in Figure 6.11. In all these publications, CNVs were derived from data of SNP genotyping arrays of different type and resolution. Only 1.66% of all reported CNVs can be consistently found in all three publications, and 11.44% are consistent throughout exactly two publications. Thus, the majority (86.90%) of previously published HapMap CNVs was only stated once. Consequently, false positive and false negative estimates being calculated with respect to previous results, can only be seen as a rough potentially biased estimate.

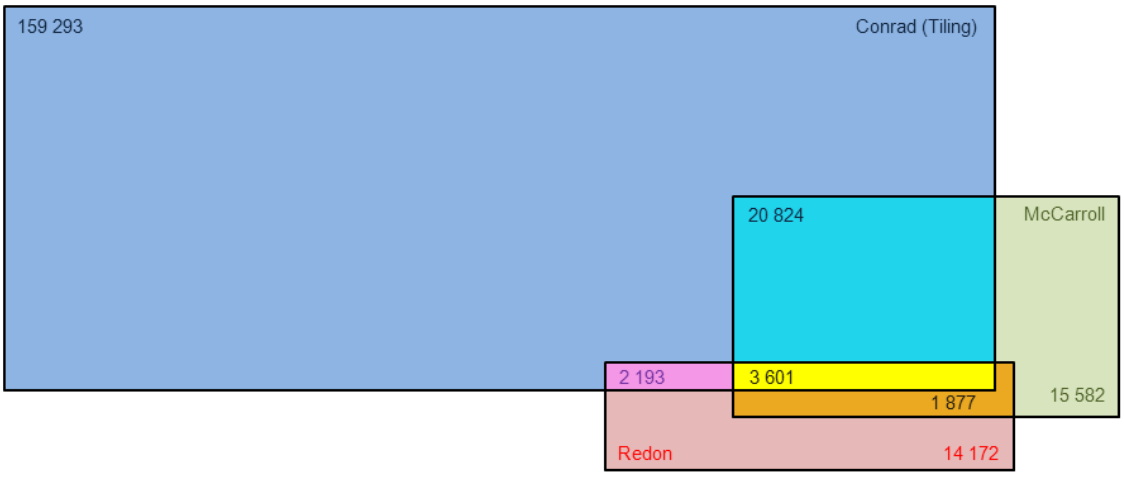

Figure 6.11: Venn diagram for CNVs of 270 HapMap individuals from three studies (Conrad et al. (2010), McCarroll et al. (2008) and Redon et al. (2006)).

The percentage of false positive and false negative PennCNV's CNV calls was calculated for CNV detection thresholds of more than $3, 5, 10, 15, 20$ and 25 probe sets per PennCNV and gold standard CNVs. Moreover, confirmation thresholds of more than $0\%, 50\%, 80\%$ and exactly 100% reciprocal overlap were considered with respect to each as well as to all previous publications as gold standard. Differences for Default and MCMR calling were directionally consistent across different CNV detection and confirmation thresholds. With regard to results of previous chapters (see Table 6.1 and Table 6.2), comparative results on false positive and false negative estimates are presented in Table 6.6 and Figure 6.12 for a CNV detection threshold of more than 20 probe sets per CNV call and for a confirmation threshold that implicates any overlap between experiment (PennCNV) and gold standard as a confirmation event.

Table 6.6: Genome-wide false positive and false negative estimates for CNV calls of up to 270 HapMap individuals.

Gold Standard [# of samples[a]]	Calling	# of CNV calls			% of false positives[b]			% false negatives[c]		
		total	sample's median [mean]	p-value[d]	total	sample's median [mean]	p-value[d]	total	sample's median [mean]	p-value[d]
Kidd [8]	Gold	3 917	470.5 [489.62]	-	-	- [-]	-	-	- [-]	-
	Default	338	41.5 [42.25]	0.1090	53.85	56.32 [54.27]	0.1028	94.64	92.12 [88.80]	0.6758
	MCMR	325	39.5 [40.62]		51.69	49.89 [52.01]		94.59	92.12 [89.02]	
Kidd - validated [8]	Gold	1 077	152.5 [134.62]	-	-	- [-]	-	-	- [-]	-
	Default	338	41.5 [42.25]	0.1090	55.92	57.27 [56.37]	0.1064	81.24	81.77 [78.48]	0.8687
	MCMR	325	39.5 [40.62]		53.85	51.33 [54.13]		81.06	80.47 [78.59]	
Park [20]	Gold	1 734	71.5 [86.70]	-	-	- [-]	-	-	- [-]	-
	Default	773	40 [38.65]	0.8180	28.59	26.38 [28.36]	0.1450	67.24	58.75 [63.16]	0.9258
	MCMR	771	40.5 [38.55]		27.37	26.28 [27.04]		67.19	57.73 [63.11]	
Perry [30]	Gold	2 598	86 [86.60]	-	-	- [-]	-	-	- [-]	-
	Default	1 231	43 [41.03]	0.1364	30.22	30.5 [29.95]	0.0041	67.51	67.60 [67.32]	0.2689
	MCMR	1 215	40.5 [40.50]		28.97	29.8 [28.70]		67.90	67.07 [67.68]	
Conrad (Mendel) [60]	Gold	618	10 [10.47]	-	-	- [-]	-	-	- [-]	-
	Default	1 642	28 [27.83]	0.3842	84.10	84.85 [83.69]	0.0883	47.25	50 [44.84]	0.5208
	MCMR	1 649	28 [27.95]		84.29	84.85 [83.98]		47.41	50 [45.30]	
Shaikh [112]	Gold	919	8 [8.28]	-	-	- [-]	-	-	- [-]	-
	Default	4 777	43 [43.04]	1.21E-08	76.43	76.47 [76.19]	2.09E-05	7.18	0 [6.53]	0.9040
	MCMR	4 635	41 [41.76]		75.71	75.68 [75.53]		7.18	0 [6.57]	

Table 6.6: continued.

Gold Standard [# of samples[a]]	Calling	# of CNV calls total	# of CNV calls sample's median [mean]	# of CNV calls p-value[d]	% of false positives[b] total	% of false positives[b] sample's median [mean]	% of false positives[b] p-value[d]	% false negatives[c] total	% false negatives[c] sample's median [mean]	% false negatives[c] p-value[d]
Shaikh - validated [112]	Gold	867	7 [7.81]	-	-	- [-]	-	-	- [-]	-
	Default	4 777	43 [43.04]	1.21E-08	77.35	76.92 [77.03]	5.76E-05	2.08	0 [1.81]	0.1999
	MCMR	4 635	41 [41.76]		76.68	77.08 [76.43]		2.42	0 [2.17]	
Conrad (Tiling) [270]	Gold	21 529	80 [80.33]	-	-	- [-]	-	-	- [-]	-
	Default	12 330	43 [46.01]	4.99E-12	53.50	49.94 [49.57]	0.0024	70.73	71.16 [70.89]	6.55E-26
	MCMR	12 017	41 [44.84]		53.84	50 [50.02]		71.95	72.22 [72.11]	
McCarroll [270]	Gold	14 922	55 [55.47]	-	-	- [-]	-	-	- [-]	-
	Default	12 357	43 [45.94]	6.03E-12	43.13	37.93 [38.49]	9.56E-19	52.20	51.72 [52.35]	0.1571
	MCMR	12 045	41 [44.78]		41.70	36.54 [36.94]		52.41	52.17 [52.57]	
Redon [270]	Gold	17 302	63 [64.32]	-	-	- [-]	-	-	- [-]	-
	Default	12 357	43 [45.94]	6.03E-12	61.23	57.69 [57.83]	2.38E-06	82.77	82.50 [81.81]	0.0285
	MCMR	12 045	41 [44.78]		60.50	56.86 [57.15]		82.87	82.81 [81.99]	
any [270]	Gold	66 696	225 [247.94]	-	-	- [-]	-	-	- [-]	-
	Default	12 357	43 [45.94]	6.03E-12	24.36	17.31 [18.19]	5.97E-05	67.07	65.52 [65.77]	1.99E-09
	MCMR	12 045	41 [44.78]		23.76	16.67 [17.64]		67.53	65.96 [66.29]	

[a] Only CNV calls of samples that were analysed in the gold standard study were considered in the comparison of Default and MCMR calling.
[b] CNV calls with any overlap between gold standard and Default or MCMR calling were considered as being recovered, respectively.
[c] CNV calls with no overlap between gold standard and Default or MCMR calling were considered as being unrecovered, respectively.
[d] Two-sided p-values of paired t-tests are reported. P-values < 0.05 with a superior (inferior) effect of MCMR references relative to Default ones are printed red (green).

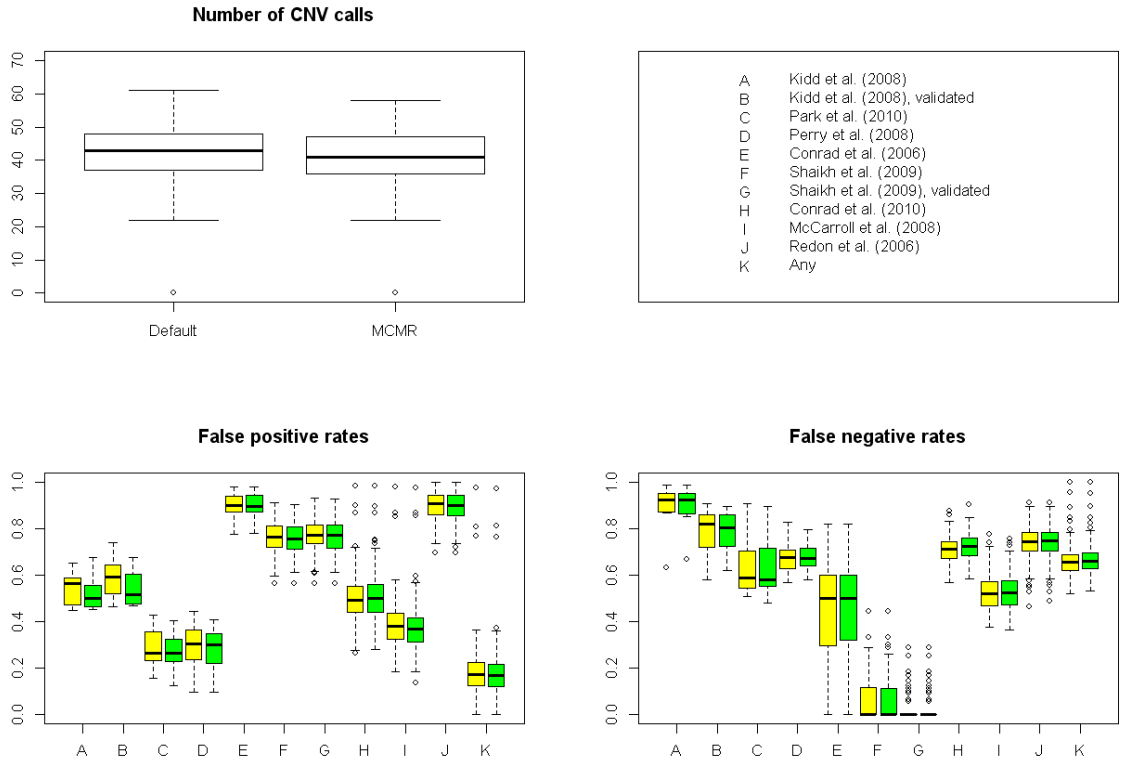

Figure 6.12: Number, false positive and negative rates of Default and MCMR based CNV calls of up to 270 HapMap individuals.

Regarding the number of individual CNV calls, an MCMR based CNV detection yields statistically significantly less CNV calls compared to the Default approach. On average, the mean number of CNV calls per individual was reduced by 0.998 calls (2.26%) when applying alternative MCMR reference intensity values. Moreover, the percentage of false positive MCMR CNV calls was reduced with respect to almost all considered gold standards. With increasing sample size, statistical significance for a reduction of the false positive rate was increasingly observable. In particular, for all studies that incorporated more than 100 individuals, the percentage of false positives was statistically significantly reduced. With MCMR calling, a 0.94% lower mean false positive rate was reached on average. Mean false negative rates tend to be higher with MCMR reference intensity values by on average 0.33%.

6.2.3.5 Mendelian Inconsistency Rates for HapMap CNVs

Another measure for comparing the accuracy of CNV detection algorithms is the rate of Mendelian inconsistent CNV calls in family data. Previous results of McCarroll et al. (2008) and Redon et al. (2006) demonstrate that more than 99% of all CNV events are derived from inheritance rather than from new mutations. Thus, observed

Mendelian discordance across estimated familial CNVs is most likely caused by a misclassification of CNVs rather than by de novo events at these loci.

For a total of 60 HapMap trios (30 CEPH & 30 YRI), Mendelian inconsistency rates were investigated for CNV calls based either on Default or MCMR reference intensity values. For each trio, the proportion of those offspring's CNVs was determined that were not estimated to be derived from parental CNVs in a Mendelian mode of inheritance. As shown in Table 6.7, there was no statistically significant difference in Mendelian inconsistency rates across 60 HapMap trios between those CNVs that were derived from the Default or from the alternative (MCMR) approach, respectively.

Table 6.7: Mendelian Inconsistency Rates in offspring's CNVs of 60 HapMap trios.

number of CNV calls in offspring	mean	Default Calling	46.33
		MCMR Calling	45.47
	median	Default Calling	45.0
		MCMR Calling	44.0
Mendelian inconsistency rate* [%]	mean	Default Calling	17.2
		MCMR Calling	17.6
	median	Default Calling	15.9
		MCMR Calling	17.1
	p-value, two-sided paired t-test		0.4253

* The Mendelian inconsistency rate was calculated as the proportion of offspring's CNVs that are not estimated to be derived from Mendelian inheritance.

6.2.4 Discussion

Validity of CNV calls is an essential component of association studies of CNVs with disease. Modification of strategy S2 into the proposed analysis strategy PS2 was motivated by the desire to improve CNV genotyping accuracy and thereby improving the meaningfulness of subsequent association test results. Consequently, the separate investigation of the isolated CNV calling step of both strategies was considered to be appropriate prior to the presentation of the whole strategy's implementation, which is given in the following chapter. In our evaluation of the two involved intensity reference models for CNV identification, we found considerable variation among the reference models in terms of the number of CNVs called, their stability and reproducibility rates as well as false positive and false negative estimates.

For stability rates of replicate CNVs detected by use of MCMR and Default reference values, alternative reference intensity values were overall superior to the most

widely used median ones. The differences declined when the number of minimal probe sets per CNV call increased. However, stability rates for both references were poor when an absolute identity between individual replicate CNV calls was considered as a pairwise concordance event. For CNVs containing more than 20 probe sets and any pairwise overlap being considered as a stable calling event, MCMR and Default references showed highest median stability rates of 89.97% and 89.51%. To some extent, the comparability between stability rates of MCMR and Default is limited by the small number of available replicate data sets, which was only five replicates for each of five individuals.

Consistently, MCMR calling showed higher mean reproducibility rates compared to Default calling. As with stability rates, the superiority of MCMR calling decreased with increasing CNV detection threshold, i.e. with higher minimal number of involved array probe sets. In case of Default calling, mean reproducibility rates were lowest for the maximal set of all called CNVs throughout all concordance levels ($40\%, 60\%, 80\%$ and 100%). For MCMR calling, no connection was observed between lowest reproducibility rates and CNV detection thresholds. Highest reproducibility rates were, however, concordantly found when including only CNV calls with more than 20 probe sets for both calling approaches (Default: 77.6%, MCMR: 79.8%). At least for the five individuals with available replicate data, the number of MCMR and Default CNV calls and subCNVRs reached comparable levels. Thus, higher stability and reproducibility of MCMR calling may indeed indicate its improved reliablity over Default calling. However, generality of this ovservation is limited by the small number of available replicate data ($n = 5$).

Concerning sample NA15510, MCMR calling produced six fewer calls relative to the 49 standard derived CNVs. Being confronted with a variety of differently composed CNV gold standards, MCMR calls were continuously assessed to include a lower percentage of false positives with simultaneously unchanged false negative rates in comparison to Default calling. Relatively poor consistence was found between the four considered gold standards. Investigation of CNV call's false rates is limited by this lack of a gold standard, since some type of CNVs might be over- or under-represented in recent reports. It is challenging to assess whether both calling approaches are likewise affected by these inadequacies. Future next-generation sequencing might overcome this concern in providing an ultimate gold standard.

Observations from the investigation of sample NA15510 could be strengthened by respective evaluation of false CNV rates for up to 270 HapMap samples. Again, the number of CNV calls tended to be lower with MCMR calling. Moreover, false positive rates were reduced compared to standard calling for almost all considered gold

standards. However, false negative rates tend to be higher by alternative MCMR reference values. Thus, the reduced number of CNV calls might end up with loosing true CNV findings when switching from standard to MCMR calling.

Finally, an absence of any difference in genome-wide Mendelian inconsistency rates was observed across 60 HapMap trios. This underlines that applying MCMR calling instead of Default calling, which was shown to offer the potential to improve validity of CNV calls, might overall lead to no more than moderately positive effects. However, on the one hand the number of available HapMap trios was limited to a relatively small number, which might cause a lack of statistical power to detect true underlying quality differences between Default and MCMR calling. On the other, it is quite conceivable that the correct classification of only one causal CNV loci might suffice to detect CNV association effects on disease status.

6.3 Application of Strategy PS2 to the Phenotype Obesity

The previous chapter focused on one particular aspect of the proposed analysis strategy PS2, namely the effect of a sophisticated intensity reference model on the quality of CNV calls. Contrarily, this chapter attempts to assess advantages and disadvantages of the whole strategy PS2. Due to their similar nature, a comparison of strategy PS2 with strategy S2 will be of particular interest. Aiming to asses potential consequences of the choice of the CNV analyses strategy on genetic results, the genome-wide data set of raw CNV data for 424 obesity trios was re-analysed in application of strategy PS2.

6.3.1 Data Set

The family-based sample was made up of 424 obesity trios, each comprising one extremely obese child or adolescent and both biological parents. Details on recruitment and phenotypical characteristics are given in chapter 3.5.1 and in Jarick et al. (2011) (Supplementary Table S1). Moreover, follow-up association analyses were additionally performed in a second family-based obesity sample of further 281 independent obesity trios that were similarly recruited and composed as the first sample (for details see Jarick et al. (2011), Supplementary Table S1). For both samples genotyping was performed on the Affymetrix 6.0 chip by the ATLAS Biolabs GmbH (Berlin, Germany) (for details see chapter 3.5.2).

6.3.2 Methods

For each individual of both trio samples, CNVs were detected by use of the PennCNV software (Wang et al., 2007) in application of MCMR reference intensity values (see chapter 6.2.2 for details). In the previous chapter, CNV calls based on Affymetrix 6.0 data with more than 20 consecutive informative probe sets were shown to be most reliable with regard to stability and recovery rates. Consequently, each CNV call that covered less than 21 consecutive informative probe sets was discarded from subsequent association tests.

As secondary analyses, Mendelian inconsistency rates were determined for each trio of both samples as the proportion of offspring's CNVs that were not called to be derived from parental CNVs. As primary analyses, estimated CNVs were tested for an association with the binary trait obesity in application of the FBAT approach by assuming an additive genetic effect model. In more detail, the coding for the different marker genotypes was specified to 0, 1, 2, 3, 4 in condordance with the estimated total unphased number of DNA segment copies. As described previously in detail, exclusively the set of unique CNV's start and end sites in 244 pre-specified genomic regions with a minimal copy number variability of five percent were tested for an association in order to avoid redundancies and to ensure a minimal number of informative families across FBATs (see chapter 3.6.1). Moreover, significance was assessed by use of the lfdr method, which accounts for the fact that multiple hypotheses were tested simultaneously (Efron et al. (2001), see chapter 3.5.2).

The most promising CNV region on chr 11q11 was technically validated by use of qPCR (Applied Biosystems, TaqMan assay Hs03802074_cn at chr 11: $55\,203\,791 \pm 50$ bp) by the Department of Child and Adolescent Psychiatry of the University of Duisburg-Essen. Validity was assessed in comparing array-based and qPCR-based results. Finally, CNV FBATs were re-calculated for all sites of this region with respect to qPCR derived CNV genotypes on all 705 obesity trios of both family-based samples.

6.3.3 Results

Genome-wide CNV calling in 424 *obesity trios*

A total of 47 825 CNVs were detected in the 1 272 individuals of the first family-based sample of 424 obesity trios, out of which 15 820 CNVs were observed in the offspring and 32 005 in the parents. Among those CNVs, 40 050 were located in the 244 pre-specified CNVRs with previously observed minimal copy number variability of five percent (offspring: n = 13 427, parents: n = 26 623).

Table 6.8: Mendelian Inconsistency Rates for CNVs of 705 Obesity Trios with Default and MCMR calling, respectively.

			in 424 obesity trios	in 281 obesity trios	in 705 obesity trios
number of CNV calls in offspring	mean	Default Calling	37.4	39.2	38.1
		MCMR Calling	37.3	39.8	38.3
	median	Default Calling	36.0	39.0	37.0
		MCMR Calling	36.0	39.0	38.0
Mendelian inconsistency rate* [%]	mean	Default Calling	24.8	19.0	22.4
		MCMR Calling	23.0	17.8	20.9
	median	Default Calling	23.1	18.6	21.1
		MCMR Calling	21.6	17.1	20.0
	p-value, two-sided paired t-test		4.06E-09	2.58E-05	5.41E-13

* The Mendelian inconsistency rate was calculated as the proportion of offspring's CNVs that were not estimated to be derived from Mendelian inheritance.

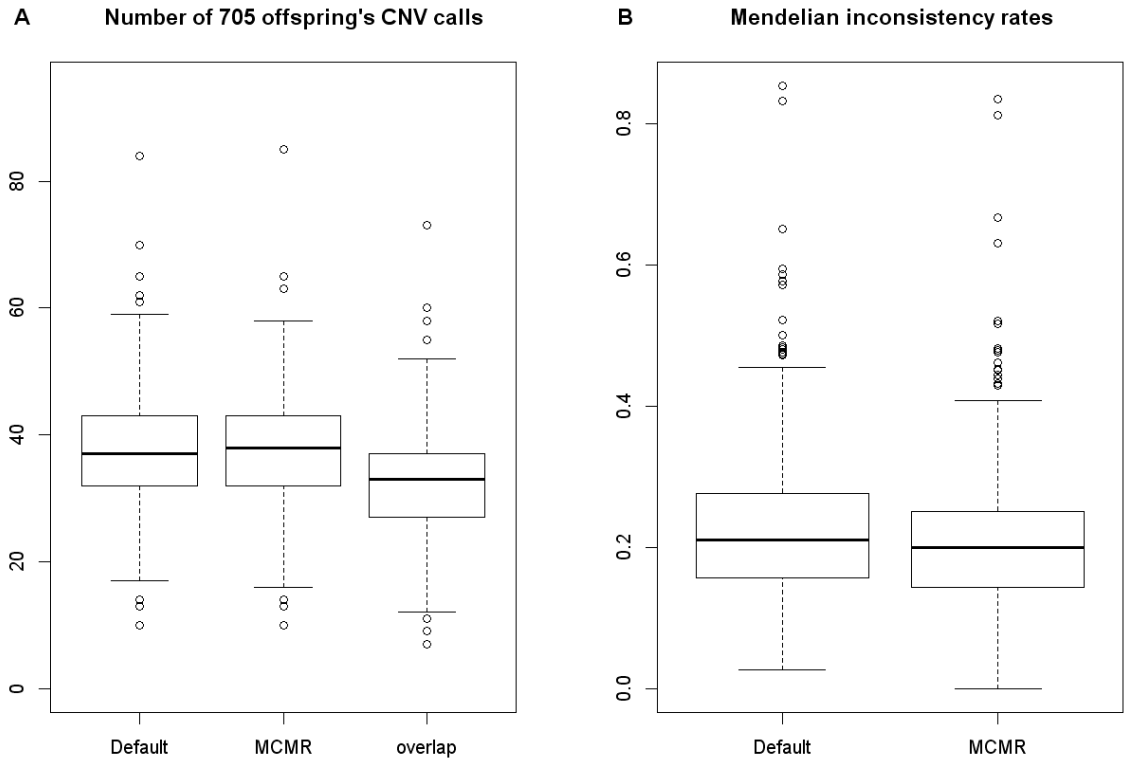

Figure 6.13: Number (Panel A) and Mendelian inconsistency rates (Panel B) of CNV calls from 705 obesity trios with Default and MCMR calling, respectively.

Compared to CNVs that were called in application of default parameters and Default median reference intensity values (see chapter 3.6 for details), the mean number of identified MCMR-based CNVs was similar in the group of 424 offspring. However, MCMR-based CNVs showed statistically significantly reduced Mendelian inconsistency rates in comparison to CNVs from a standard procedure ($p = 4.06 \times 10^{-9}$, see Table 6.8 and Figure 6.13).

Genome-wide association testing in 424 obesity trios

In the first family-based sample of 424 obesity trios, FBATs were performed for a total of 3 199 unique CNV's start and end sites at 244 CNVRs in the association testing step (Figure 6.14). 49 sites, reflecting 32 sub-CNVRs in seven CNVRs, yielded lfdr values below 0.2, which is a threshold to be sensible as proposed by Efron (2004) (see Figure 6.14 and Table 6.9).

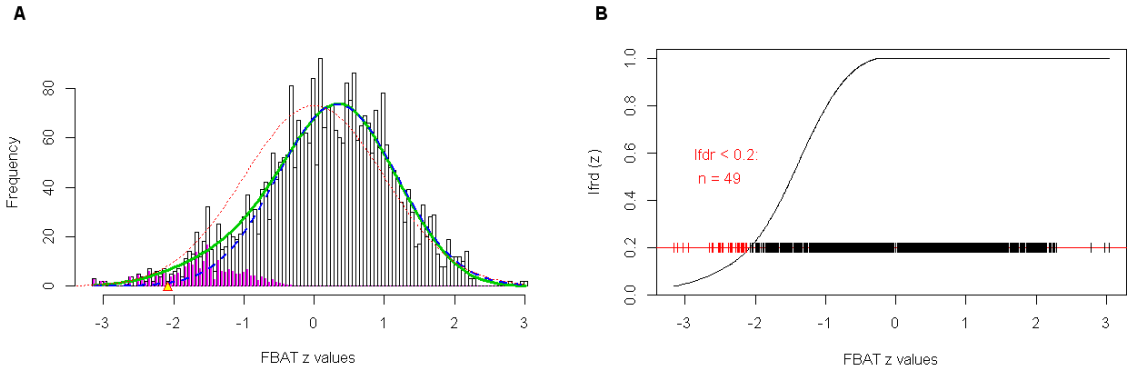

Figure 6.14: Histogram and lfdr curve of CNV FBAT z-values for the genome-wide analysis of 424 obesity trios at 3 199 unique CNV's start and end sites in 244 CNVRs.
Panel A: Histogram. The red dashed curve depicts the standard normal distribution, the dashed blue line is $\hat{p_0}\hat{f_0}$, the empirical null density, $\mathcal{N}(0.332, 0.851^2)$, and the green line is the empirically estimated mixture density. The small pink bars represent the estimated non-null counts.
Panel B: Lfdr curve as derived from empirical estimates of f_0, f and p_0 (Panel A). Observed CNV FBAT z-values are illustrated as ticks on the horizontal lines, those with lfdr < 0.2 are printed in red.

Table 6.9: Results for 32 sub-CNVRs reflecting 49 CNV's start and end sites with lfdr < 0.2 across 3 199 FBATs in 424 obesity trios at a total of 244 CNVRs. (in chromosomal order)

Chr: Position [hg18]	FBAT z-value	FBAT p-value	lfdr	in CNVR [chr: bp, hg18]
2: 41 091 947 - 41 099 391	−2.385	0.0171	0.124	2: 41 091 935 - 41 099 404
3: 131 245 549 - 131 246 387	−2.283	0.0224	0.145	3: 131 245 537 - 131 290 979
3: 131 269 711 - 131 269 888	−2.160	0.0308	0.176	3: 131 245 537 - 131 290 979
3: 131 269 889 - 131 271 914	−2.279	0.0227	0.146	3: 131 245 537 - 131 290 979
3: 131 271 915 - 131 273 779	−2.332	0.0197	0.134	3: 131 245 537 - 131 290 979
3: 131 273 780 - 131 274 037	−2.337	0.0194	0.134	3: 131 245 537 - 131 290 979
3: 131 274 038 - 131 274 200	−2.259	0.0239	0.150	3: 131 245 537 - 131 290 979
3: 131 274 201 - 131 274 319	−2.197	0.0280	0.166	3: 131 245 537 - 131 290 979
3: 131 274 320 - 131 276 344	−2.239	0.0252	0.155	3: 131 245 537 - 131 290 979
3: 131 276 345 - 131 276 696	−2.193	0.0283	0.167	3: 131 245 537 - 131 290 979
3: 131 276 697 - 131 277 782	−2.143	0.0321	0.181	3: 131 245 537 - 131 290 979
3: 131 277 783 - 131 281 471	−2.455	0.0141	0.112	3: 131 245 537 - 131 290 979
3: 131 281 472 - 131 282 357	−2.360	0.0183	0.129	3: 131 245 537 - 131 290 979
3: 131 282 358 - 131 288 741	−3.111	0.0019	0.039	3: 131 245 537 - 131 290 979
3: 131 288 742 - 131 288 926	−3.155	0.0016	0.038	3: 131 245 537 - 131 290 979
3: 131 289 689 - 131 291 500	−2.137	0.0326	0.182	3: 131 245 537 - 131 290 979
3: 196 868 323 - 196 875 860	−2.341	0.0192	0.133	3: 196 868 311 - 196 946 380
3: 196 875 861 - 196 884 255	−2.333	0.0197	0.134	3: 196 868 311 - 196 946 380
3: 196 884 256 - 196 895 126	−2.355	0.0185	0.130	3: 196 868 311 - 196 946 380
3: 196 895 127 - 196 901 833	−2.621	0.0088	0.088	3: 196 868 311 - 196 946 380
3: 196 901 834 - 196 907 468	−2.949	0.0032	0.052	3: 196 868 311 - 196 946 380
3: 196 907 469 - 196 914 787	−2.120	0.0340	0.187	3: 196 868 311 - 196 946 380
3: 196 914 788 - 196 928 237	−2.176	0.0296	0.171	3: 196 868 311 - 196 946 380
3: 196 928 238 - 196 928 253	−2.226	0.0260	0.158	3: 196 868 311 - 196 946 380
7: 133 446 382 - 133 448 649	−3.032	0.0024	0.045	7: 133 435 705 - 133 449 750
7: 133 448 650 - 133 449 098	−2.620	0.0088	0.088	7: 133 435 705 - 133 449 750
10: 58 186 381 - 58 191 255	−2.656	0.0079	0.083	10: 58 186 369 - 58 196 856
10: 58 191 256 - 58 196 843	−2.524	0.0116	0.101	10: 58 186 369 - 58 196 856
11: 55 209 586 - 55 210 152	−2.134	0.0329	0.183	11: 55 130 596 - 55 210 165
12: 9 529 176 - 9 542 559	−2.597	0.0094	0.091	12: 9 525 125 - 9 604 954
12: 9 542 560 - 9 604 941	−2.469	0.0136	0.110	12: 9 525 125 - 9 604 954
12: 9 604 942 - 9 606 831	−2.508	0.0121	0.104	12: 9 525 125 - 9 604 954

Genome-wide CNV calling in further 281 *obesity trios*

Statistically significant findings from the first family-based GWAS discovery sample of 424 obesity trios were followed up by investigating a second independent family-based sample of further 281 obesity trios: Concerning this second sample, a total of 33 372 CNVs were identified in the 843 individuals (offspring: n = 11 177, parents: n = 22 195). A subset of 29 112 CNVs were located in the 244 pre-specified CNVRs (offspring: n = 9 774, parents: n = 19 338).

As has been observed for the initially considered sample, Mendelian inconsistency rates for MCMR-based CNVs of the second sample were statistically significantly lower than for CNVs being estimated by use of the Default procedure ($p = 2.58 \times 10^{-5}$, Table 6.8). The reduced Mendelian inconsistency rates across MCMR-based CNVs become even more evident when both samples were considered in a combined manner ($p = 5.41 \times 10^{-13}$, Table 6.8).

Genome-wide association testing in further 281 *obesity trios*

In the genome-wide association testing step of the second sample, a total of 3 718 FBATs were performed at all unique CNV's start and end sites in the 244 pre-specified CNVRs. 381 sites, reflecting 47 CNVRs, showed significance with lfdr values below 0.2 (Figure 6.15).

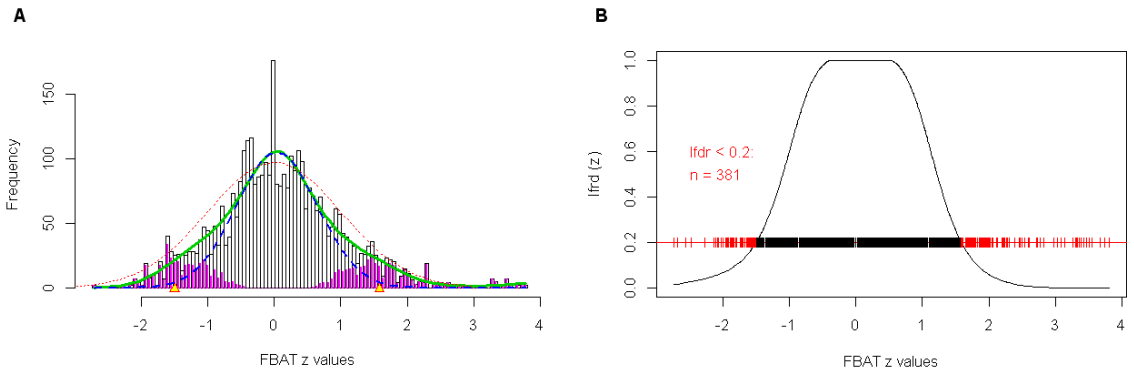

Figure 6.15: Histogram and lfdr curve of CNV FBAT z-values for the genome-wide analysis of 281 obesity trios at 3 718 unique CNVs start and end sites in 244 CNVRs. See Figure 6.14 for a detailed description. The empirical null density is $\mathcal{N}(0.042, 0.610^2)$.

Three of the seven CNVRs with significance in 424 obesity trios also indicated evidence for an association with obesity in the additional 281 obesity trios (Table 6.10). Of note, only one CNVR (at chr 11: 55 130 596 - 55 210 165) showed statistically significant and continuously directionally consistent effects in both trio samples.

Table 6.10: Results for eleven sub-CNVRs reflecting 25 CNV's start and end sites at three CNVRs with lfdr values < 0.2 in genome-wide FBATs accounting for 281 as well as for further 424 obesity trios. (in chromosomal order)

Chr: Position [hg18]	FBAT z-value	FBAT p-value	lfdr	in CNVR [chr: bp, hg18]
3: 131 288 927 - 131 289 688	2.266	0.0234	0.026	3: 131 245 537 - 131 290 979
10: 58 186 381 - 58 186 526	2.263	0.0237	0.026	10: 58 186 369 - 58 196 856
10: 58 186 527 - 58 196 843	2.380	0.0173	0.017	10: 58 186 369 - 58 196 856
11: 55 133 074 - 55 134 453	−1.871	0.0614	0.085	11: 55 130 596 - 55 210 165
11: 55 134 454 - 55 142 243	−1.824	0.0681	0.094	11: 55 130 596 - 55 210 165
11: 55 142 244 - 55 142 244	−1.674	0.0941	0.131	11: 55 130 596 - 55 210 165
11: 55 142 245 - 55 143 361	−1.729	0.0837	0.116	11: 55 130 596 - 55 210 165
11: 55 149 884 - 55 178 915	−1.611	0.1072	0.152	11: 55 130 596 - 55 210 165
11: 55 178 916 - 55 187 640	−1.737	0.0824	0.114	11: 55 1305 96 - 55 210 165
11: 55 209 586 - 55 210 152	1.742	0.0816	0.127	11: 55 1305 96 - 55 210 165
11: 55 210 153 - 55 217 258	2.000	0.0455	0.059	11: 55 1305 96 - 55 210 165

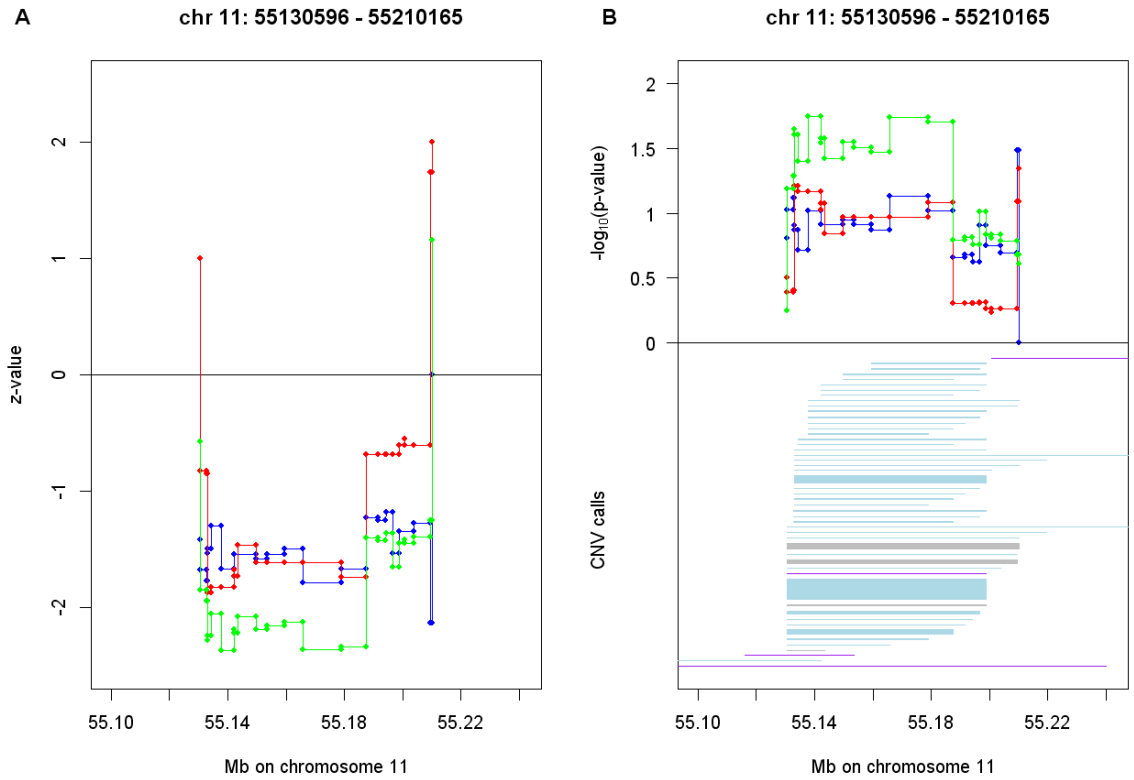

Figure 6.16: CNV FBAT z-values (Panel A), p-values and CNV calls (Panel B) for two obesity trio samples at CNVR on chr 11q11. blue: 424 trios, red: 281 trios, green: all 705 trios. CNV calls of all 705 trios are presented as vertical bars and the bar thickness reflects the CNV frequency. gray: homozygous deletion, blue: heterozygous deletion, purple: duplication.

This CNVR at chr 11q11 was already identified to be associated with the trait obesity in application of strategy PS1 (see chapter 5.2). Details on FBAT z-values, p-values and CNV calls in this most promising CNVR at chr 11q11 are presented in Figure 6.16 for both obesity trio samples as well as for the combined sample of all 705 nuclear families.

CNV validation

Finally, the array-based observation of deletions in the described CNVR at chr 11q11 being associated with obesity in both family-based obesity samples was technically validated by qPCR analyses (see Table 6.11). The array-derived association effect was consistently re-observed for qPCR-based CNVs (FBAT p-value = 0.023). Moreover, the majority of MCMR-based copy number states could by validated as true findings (false positives = 0.35%, false negatives = 11.55%).

Table 6.11: Locus-specific false positive and false negative estimates for CNV calls of 705 obesity trios at chromosome 11 : 55 130 596 − 55 210 165 (hg18).

	qPCR validated CNVs	CNV calls of MCMR calling[a]	CNV calls of Default calling[a]
# of homozygous deletions (cn=0)	169	169 [168]	169 [168]
# of heterozygous deletions (cn=1)	830	687 [687]	342 [342]
# of copy number neutrals (cn=2)	1 077	1 211 [1 072]	1 098 [614]
# of heterozygous duplications (cn=3)	0	3 [0]	461 [0]
# of complex CNVs	-	6	6
# of CNVs failed to be assigned	39	39	39
% of false positive CNV calls[b]	-	0.35%	47.48%
% of false negative CNV calls[c]	-	11.55%	44.13%
p-value, FBAT, two-sided	0.0231	0.0295	0.5114
z-value, FBAT	−2.2711	−2.1772	−0.6566

[a] The number of CNV calls with identical individual called copy number state (cn) and qPCR result is given in brackets.
[b] False positives: the percentage of individuals who were called to have a CNV, but were confirmed to be copy number neutral by qPCR analyses.
[c] False negatives: the percentage of individuals who were called to be copy number neutral, but were confirmed to be copy number variable by qPCR analyses.

Contrarily, the association effect at chr 11q11 would not have been detected, if a Default calling procedure was applied, that is if median reference intensity values

were used (Default FBAT p-value = 0.5114 in 705 obesity trios). The main reason for this fundamental weakness of strategy S2 might be the high rates of misclassified CNV calls (false positives = 47.48%, false negatives = 44.13%).

6.3.4 Discussion

In this chapter, the data of two family-based samples for genome-wide association studies (GWASs) on early-onset extreme obesity were analysed in application of strategy PS2. Directionally consistent and genome-wide statistically significant association was exclusively observed for deletions in a region on chromosome 11q11. This association effect could be strengthened by validation data of qPCR experiments. In chapter 3.6, the analogous application of strategy S2 to the same data example did not reveal evidence for an association of any CNV with the trait obesity. It was demonstrated that high false positive and false negative rates of CNV calls identified in use of strategy S2 caused invalid association test results especially for the CNVR at chr 11q11. Consequently, a superiority of strategy PS2 over strategy S2 could be shown for the analysed obesity data example. As mentioned previously, due to the complexity of any genetic genome-wide data set, this observation does not allow a general evaluative conclusion regarding power or type 1 error levels of strategy S2 and strategy PS2.

QPCR experiments were only performed for the most promising CNV region on chr 11q11, which probably reflects a realistic procedure for practical use with regard to monetary cost considerations. Thus, the genome-wide effect of using MCMR instead of Default median reference intensity values for CNV calling cannot comprehensively be assessed at a genome-wide level. However, the respective qPCR validation data impressively demonstrate how the choice of the global intensity reference set may dramatically influence association test results for common CNVs at a locus-specific level. In addition to that, estimates on genome-wide Mendelian inconsistency rates were consulted to evaluate the validity of MCMR and median-based CNV calls at a genome-wide level. With regard to previous reports, most CNV events are rather derived from inheritance than from new mutation events (McCarroll et al., 2008; Redon et al., 2006). As a consequence, low Mendelian inconsistency rates indicate a high quality of CNV calling. For the two obesity data sets, genome-wide Mendelian inconsistency rates of CNV calls were statistically significantly reduced in use of MCMR instead of median reference intensity values. Consequently, for the presented data example CNV calling validity was improved in use of strategy PS2 relative to strategy S2. Of note, Mendelian inconsistency rates are likewise far from the optimal zero value for MCMR based CNVs, which potentially reflects the

limited effect of sophisticated reference models, such as the proposed MCMR model, on validity of CNV calling. This limitation might be caused by several other biases, such as raw data quality and robustness, which are not addressed by the choice of the intensity reference model.

In sum, strategy PS2 turned to be out to be more useful as strategy S2 for the genome-wide CNV analysis of the obesity data set. However, this result should not be overestimated because it is based on only one observation. Additionally, the moderate reduction of genome-wide Mendelian inconsistency rates relative to strategy S2 rather suggests that a limited number of CNV calls is actually affected by a more precise genotype assignment. Contrarily, this small number may suffice to enormously change genome-wide results. Indeed, any single truly associated and correctly genotyped CNV region offers the potential to end up with genome-wide significance. For later practical use and with regard to a reduction of computation time, it might be particularly appropriate to follow selected genomic regions up in application of more than one reference CNV calling model. In particular, visual inspection of raw intensity data may be useful to identify those respective regions.

7 General Discussion

In this thesis I explored several whole genome-wide analysis strategies for raw CNV data using the two most widespread methods from recent research as well as two modified techniques that aim to overcome weaknesses of currently used procedures. Initially, I presented the two main existing approaches for CNV association analyses based on genome-wide SNP genotyping data that primarily differ in the extent to which individual CNV genotypes are assessed prior to genome-wide association testing. In application of the existing strategy, which completely skips CNV calling while instead focussing on raw continuous intensity CNV measurements in the association testing step, I studied the role of common CNVs in severe early-onset obesity. In use of this strategy, no CNVs could be identified as being causal for that phenotype. Using the other existing strategy, which bases association testing on discrete CNV genotypes being obtained from previous application of CNV calling software tools, I investigated the role of both, common CNVs in severe early-onset obesity and rare CNVs in childhood ADHD. Again, no evidence for any association between CNVs and the trait obesity were detected. Contrarily, this second standard CNV analysis strategy turned out to be useful to discover rare CNVs at the *PARK2* locus as being statistically genome-wide significantly associated with ADHD in children. Secondly, I presented two modified approaches for genome-wide CNV association analyses, which are both motivated by previous concerns with regard to the two above mentioned commonly applied strategies (Ionita-Laza et al., 2008; Zhang et al., 2011). On the one hand, Ionita-Laza et al. (2008) proposed that a refinement on the marker selection before the association testing of raw hybridization intensity CNV measurements might be useful to alleviate the multiple testing issue, and thus to simplify the detection of underlying CNV association effects. Here, this proposal was taken up and the corresponding re-analysis of potential CNV - obesity association effects led to the identification of CNVs at chromosome 10q11.22 and at chromosome 11q11 being positively associated with early-onset obesity. One of these two findings is strengthened by a previous study, which concordantly reports on CNVs at chromosome 10q11.22 being associated with BMI in a sample of 597 elderly Chinese Han subjects (Sha et al., 2009). The other finding related to CNVs

at chromosome 11q11 remains to be proven in large-scale meta-analyses. On the other hand, several authors, such as Zhang et al. (2011), realized that especially array-based CNV calling of common CNVs is greatly influenced and to some extent distorted by the use of $\log_2$ ratios that were calculated with respect to the sample's mean or median hybridization intensity values, which is common practice in most currently available CNV calling software suites. Alternatively, I propose to estimate copy number neutral reference hybridization intensity values for each probe set in a Gaussian mixture model framework prior to the determination of $\log_2$ ratios and subsequent CNV calling and association testing. In an isolated application to CNV calling, this approach showed slightly better quality of CNV calls compared to standard derived CNV genotypes. As a demonstration and to allow a comparison across all presented strategical approaches, I applied the lastly suggested strategy to re-re-analyse associations of CNVs on obesity. Not only were again CNVs at chromosome 11q11 found to be statistically genome-wide significantly associated with obesity, but also it became apparent how the misclassification of eventually only one single CNV region can dramatically change genome-wide association test results. The latter aspect was demonstrated with respect to qPCR experiment data for the chromosome 11q11 region, which was nearly consistent with MCMR based CNV calls but substantially differed from respective standard derived CNVs. To my knowledge, this is the first time that such a variety of alternative whole genome CNV analysis strategies has been investigated and comparably been applied to real data examples.

Genome-wide CNV analyses strategies based on raw genotyping array data are complex procedures including several partial steps, such as data pre-processing, CNV identification, association testing and validation experiments. It is worth considering overall limitations of the conclusions towards advantages and disadvantages across the presented strategies, which have already been discussed in a comparative sense at the end of each data example chapter. In order to judgmentally embrace the entire spectrum of different CNV analyses strategies, a much larger variety of different approaches addressing each single aspect of the whole analyses pipeline is needed. Here, I explicitly concentrated on two characteristics, the marker selection for testing raw CNV measurements and improving CNV calling quality by alternate reference models, which leaves a lot of room for investigating further facets. Many of the remaining aspects have already been evaluated in an isolated way, that is in picking out and concentrating on a certain sub-step of the whole genome-wide CNV analysis. For instance, the effect of different normalization methods in data pre-processing (Bolstad et al., 2003) or the optimal choice of an appropriate soft-

ware algorithm for CNV detection (Winchester et al., 2009; Koike et al., 2011; Zhang et al., 2011; Dellinger et al., 2010) were investigated in much detail. When separately evaluating selected sub-steps of a whole genome-wide CNV analysis strategy, the actual effects on genome-wide association test results are left to speculation. However, any improvement on each single pipeline part offers the potential to substantially increase validity of the whole analysis. Contrarily, completely implemented pipelines focussing on the modification of single aspects are presented here.

Due to its availability, all analyses were restricted to one selected type of raw CNV data, namely to those being collected from the Affymetrix 6.0 SNP genotyping arrays. As outlined in chapter 3, each presented strategy can with slight array-specific adaptations likewise be applied to alternatively derived raw CNV data. However, especially the observed superior performance of strategy PS2 over strategy S2 might in parts be affected by the characteristics of Affymetrix 6.0 data, which is known to require much more robust algorithms than those of Illumina SNP arrays or CGH arrays (Koike et al., 2011).

Moreover, practical investigations of strategies S2 and PS2 were limited to the use of the PennCNV software for calling CNVs with standard and MCMR reference intensity values. Apart from the consequence of suffering from a restricted transferability to other CNV calling software, this design guarantees a valid comparison between both approaches. Of note, most commonly used CNV detection software tools start their estimating calculations with $\log_2$ ratios, and can thus likewise be applied in use of several alternate reference intensity models.

Finally, the evaluation of the four presented CNV analysis strategies was performed in application to real data examples for seemingly healthy HapMap and replicate samples as well as for the phenotypes obesity and ADHD. Although the investigation of HapMap and replicate CNV calls and additionally the repeated whole genome-wide analyses of one and the same obesity data set admitted meaningful statements concerning the superiority of the proposed strategies over standard strategies, these results have to be handled with caution and cannot be understood as general conclusions or recommendations. Below the line, it has been shown that targeted modification of standard CNV analyses approaches may reveal useful association results. In this respect, it is important to remember that in particular the proposed strategy PS1 did not only end up with a new CNV - obesity association finding at chromosome 11q11, which still has to prove its validity in future large scale-meta analyses. Most importantly, the previous finding of a CNV - BMI association at chromosome 10q11.22 was re-identified (Sha et al., 2009). However, all this does not allow any general conclusion regarding power or type 1 error levels

for which, as already mentioned in each single previous discussion, target-orientated and informative simulation studies have to be conducted. Nevertheless, since an appropriate simulation design might realistically only account for selected aspects of a whole genome-wide CNV analysis, a broad range of such simulation studies is needed to allow recommendations for a genome-wide analyses that starts with the extraction of raw hybridization intensities and ends up with association p-values. Last but not least, the determination of false positive and false negative rates for HapMap and replicate CNV calls in chapter 6.2 has provided an insight into the difficulties and complexity in finding an appropriate gold standard CNV set. For example, out of a total of 681 CNV segments that were reported in at least one of four publications (Tuzun et al., 2005; Korbel et al., 2007; Redon et al., 2006; Kidd et al., 2008) for sample NA15510, only four CNVs (= 0.59%) were consistently found in all four reports. These and other CNV-specific challenges, such as the complex correlation structure of involved CNV markers, the lack of full knowledge on causes for CNV occurrence, their inheritance mechanism or their genome-wide interrelationship, were the reasons to restrict strategy comparisons to real data applications.

With regard to recent advances in next-generation sequencing techniques that will provide highly reliable CNV information for thousands of individuals at moderate fee and time, many of the addressed difficulties of SNP array based CNV association analyses will be eliminated in future. Until then, however, there might still be great interest in using the variety of collected genome-wide SNP genotyping data by the largest possible amount. Now, data from whole genome-wide SNP association studies can be used for dual purposes, SNP and CNV analyses. Consequently, many genome-wide SNP genotyping data sets that were so far only investigated towards SNP associations will be remembered with respect to CNV re-analyses. Towards an implementation of any of the presented strategies, I would recommend using a second approach maybe only to selected genomic regions on a single data set to generate most informative results. It is also important, to let as many CNV results as possible be validated by independent techniques, such as qPCR experiments. Against this background and keeping in mind that the validity of genome-wide results are affected by the interaction of locus-specific reliability, a collection of several validation experiments accounting for a variety of selected genomic loci on a random basis might be an optimal approach.

References

Affymetrix. Whitepaper: CNAT 4.0. Copy Number and Loss of Heterozygosity Estimation Algorithms for the GeneChip® Human Mapping 10/50/100/250/500K Array Set. *http://media.affymetrix.com/support/technical/whitepapers/cnat_4_algorithm_whitepaper.pdf*, 2007.

Affymetrix. Whitepaper: Genome-Wide Human SNP Array 6.0. *http://media.affymetrix.com/support/technical/datasheets/genomewide_snp6_datasheet.pdf*, 2009.

American Psychiatric Association. Diagnostic and Statistical Manual of Mental Diseases (DSM-IV), 4th ed. *American Psychiatric Publishing; Washington, DC*, 1994.

C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and M. E. Hurles. A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.*, 40(10):1245–1252, 2008.

H. Bengtsson, J. Bullard, and K. D. Hansen. affxparser: Affymetrix file parsing sdk. *R Manual*, R package version 1.14.2, 2008a.

H. Bengtsson, R. Irizarry, B. Carvalho, and T. P. Speed. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics*, 24(6):759–767, 2008b.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 57:289–300, 1995.

A. v. Berg, U. Kramer, E. Link, C. Bollrath, J. Heinrich, I. Brockow, S. Koletzko, A. Grubl, B. Filipiak-Pittroff, H. E. Wichmann, et al. Impact of early feeding on childhood eczema: development after nutritional intervention compared with the natural course - the GINIplus study up to the age of 6 years. *Clin. Exp. Allergy*, 40(4):627–636, 2010.

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

C. Brasch-Andersen, L. Christiansen, Q. Tan, A. Haagerup, J. Vestbo, and T. A. Kruse. Possible gene dosage effect of glutathione-S-transferases on atopic asthma: using real-time PCR for quantification of GSTM1 and GSTT1 gene copy numbers. *Hum. Mutat.*, 24(3):208–214, 2004.

C. E. Bruder, A. Piotrowski, A. A. Gijsbers, R. Andersson, S. Erickson, T. Diaz de Stahl, U. Menzel, J. Sandgren, D. von Tell, A. Poplawski, et al. Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.*, 82(3):763–771, 2008.

H. S. Chai, T. M. Therneau, K. R. Bailey, and J. P. Kocher. Spatial normalization improves the quality of genotype calling for Affymetrix SNP 6.0 arrays. *BMC Bioinformatics*, 11:356, 2010.

S. Colella, C. Yau, J. M. Taylor, G. Mirza, H. Butler, P. Clouston, A. S. Bassett, A. Seller, C. C. Holmes, and J. Ragoussis. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, 35(6):2013–2025, 2007.

D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard. A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, 38(1):75–81, 2006.

D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010.

1000 Genomes Project Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319): 1061–1073, 2010a.

1000 Genomes Project Consortium, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. dePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

Wellcome Trust Case Control Consortium, N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–20, 2010b.

G. M. Cooper, T. Zerr, J. M. Kidd, E. E. Eichler, and D. A. Nickerson. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat. Genet.*, 40(10):1199–1203, 2008.

N. Day, A. Hemmaplardh, R. E. Thurman, J. A. Stamatoyannopoulos, and W. S. Noble. Unsupervised segmentation of continuous genomic data. *Bioinformatics*, 23(11):1424–1426, 2007.

A. E. Dellinger, S. M. Saw, L. K. Goh, M. Seielstad, T. L. Young, and Y. J. Li. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.*, 38(9):e105, 2010.

B. Efron. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association*, 99:96–104, 2004.

B. Efron. Correlation and Large-Scale Simultaneous Significance Testing. *J. Amer. Statist. Assoc.*, 102:93–103, 2007a.

B. Efron. Size, power and false discovery rates. *The Annals of Statistics*, 35(4): 1351–1377, 2007b.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160, 2001.

B. Efron, B. B. Turnbull, and B. Narasimhan. locfdr: Computes local false discovery rates. *R Manual*, R package version 1.1-7, http://CRAN.R-project.org/package=locfdr, 2011.

J. Elia, X. Gai, H. M. Xie, J. C. Perin, E. Geiger, J. T. Glessner, M. D'arcy, R. deBerardinis, E. Frackelton, C. Kim, et al. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. *Mol. Psychiatry*, 15(6):637–646, 2010.

S. Eyheramendy, C. Gieger, M. Laan, T. Illig, T. Meitinger, and E. Wichmann. Effect of genome-wide simultaneous hypotheses tests on the discovery rate. *Int J Mol Epidemiol Genet*, 2(2):163–177, 2011.

L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nat. Rev. Genet.*, 7(2):85–97, 2006a.

L. Feuk, C. R. Marshall, R. F. Wintle, and S. W. Scherer. Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.*, 15 Spec No 1:57–66, 2006b.

C. Fraley, A. Raftery, T. B. Murphy, and L. Scrucca. MCLUST Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation. *Technical Report*, Department of Statistics, University of Washington (597), 2012.

L. Franke, C. G. de Kovel, Y. S. Aulchenko, G. Trynka, A. Zhernakova, K. A. Hunt, H. M. Blauw, L. H. van den Berg, R. Ophoff, P. Deloukas, et al. Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am. J. Hum. Genet.*, 82(6):1316–1333, 2008.

C. M. Freitag, L. A. Rohde, T. Lempp, and M. Romanos. Phenotypic and measurement influences on heritability estimates in childhood ADHD. *Eur Child Adolesc Psychiatry*, 19(3):311–323, 2010.

J. Hardy and A. Singleton. Genomewide association studies and human disease. *N. Engl. J. Med.*, 360(17):1759–1768, 2009.

J. Hebebrand, H. Heseker, G. W. Himmelmann, H. Schäfer, and Remschmidt H. Altersperzentilen für den Body Mass Index aus Daten der Nationalen Verzehrsstudie einschließlich einer Übersicht zu relevanten Einflußfaktoren. *Aktuel. Ernährungsmed.*, 19:259–265, 1994.

A. Hinney, A. Scherag, I. Jarick, O. Albayrak, C. Putter, S. Pechlivanis, M. R. Dauvermann, S. Beck, H. Weber, S. Scherag, et al. Genome-wide association study in German patients with attention deficit/hyperactivity disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 156B(8):888–897, 2011.

J. Huang, W. Wei, J. Zhang, G. Liu, G. R. Bignell, M. R. Stratton, P. A. Futreal, R. Wooster, K. W. Jones, and M. H. Shapero. Whole genome DNA copy number changes identified by high density oligonucleotide arrays. *Hum. Genomics*, 1(4): 287–299, 2004.

Illumina. Technical Note: DNA Copy Number and Loss of Heterozygosity Analysis Algorithms. *http://www.illumina.com/Documents/products/technotes/technote_cnv_algorithms.pdf*, 2010.

K. Inoue and J. R. Lupski. Molecular mechanisms for genomic disorders. *Annu Rev Genomics Hum Genet*, 3:199–242, 2002.

I. Ionita-Laza, G. H. Perry, B. A. Raby, B. Klanderman, C. Lee, N. M. Laird, S. T. Weiss, and C. Lange. On the analysis of copy-number variations in genome-wide

association studies: a translation of the family-based association test. *Genet. Epidemiol.*, 32(3):273–284, 2008.

I. Ionita-Laza, A. J. Rogers, C. Lange, B. A. Raby, and C. Lee. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*, 93(1):22–26, 2009.

I. Jarick, C. I. Vogel, S. Scherag, H. Schäfer, J. Hebebrand, A. Hinney, and A. Scherag. Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum. Mol. Genet.*, 20(4):840–852, 2011.

I. Jarick, A. L. Volckmar, C. Putter, S. Pechlivanis, T. T. Nguyen, M. R. Dauvermann, S. Beck, O. Albayrak, S. Scherag, S. Gilsbach, et al. Genome-wide analysis of rare copy number variations reveals PARK2 as a candidate gene for attention-deficit/hyperactivity disorder. *Mol. Psychiatry*, doi: 10.1038/mp.2012.161, 2012.

J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008.

P. M. Kim, H. Y. Lam, A. E. Urban, J. O. Korbel, J. Affourtit, F. Grubert, X. Chen, S. Weissman, M. Snyder, and M. B. Gerstein. Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res.*, 18(12):1865–1874, 2008.

A. Koike, N. Nishida, D. Yamashita, and K. Tokunaga. Comparative analysis of copy number variation detection methods and database construction. *BMC Genet.*, 12: 29, 2011.

D. Komura, F. Shen, S. Ishikawa, K. R. Fitch, W. Chen, J. Zhang, G. Liu, S. Ihara, H. Nakamura, M. E. Hurles, et al. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.*, 16(12): 1575–1584, 2006.

J. O. Korbel, A. E. Urban, J. P. Affourtit, B. Godwin, F. Grubert, J. F. Simons, P. M. Kim, D. Palejev, N. J. Carriero, L. Du, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007.

J. M. Korn, F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J. Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly, and D. Altshuler. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.*, 40(10):1253–1260, 2008.

M. Krawczak, S. Nikolaus, H. von Eberstein, P. J. Croucher, N. E. El Mokhtari, and S. Schreiber. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet*, 9 (1):55–61, 2006.

N. M. Laird and C. Lange. Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.*, 7(5):385–394, 2006.

C. Lange, D. L. DeMeo, and N. M. Laird. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am. J. Hum. Genet.*, 71(6):1330–1341, 2002.

K. P. Lesch, S. Selch, T. J. Renner, C. Jacob, T. T. Nguyen, T. Hahn, M. Romanos, S. Walitza, S. Shoichet, A. Dempfle, et al. Genome-wide copy number variation analysis in attention-deficit/hyperactivity disorder: association with neuropeptide Y gene dosage in an extended pedigree. *Mol. Psychiatry*, 16(5):491–503, 2011.

S. Levy, G. Sutton, P. C. Ng, L. Feuk, A. L. Halpern, B. P. Walenz, N. Axelrod, J. Huang, E. F. Kirkness, G. Denisov, et al. The diploid genome sequence of an individual human. *PLoS Biol.*, 5(10):e254, 2007.

C. Li, R. Beroukhim, B. A. Weir, W. Winckler, L. A. Garraway, W. R. Sellers, and M. Meyerson. Major copy proportion analysis of tumor samples using SNP arrays. *BMC Bioinformatics*, 9:204, 2008.

D. Lin, I. B. Gibson, J. M. Moore, P. C. Thornton, S. M. Leal, and P. J. Hastings. Global chromosomal structural instability in a subpopulation of starving Escherichia coli cells. *PLoS Genet.*, 7(8):e1002223, 2011.

A. C. Lionel, J. Crosbie, N. Barbosa, T. Goodale, B. Thiruvahindrapuram, J. Rickaby, M. Gazzellone, A. R. Carson, J. L. Howe, Z. Wang, et al. Rare copy number variation discovery and cross-disorder comparisons identify risk genes for ADHD. *Sci Transl Med*, 3(95):95ra75, 2011.

D. P. Locke, A. J. Sharp, S. A. McCarroll, S. D. McGrath, T. L. Newman, Z. Cheng, S. Schwartz, D. G. Albertson, D. Pinkel, D. M. Altshuler, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.*, 79(2):275–290, 2006.

K. L. Lunetta, S. V. Faraone, J. Biederman, and N. M. Laird. Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. *Am. J. Hum. Genet.*, 66(2):605–614, 2000.

J. R. Lupski. Genomic rearrangements and sporadic disease. *Nat. Genet.*, 39(7 Suppl):S43–47, 2007.

H. H. Maes, M. C. Neale, and L. J. Eaves. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.*, 27(4):325–351, Jul 1997.

T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

J. C. Marioni, N. P. Thorne, A. Valsesia, T. Fitzgerald, R. Redon, H. Fiegler, T. D. Andrews, B. E. Stranger, A. G. Lynch, E. T. Dermitzakis, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol.*, 8(10):R228, 2007.

O. Mayo. The rise and fall of the common disease-common variant (CD-CV) hypothesis: how the sickle cell disease paradigm led us all astray (or did it?). *Twin Res Hum Genet*, 10(6):793–804, 2007.

S. A. McCarroll. Copy number variation and human genome maps. *Nat. Genet.*, 42 (5):365–366, 2010.

S. A. McCarroll and D. M. Altshuler. Copy-number variation and association studies of human disease. *Nat. Genet.*, 39(7 Suppl):37–42, 2007.

S. A. McCarroll, F. G. Kuruvilla, J. M. Korn, S. Cawley, J. Nemesh, A. Wysoker, M. H. Shapero, P. I. de Bakker, J. B. Maller, A. Kirby, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat. Genet.*, 40(10):1166–1174, 2008.

R. McGinnis, S. Shifman, and A. Darvasi. Power and efficiency of the TDT and case-control design for association scans. *Behav. Genet.*, 32(2):135–144, 2002.

D. A. Oldridge, S. Banerjee, S. R. Setlur, A. Sboner, and F. Demichelis. Optimizing copy number variation analysis using genome-wide short sequence oligonucleotide arrays. *Nucleic Acids Res.*, 38(10):3275–3286, 2010.

A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

J. Ott. Statistical properties of the haplotype relative risk. *Genet. Epidemiol.*, 6(1): 127–130, 1989.

H. Park, J. I. Kim, Y. S. Ju, O. Gokcumen, R. E. Mills, S. Kim, S. Lee, D. Suh, D. Hong, H. P. Kang, et al. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat. Genet.*, 42(5):400–405, 2010.

D. A. Peiffer, J. M. Le, F. J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. A. Shaw, J. Belmont, et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, 16 (9):1136–1148, 2006.

G. H. Perry, A. Ben-Dor, A. Tsalenko, N. Sampas, L. Rodriguez-Revenga, C. W. Tran, A. Scheffer, I. Steinfeld, P. Tsang, N. A. Yamada, et al. The fine-scale and complex architecture of human copy-number variation. *Am. J. Hum. Genet.*, 82 (3):685–695, 2008.

A. Piotrowski, C. E. Bruder, R. Andersson, T. Diaz de Stahl, U. Menzel, J. Sandgren, A. Poplawski, D. von Tell, C. Crasto, A. Bogdan, et al. Somatic mosaicism for copy number variation in differentiated human tissues. *Hum. Mutat.*, 29(9):1118–1124, 2008.

R. Pique-Regi, J. Monso-Varona, A. Ortega, R. C. Seeger, T. J. Triche, and S. Asgharzadeh. Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics*, 24(3):309–318, 2008.

R. Pique-Regi, A. Ortega, and S. Asgharzadeh. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. *Bioinformatics*, 25(10):1223–1230, 2009.

G. Polanczyk, M. S. de Lima, B. L. Horta, J. Biederman, and L. A. Rohde. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. *Am J Psychiatry*, 164(6):942–948, 2007.

R. L. Prentice and R. Pyke. Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411, 1979.

K. Prüfer, K. Munch, I. Hellmann, K. Akagi, J. R. Miller, B. Walenz, S. Koren, G. Sutton, C. Kodira, R. Winer, et al. The bonobo genome compared with the chimpanzee and human genomes. *Nature*, 486(7404):527–531, 2012.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, 81(3):559–575, 2007.

J. Quackenbush. Microarray data normalization and transformation. *Nat. Genet.*, 32 Suppl:496–501, 2002.

D. Rabinowitz and N. Laird. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.*, 50(4):211–223, 2000.

R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, et al. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

G. Rigaill, P. Hupe, A. Almeida, P. La Rosa, J. P. Meyniel, C. Decraene, and E. Barillot. ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays. *Bioinformatics*, 24(6):768–774, 2008.

A. Schmermund, S. Mohlenkamp, A. Stang, D. Gronemeyer, R. Seibel, H. Hirche, K. Mann, W. Siffert, K. Lauterbach, J. Siegrist, et al. Assessment of clinically silent atherosclerotic disease and established and novel risk factors for predicting myocardial infarction and cardiac death in healthy middle-aged subjects: rationale and design of the Heinz Nixdorf RECALL Study. Risk Factors, Evaluation of Coronary Calcium and Lifestyle. *Am. Heart J.*, 144(2):212–218, 2002.

B. Y. Sha, T. L. Yang, L. J. Zhao, X. D. Chen, Y. Guo, Y. Chen, F. Pan, Z. X. Zhang, S. S. Dong, X. H. Xu, and H. W. Deng. Genome-wide association study suggested copy number variation may be associated with body mass index in the Chinese population. *J. Hum. Genet.*, 54(4):199–202, 2009.

T. H. Shaikh, X. Gai, J. C. Perin, J. T. Glessner, H. Xie, K. Murphy, R. O'Hara, T. Casalunovo, L. K. Conlin, M. D'Arcy, et al. High-resolution mapping and

analysis of copy number variations in the human genome: a data resource for clinical and research applications. *Genome Res.*, 19(9):1682–1690, 2009.

Y. Shen and B. L. Wu. Microarray-based genomic DNA profiling technologies in clinical molecular diagnostics. *Clin. Chem.*, 55(4):659–669, 2009.

E. K. Speliotes, C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, A. U. Jackson, H. Lango Allen, C. M. Lindgren, J. Luan, R. Magi, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, 42(11):937–948, 2010.

A. J. Stunkard, T. T. Foch, and Z. Hrubec. A twin study of human obesity. *JAMA*, 256(1):51–54, Jul 1986a.

A. J. Stunkard, T. I. Sorensen, C. Hanis, T. W. Teasdale, R. Chakraborty, W. J. Schull, and F. Schulsinger. An adoption study of human obesity. *N. Engl. J. Med.*, 314(4):193–198, Jan 1986b.

I. Subirana, R. Diaz-Uriarte, G. Lucas, and J. R. Gonzalez. CNVassoc: Association analysis of CNV data using R. *BMC Med Genomics*, 4:47, 2011.

The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437(7055):69–87, 2005.

M. Turula, J. Kaprio, A. Rissanen, and M. Koskenvuo. Body weight in the Finnish Twin Cohort. *Diabetes Res. Clin. Pract.*, 10 Suppl 1:S33–36, 1990.

E. Tuzun, A. J. Sharp, J. A. Bailey, R. Kaul, V. A. Morrison, L. M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, et al. Fine-scale structural variation of the human genome. *Nat. Genet.*, 37(7):727–732, 2005.

A. J. Walley, J. E. Asher, and P. Froguel. The genetic contribution to non-syndromic human obesity. *Nat. Rev. Genet.*, 10(7):431–442, Jul 2009.

T. Walsh, J. M. McClellan, S. E. McCarthy, A. M. Addington, S. B. Pierce, G. M. Cooper, A. S. Nord, M. Kusenda, D. Malhotra, A. Bhandari, et al. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, 320(5875):539–543, 2008.

K. Wang, M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan. PennCNV: an integrated hidden Markov model designed for high-

resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.*, 17(11):1665–1674, 2007.

J. Wardle, S. Carnell, C. M. Haworth, and R. Plomin. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *Am. J. Clin. Nutr.*, 87(2):398–404, Feb 2008.

H. E. Wichmann, C. Gieger, and T. Illig. KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*, 67 Suppl 1:26–30, 2005.

C. J. Willer, E. K. Speliotes, R. J. Loos, S. Li, C. M. Lindgren, I. M. Heid, S. I. Berndt, A. L. Elliott, A. U. Jackson, C. Lamina, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.*, 41(1):25–34, 2009.

N. M. Williams, I. Zaharieva, A. Martin, K. Langley, K. Mantripragada, R. Fossdal, H. Stefansson, K. Stefansson, P. Magnusson, O. O. Gudmundsson, et al. Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis. *Lancet*, 376(9750):1401–1408, 2010.

L. Winchester, C. Yau, and J. Ragoussis. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic*, 8(5):353–366, 2009.

D. Zhang, Y. Qian, N. Akula, N. Alliey-Rodriguez, J. Tang, The Bipolar Genome Study, E. S. Gerschon, and C. Liu. Accuracy of CNV Detection from GWAS Data. *PLoS ONE*, 6(1):e14511, 2011.

A. Zutavern, I. Brockow, B. Schaaf, G. Bolte, A. von Berg, U. Diez, M. Borte, O. Herbarth, H. E. Wichmann, and J. Heinrich. Timing of solid food introduction in relation to atopic dermatitis and atopic sensitization: results from a prospective birth cohort study. *Pediatrics*, 117(2):401–411, 2006.

Tabellarischer Lebenslauf

Persönliche Daten

Ivonne Jarick
Cappeler Str. 2
35039 Marburg

Tel.: 06421 18 67 870
Email: ivonne.jarick@gmx.de

Geb. am 17.04.1981 in Cottbus
ledig, keine Kinder, deutsche Nationalität

Beruflicher Werdegang

seit 06/2008	Wissenschaftliche Mitarbeiterin am Institut für Medizinische Biometrie und Epidemiologie (IMBE), Philipps-Universität Marburg

Hochschulstudium

10/2002 - 05/2008	Diplom-Studium der Mathematik mit Nebenfach VWL, Ruprecht-Karls-Universität Heidelberg
	Diplomarbeit: "Nonparametric Estimation in the Time Domain for Simple Stationary Point Processes" (Prof. Rannacher)
	Diplom in Mathematik (Note: sehr gut)

Berufsausbildung

09/2000 - 07/2002	Kaufmännische Ausbildung zu Industriekauffrau (IHK)
	bei Universal Music Germany, Hamburg im Rahmen des dualen Hamburger Modells mit der Wirtschaftsakademie Hamburg (WAH), Hamburg

Schulausbildung

08/1993 - 06/2000	Abitur am Heinrich-Heine Gymnasium, Cottbus
09/1987 - 06/1993	1. Grundschule, Cottbus

Eigene Publikationen

N. Knoll, I. Jarick, A. L. Volckmar, M. Klingenspor, T. Illig, H. Grallert, C. Gieger, H. E. Wichmann, A. Peters, J. Hebebrand, et al. Gene set of nuclear-encoded mitochondrial regulators is enriched for common inherited variation in obesity. *PLoS One*, 8(2):e55884, 2013.

A. L. Volckmar, F. Bolze, I. Jarick, N. Knoll, A. Scherag, T. Reinehr, T. Illig, H. Grallert, H. E. Wichmann, S. Wiegand S, et al. Mutation screen in the GWAS derived obesity gene SH2B1 including functional analyses of detected variants. *BMC Med Genomics*, 5:65, 2012.

I. Jarick, A. L. Volckmar, C. Pütter, S. Pechlivanis, T.T Nguyen, M.R. Dauvermann, S. Beck, Ö. Albayrak, S. Scherag, S. Gilsbach, et al. Genome-wide analysis of rare copy number variations reveals PARK2 as a candidate gene for attention-deficit/hyperactivity disorder. *Mol. Psychiatry*, doi: 10.1038/mp.2012.161, 2012.

J.P. Bradfield, H.R. Taal, N.J. Timpson, A. Scherag, C. Lecoeur, N.M. Warrington, E. Hypponen, C. Holst, B. Valcarcel, E. Thiering, ..., I. Jarick et al. A genome-wide association meta-analysis identifies new childhood obesity loci. *Nat. Genet.*, 44(5):526-31, 2012.

A. Hinney, A. Scherag, I. Jarick, Ö. Albayrak, C. Pütter, S. Pechlivanis, M.R. Dauvermann, S. Beck, H. Weber, S. Scherag, et al. Genome-wide association study in German patients with attention deficit/hyperactivity disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.*, 156B(8):888-97, 2011.

I. Jarick, C.I. Vogel, S. Scherag, H. Schäfer, J. Hebebrand, A. Hinney, A. Scherag. Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum. Mol. Genet.*, 20(4):840-52, 2011.

A. Scherag, I. Jarick, J. Grothe, H. Biebermann, S. Scherag, A.L. Volckmar, C.I. Vogel, B. Greene, J. Hebebrand, A. Hinney. Investigation of a genome wide association signal for obesity: synthetic association and haplotype analyses at the melanocortin 4 receptor gene locus. *PLoS One*, 5(11):e13967, 2010.

T.D. Müller, M.H. Tschöp, I. Jarick, S. Ehrlich, S. Scherag, B. Herpertz-Dahlmann, S. Zipfel, W. Herzog, M. de Zwaan, R. Burghardt, et al. Genetic variation of the ghrelin activator gene ghrelin O-acyltransferase (GOAT) is associated with anorexia nervosa. *J. Psychiatr. Res.*, 45(5):706-11, 2010.

E.K. Speliotes, C.J. Willer, S.I. Berndt, K.L. Monda, G. Thorleifsson, A.U. Jackson, H. Lango Allen , C.M. Lindgren, J. Luan, R. Mägi, ..., I. Jarick, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.*, 42(11):937-48, 2010.

A. Scherag, C. Dina, A. Hinney, V. Vatin, S. Scherag, C.I. Vogel, T.D. Müller, H. Grallert, H.E. Wichmann, B. Balkau, ..., I. Jarick, et al. Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS Genet.*, 6(4):e1000916, 2010.

Verzeichnis akademischer Lehrer

Meine akademischen Lehrer waren die Damen und Herren

in Heidelberg

Bastian, Banagl, Bell, Dahlhaus, Eichberger, Gloede, Jäger, Johannes, Kanschat, Kieser, Kogelschatz, Leinert, Oelschläger, Richter, Rannacher, Sawitzki, Venjakob, Wichelhaus, Wingberg

in Marburg

Schäfer.

Danksagung

Die vorliegende Arbeit wurde am Institut für Medizinische Biometrie und Epidemiologie (IMBE) der Philipps-Universität Marburg angefertigt. Mein besonderer Dank gilt meinem Betreuer Prof. Dr. Helmut Schäfer, ohne den die Erstellung dieser Doktorarbeit nicht möglich gewesen wäre. Insbesondere bin ich dankbar für die fortwährende Möglichkeit an einer Vielzahl spannender und interessanter Fragestellungen arbeiten zu können. Darüber hinaus danke ich allen derzeitigen und ehemaligen Kollegen des IMBE für die gute Zusammenarbeit und das nette Arbeitsklima.

Besonders danke ich Prof. Dr. Johannes Hebebrand, Dr. Anke Hinney, Dr. Andre Scherag sowie allen weiteren klinischen und molekulargenetischen Kooperations-Partnern des NGFNplus Adipositasnetzes für die konstruktive und produktive Zusammenarbeit. Die Vielzahl an gemeinsamen Projekten, die mich stets inhaltlich bereicherten und häufig fachlich herausforderten, haben diese Arbeit wesentlich geprägt.

Ehrenwörtliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die dem Fachbereich Medizin Marburg zur Promotionsprüfung eingereichte Arbeit mit dem Titel „Strategies for Genome-Wide Association Analyses of Raw Copy Number Variation Data“ im Institut für Medizinische Biometrie und Epidemiologie unter Leitung von Prof. Dr. H. Schäfer ohne sonstige Hilfe selbst durchgeführt und bei der Abfassung der Arbeit keine anderen als die in der Dissertation aufgeführten Hilfsmittel benutzt habe. Ich habe bisher an keinem in- oder ausländischen Medizinischen Fachbereich ein Gesuch um Zulassung zur Promotion eingereicht, noch die vorliegende oder eine andere Arbeit als Dissertation vorgelegt.

Teile der vorliegenden Arbeit wurden in folgenden Publikationsorganen
Human Molecular Genetics (2011, 20(4): 840-852)
Molecular Psychiatry (2012, doi: 10.1038/mp.2012.161)
veröffentlicht.

Zudem wurden Teile der vorliegenden Arbeit als Poster oder Vortrag auf folgenden Tagungen
2nd Annual Meeting of NGFN-Plus and NGFN-Transfer, Berlin, 26.-28.11.2009
3rd Annual Meeting of NGFN-Plus and NGFN-Transfer, Berlin, 25.-27.11.2010
26. Jahrestagung der DAG e.V., Berlin, 4.-6.11.2010
(Posterpreis)
56. GMDS-Jahrestagung 2011 & 6. DGEpi-Jahrestagung, Mainz, 26.-29.09.2011
27. Jahrestagung der DAG e.V., Bochum, 6.-8.11.2011
20th World Congress of Psychiatric Genetics, Hamburg, 14.-18.10.2012
(Richard Todd Award)
vorgestellt.

Marburg, den 23.05.2013

Dipl. Math. Ivonne Jarick

Appendix

R-function for probe-wise specification of MCMR reference intensity values

```
find.MCMR.reference <- function( all.probewise.intensities ){

  # The input 'all.probewise.probewise' should equal a vector of
  # sample-wide probe-wise pre-processed hybridization intensity
  # values from SNP arrays.

  # As output the probe-specific MCMR reference intensity value will
  # be given.

  # For the estimation of the underlying probe-wise Gaussian mixture
  # model those vector of intensities that is free of outliers is
  # used.

  probewise.intensities <-
    all.probewise.intensities[ !all.probewise.intensities %in%
                        boxplot.stats(all.probewise.intensities)$out ]

  # The underlying Gaussian mixture model is estimated by use of the
  # R-package 'Mclust' which applies the EM algorithm for model
  # parameter estimation and the BIC for model selection.

    library( mclust )

  ints.clust.est <- Mclust( probewise.intensities )

  # The MCMR reference intensity value is selected to equal the mean
  # of those component that the samples mean is most probably
  # underlying.

  sample.mean <- mean( probewise.intensities )

  numb.clusts <- ints.clust.est$G

  lik.mean       <- c(1:numb.clusts)
  post.prob.mean <- c(1:numb.clusts)
```

```
  if( numb.clusts > 1 ){

    for( c in 1:numb.clusts ){

      m <- ints.clust.est$parameters$mean[c]

      if( ints.clust.est$parameters$variance$modelName == "E" ) {
          sd <- sqrt(ints.clust.est$parameters$variance$sigmasq)
      } else {
          sd <- sqrt(ints.clust.est$parameters$variance$sigmasq[c])
        }

      prop        <- ints.clust.est$parameters$pro[c]
      lik.mean[c] <- dnorm( sample.mean), m, sd ) * prop
    }

    for( c in 1:numb.clusts ){
      post.prob.mean[c] <- lik.mean[c] / sum(lik.mean)
    }

    mean.clust <- which.max( post.prob.mean )

    MCMR.reference <- ints.clust.est$parameters$mean[mean.clust]

  } else {

      MCMR.reference <- ints.clust.est$parameters$mean[1]
    }

  return( MCMR.reference )

}
```