METHODS TO INVESTIGATE GENE-STRATA INTERACTION IN GENOME-WIDE ASSOCIATION META-ANALYSES ON THE EXAMPLE OF OBESITY

Dissertation
zur Erlangung des Doktorgrades
der Biomedizinischen Wissenschaften
(Dr. rer. physiol.)

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Thomas Winkler
aus
Regensburg

im Jahr
2015

Dekan:	Prof. Dr. Dr. Torsten E. Reichert
Betreuer:	*Prof. Dr. Iris Heid*
Tag der mündlichen Prüfung:	29.07.2015

TABLE OF CONTENTS

1 Introduction

The overarching goal of epidemiology is to study the etiology of diseases. Diseases with substantial genetic components are called heritable diseases and are clustered into monogenic and complex diseases.

Monogenic diseases are typically rare and caused by mutations in a single gene. Examples for monogenic diseases are Cystic Fibrosis, Huntington's disease, Sickle cell anemia or the fragile X syndrome.

In contrast, complex diseases are characterized by a complicated interplay of multiple genetic and environmental factors. They are also referred to as multifactorial diseases. Examples for complex diseases are common diseases such as cancer, diabetes, cardiovascular disease, asthma, psychiatric illnesses, inflammatory diseases or obesity. Single genetic factors typically contribute very little to the development of complex diseases. However, an accumulation of multiple small but disadvantageous genetic factors in combination with environmental factors that may further be interacting with each other contributes substantially to the development of complex diseases.

Genetic epidemiology is the scientific field that aims to unravel the complicated interplay of genetic and environmental factors that influence complex disease development (Khoury, Beaty, & Cohen, 1993). Revealing the underlying genetic mechanism is pivotal for understanding disease etiology and may lead to novel therapies, improved prediction or targeted prevention programs.

1.1 Obesity and genetics of obesity

Over the past decades, obesity has become one of the world's major healthcare problems (Caballero, 2005). In particular the westernized countries have developed highly obesogenic environments that have led to a sharp increase in obesity prevalence: In Germany, more than 37% of individuals are estimated to be overweight (body mass index, BMI ≥ 25 kg/m^2 and BMI < 30 kg/m^2) and another 20% are estimated to be obese (BMI ≥ 30 kg/m^2) (Rubner-Institut, 2008). Obesity is strongly associated with mortality, an association that is mediated through increased risk for different morbidities, such as cardiovascular diseases (e.g., coronary heart disease), metabolic diseases (e.g., type 2 diabetes), psychiatric diseases (e.g., depression), or cancer (Haslam & James, 2005; Samanic, Chow, Gridley, Jarvholm, & Fraumeni, 2006). Due to its severe consequences, obesity has overcome the impact of smoking and drinking on individual health and on total healthcare costs (Moriarty et al., 2012; Sturm, 2002).

1. Introduction

Obesity can generally be classified into two categories that are independently associated with increased risk for morbidity and mortality (Pischon et al., 2008): *Overall obesity* as measured by BMI reflects total body mass and *central obesity* as measured by waist circumference or waist-hip ratio (WHR) reflects abdominal obesity or body fat distribution.

Obesity risk as well as BMI and WHR are known to be influenced by environmental factors, such as sex, age, smoking, nutritional factors and physical activity. For example, the obesity prevalence is generally higher in women than in men (Lovejoy, Sainsbury, & Stock Conference Working, 2009). Considering life course, men slowly accumulate fat around the waist, whereas women store more fat around hips at younger ages and begin to accumulate more fat around the waist after menopause, when estrogen levels drop (Kirchengast, 2010; Loomba-Albrecht & Styne, 2009). Furthermore, moderate smokers display lower body weight, lower BMI, but higher waist circumference than non-smokers and gain weight after smoking cessation (Chiolero, Faeh, Paccaud, & Cornuz, 2008). Notably, heavy smokers display increased weight and BMI, which may be due to an accumulation of risky behaviors, such as low physical activity or high caloric intake that overcomes the BMI decreasing effect of moderate smoking on obesity measures. Finally, a reasonable diet and increased physical activity lowers the risk of obesity and related diseases (Lakka & Bouchard, 2005), but both is incredibly difficult to implement into obese persons' life styles.

Both, overall and central obesity, involve substantial genetic components comprising high estimates of heritability (i.e., phenotypic variance explained by genetics, typically >70% for BMI, and >45% for WHR) (Farooqi & O'Rahilly, 2000; Rose, Newman, Mayer-Davis, & Selby, 1998; Zaitlen et al., 2013).

Although obesity can generally be classified as a common complex disease, there are also some rare monogenic forms of obesity.

The prominent melanocortin-4 receptor (*MC4R*) gene has multiple implications in obesity. Generally, *MC4R* acts in the central melanocortinergic system and regulates food and energy intake (Adan et al., 2006). On the one hand, a single rare mutation in *MC4R* gene was shown to cause a specific monogenic form of childhood obesity (Farooqi et al., 2003). On the other hand, a common variant located near the *MC4R* gene (present in ~24% of the general population) was shown to be associated with increased BMI (Loos et al., 2008). Although the single common *MC4R* variation explained only ~0.10% of the total BMI variation, it helped highlighting an interesting biological pathway: The common risk variant showed decreased *MC4R* protein levels in the hypothalamus that leads to increased appetite and decreased satiety (Qi, Kraft, Hunter, & Hu, 2008). Therefore, *MC4R* agonists are interesting candidates for the pharmacological development of drugs to treat not only rare monogenic forms but also the common form of obesity (Adan et al., 2006).

Another prominent obesity gene that illustrates the complex interplay of genetic and environmental factors is the 'fat mass and obesity associated' (*FTO)* gene. Initially identified for its association with type 2 diabetes (T2D), follow-up analyses adjusted the T2D association for BMI and showed that the T2D association of the common *FTO* variant disappeared (Frayling et al., 2007). This suggested that the impact of the common risk variant in *FTO* (present in ~42% of the general population) on T2D was mediated through obesity. Functional follow-up studies using gene knock-out mice revealed that *FTO* variants were implicated in energy homeostasis (Fischer et al., 2009). This was one of the first successful attempts to translate genetic epidemiology association study results into functional processes and consequences affecting body weight regulation. Finally, one of the first gene-environment interactions highlighted for obesity was observed for *FTO*: The effect of *FTO* variants on obesity risk was shown to be attenuated by higher physical activity levels (Andreasen et al., 2008).

Besides physical activity, other environmental factors such as sex, age, smoking or nutrition may modify genetic effects on obesity and may – at least in part – explain known differences in obesity measures between men and women, between smokers and non-smokers, between individuals on high-calorie and low-calorie diet or explain the changes in body shape over life course.

In summary, these examples illustrate the importance to investigate the genetic underpinning of obesity to further the understanding of involved mechanism that may ultimately lead to improved therapeutic options.

1.2 Genetic association studies

The aim of genetic association studies is to identify association between genetic variation and an outcome of interest, such as disease (e.g., type 2 diabetes) or disease-relevant parameters (e.g., BMI). The outcome is referred to as *phenotype* in the following.

The most frequent forms of genetic variation in the Deoxyribonucleic Acid (DNA) are *single nucleotide polymorphisms* (SNPs). Each SNP denotes a single base-exchange that originated at some point during evolution and is located at a specific position in the DNA sequence (**Figure 1**). Most SNPs are bi-allelic comprising two possible nucleotide variations, also referred to as *alleles*. The combination of two alleles - one on the maternal and one on the paternal chromosome - makes up the so-called *genotype* of the SNP for a specific person. SNPs are called *common* if the less frequent allele (i.e., *minor allele*) is present in at least 5% of individuals in a population. Less frequent SNPs are called *rare.* More than 60 million SNPs are known to date including approximately 10 million SNPs with minor allele frequency (MAF) greater than 1% (Sherry et al., 2001).

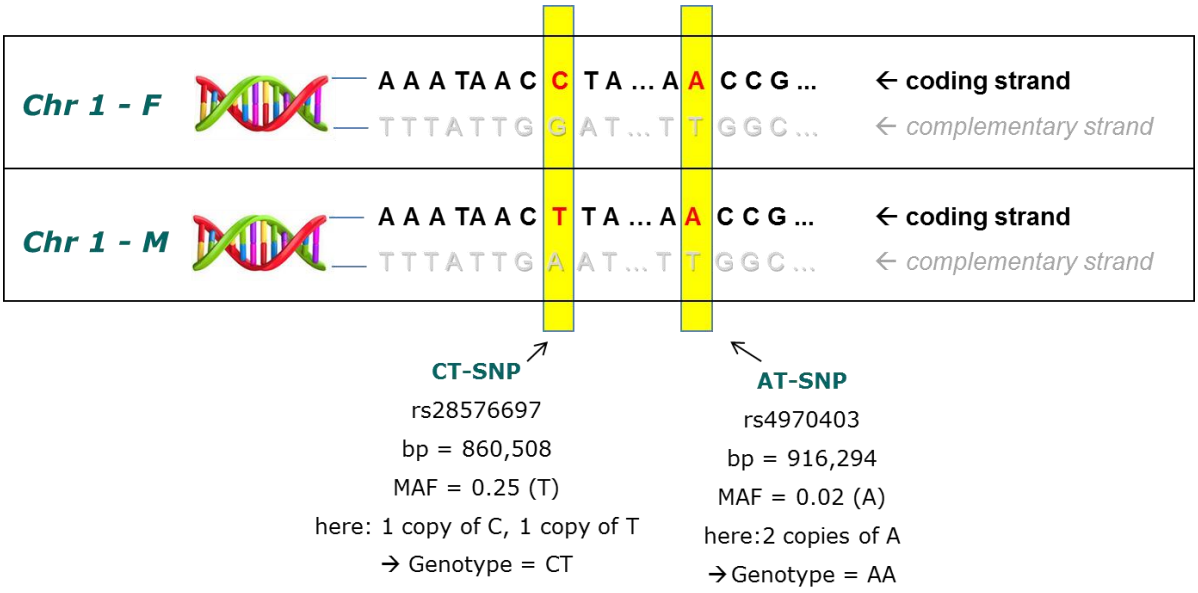


Figure 1. *Schematic presentation of two SNPs. Shown are the coding and the complementary strand for both versions of chromosome 1 (paternal and maternal chromosome 1).*

In the early 2000s, SNP-phenotype associations have mainly been identified through hypothesis-driven *candidate-gene* approaches. Such approaches require a-priori knowledge on the biology of the phenotype and on the possible implications of potential candidate genes. The problem with this approach was that it required choosing the right candidate gene in advance, a process that can be difficult because the decision has to be made on the current state of knowledge, which may be limited. Overall, candidate-gene approaches were often unsuccessful, mostly because wrong candidates were chosen or because the power of single candidate-gene studies was too low.

Instead, *genome-wide association studies* (GWAS) have recently been found to be more efficient. GWAS are hypothesis-free approaches that simultaneously screen a dense field of millions of SNPs - spread across the whole genome - for association. Importantly, to avoid large numbers of false positive findings, GWAS require rigorous control for the multiple testing of millions of variants.

GWAS were only yet enabled through technical advances in genotyping since 2005 (Hirschhorn & Daly, 2005). Improved chip-based microarray technologies nowadays allow study centres to effectively assess one million or more SNPs for large studies including thousands of individuals. A large number of genotyping chips have been developed over the past years (Distefano & Taverna, 2011). Besides genome-wide chips that cover SNPs equally spread across the whole genome, customized genotyping chips were designed that fine-map particular regions of interest. For example, the MetaboChip specifically covers regions associated with metabolic disorders, including obesity (Voight et al., 2012).

1. Introduction

Besides advances in direct genotyping chip technologies, genotype *imputation* has largely contributed to the success of GWAS (Y. Li, Willer, Sanna, & Abecasis, 2009). Imputation methods allow for inferring unmeasured variants on the basis of reference sequence data. Imputation is pivotal for comparing association results across studies with different genotyping platforms. Over the past years, almost all GWAS have used imputed data with up to ~2.5M variants based on HapMap reference panels (International HapMap, 2005).

A challenge in GWAS is that genetic effects are typically small. For example, in obesity, the largest genetic effect observed for BMI (near *FTO*) explains only ~0.34% of the total BMI variation (Speliotes et al., 2010). Medium genetic effects on BMI are even smaller by ten-fold and explain only ~0.04% of the BMI variation. Given the multiple testing burdens, identification of such subtle effects necessitates large sample sizes and single GWAS involving ~1,000 individuals are often underpowered.

One way to increase power to detect small genetic effect sizes is to pool multiple GWAS in so-called *genome-wide association meta-analyses* (GWAMAs). Meta-analysis of aggregated statistics across multiple studies and for each SNP genome-wide allows for increasing power when an individual participant data analysis is not possible. This is typically the case in genetic studies because study partners are most often not allowed to share individual level participant genotype data due to ethical constraints. Over the past years, many GWAMA consortia have emerged and multiplied the total sample size for several diseases and disease-relevant parameters. For example, one of the largest GWAMA consortia worldwide is the Genetic Investigation of Anthropometric Traits (GIANT) consortium. GIANT has set out to investigate the genetic underpinning of anthropometric and obesity traits (primarily focusing on height, BMI and WHR) involving hundreds of single GWAS and hundreds of thousands of individuals altogether.

Taken together, GWAS and GWAMAs have successfully been employed over the past years and have led to a major increase in the number of known SNP-phenotype associations (Visscher, Brown, McCarthy, & Yang, 2012; Welter et al., 2014). So far, more than 14,000 SNP-phenotype associations have been reported that are accessible through publicly available data bases, such as the GWAS catalogue (Hindorff et al., 2009).

Yet, despite the large success of GWAS, a substantial fraction of the heritability still remains unexplained for most phenotypes (Manolio et al., 2009). A fraction of the *missing heritability* might be explained by rare variants (MAF < 5%, expected to yield larger effect sizes than common variants) or by structural variation (e.g., copy number variants such as insertions or deletions), both of which have mostly been missed due to low power or due to unsuitable array designs (genotyping arrays primarily focused on common variants, i.e., SNPs with MAF > 5%). Detecting rare variant or structural variation effects requires

systematic screens that employ novel genotyping arrays as well as denser imputation reference panels.

Another fraction of the missing heritability might be explained by gene-environment interaction effects that have been missed and ignored by the commonly conducted genome-wide scans focussing on overall associations. Detecting gene-environment interaction effects requires systematic screens that employ large sample sizes, extended statistical methods and software tools that are applicable to large-scale genome-wide data sets. Yet, such methods are poorly understood and software tools are lacking.

1.3 Statistical models and methods for genetic association studies

The following chapter introduces statistical models and concepts of SNP-phenotype association testing, GWAS and GWAMAs.

1.3.1 The linear regression model

A general SNP-trait association test is based on regression methods that fit a Generalized Linear Model (GLM). A GLM allows for correcting for potential confounders by using additional covariables. For a continuous phenotype Y, a *linear regression model* is considered:

$$Y = \propto + \beta G + \beta_{C_1} C_1 + \cdots \beta_{C_k} C_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \qquad (1).$$

Here, G denotes the SNP genotype, C_i are the co-variables, α is the intercept of the regression model, β the genetic effect on Y, and ε a random error variable, also called residual. The linear regression model is based on some important assumptions that involve (i) lack of auto-correlation (i.e., residuals are assumed to be independent and to follow a normal distribution with zero mean and residual variance σ^2), (ii) homoscedasticity (i.e., the residual variance is assumed to be constant across genotype), and (iii) a linear and additive relationship between genotypes and phenotypes. To ensure comparability of the phenotype across studies, one approach is to normalize (or to standard normalize) the phenotype per study which yields $Y \sim N(\mu_Y, \sigma_Y^2)$ (or $Y \sim N(0,1)$).

In genetic epidemiology, regression models are usually adjusted for other epidemiological factors (added as co-variables) that either have an impact on the phenotype or have an impact on both, the phenotype and the genotype.

Adjusting for factors that are known to influence the phenotype (only) reduces the phenotypic variance by the proportion that is explained through the respective co-variable and as such increases the power to find the genotype-phenotype association. For example,

due to their well-known influence on anthropometric traits, GIANT requests contributing study partners to adjust for sex and age.

Adjusting for factors that are known to influence both the phenotype and the genotype (i.e., *confounders*), prevents from observing an association between genotype and phenotype that is actually driven by the hidden confounding variable. Genetic association models are usually adjusted for potential confounding through *population stratification* that reflects systematic diversities between population substructures. To avoid such confounding, in many cases the first ten independent genotype dimensions (principle components) are added to the regression model as co-variables.

1.3.2 SNP genotype models

Usual genotype models are the recessive, the dominant and the additive model. Consider a SNP with two alleles: Major allele *A* denoting the more frequent allele, and minor allele *a* denoting the less frequent allele. For a specific SNP, an individual can take three possible genotype states: *AA* (two copies of the major allele), *Aa* (one copy of the major and one of the minor allele) and *aa* (two copies of the minor allele).

The *dominant model* implies that individuals with one or two copies of the minor allele exhibit the phenotype (with equal probability). Thus, the genotype variable is coded G = 0 for genotype *AA* and coded G = 1 for genotypes *Aa* and *aa*.

The *recessive model* implies that only individuals with two copies of the minor allele exhibit the phenotype. Thus, the genotype variable is coded G = 0 for genotypes *Aa* and *AA*, and coded G = 1 for genotype *aa*.

The commonly used *additive model* implies that the probability to exhibit the phenotype increases linearly with each additional copy of the minor allele. Thus, the genotype variable is coded G = 0 for genotype *AA*, G = 1 for genotype *Aa* and G = 2 for genotype *aa*. With MAF being the minor allele frequency of a particular SNP, the additively modelled SNP genotypes 0,1 and 2 occur with probabilities $(1–MAF)^2$, $2MAF(1–MAF)$ and MAF^2, respectively, if Hardy-Weinberg equilibrium is fulfilled (Edwards, 2008). For large sample sizes, the binomial genotype distribution approximates a normal distribution with genotypic mean $\mu_G = 2 \cdot MAF$ and genotypic variance $\sigma_G^2 = 2MAF(1 - MAF)$.

1.3.3 SNP association testing

To infer whether the modelled SNP genotype is associated with the phenotype, i.e., whether the genetic effect on *Y* - estimated from the regression model - is significantly different from zero, a t test can be conducted that compares the null hypothesis $H_0: \beta = 0$ versus the alternative hypothesis $H_A: \beta \neq 0$. Herewith, a test statistic $T = b/se(b)$ is employed, where *n* is the sample size, *b* is the observed genetic effect estimate of *β,* and *se(b)* the standard

error of *b*. Assuming the nulll hypothesis, the test statistic *T* follows a t distribution with *n-2* degrees of freedom (df), i.e., $T \sim t(n-2)|H_0$. The t test yields a SNP association P-Value *P*.

1.3.4 Genome-wide association studies

In GWAS, the association testing is conducted separately and simultaneously for the millions of SNPs available for a study. To avoid huge numbers of false positive findings, SNP-specific association results are corrected for the multiple testing. Typically in HapMap imputation based GWAS, a conservative genome-wide significance threshold of α = 5 x 10^{-8} is applied that Bonferroni-corrects the usual 5% α-level for an approximate number of one million (1M) independent SNP association tests (Johnson et al., 2010).

1.3.5 Genome-wide association meta-analyses

For each SNP, each GWA study *j* provides study-specific summary estimates such as the genetic effect estimate b^j, the corresponding standard error $se(b^j)$, the association P-Value P^j and the sample size n^j. To obtain pooled genetic effect estimates and standard errors for each SNP, an *inverse-variance weighted meta-analysis* can be conducted, computing

$$b = \frac{\sum_j b^j/se(b^j)^2}{\sum_j 1/se(b^j)^2} \quad \text{and} \quad se(b) = \sqrt{\frac{1}{\sum_j 1/se(b^j)^2}} \qquad (2).$$

In the meta-analytical setting, *b* and *se(b)* are referred to as *pooled* genetic effect estimate and *pooled* standard error. As in single study association testing, a t test statistic $T = b/se(b) \sim t(n-2)|H_0$ is utilized to infer, whether the pooled genetic effect is significantly different from zero. The t test yields a pooled overall association P-Value *P.* The meta-analysis formulae (2) assume a fixed effect model across studies. For homogeneous genetic effects across studies, the *pooled* genetic effect estimates and standard errors are approximately the same as the genetic effect estimates and standard errors obtained from a single regression model using one large study involving all individuals (Behrens, Winkler, Gorski, Leitzmann, & Heid, 2011).

1.4 Genome-wide association meta-analyses for obesity traits: The GIANT consortium

One of the largest GWAMA consortia worldwide is the Genetic Investigation of ANthropometrics Traits (GIANT) consortium (**Figure 2**). This consortium has set out to describe the genetic underpinning of anthropometric and obesity traits. For the obesity traits, the primary focus is on body mass index (BMI, as a measure of overall obesity) and on waist

hip ratio adjusted for BMI (WHR_{adjBMI}, as a measure of central obesity that is independent of BMI).

Since 2006, GIANT has published multiple rounds of meta-analyses on these primary traits, with each round iteratively increasing the total sample size, increasing the total number of studies involved, adding novel genotyping chip technologies and incorporating larger imputation reference panels (Heid et al., 2010; Lindgren et al., 2009; Loos et al., 2008; Speliotes et al., 2010; Willer et al., 2009). In 2010, the number of identified loci was raised to 32 for BMI (using discovery GWAS data from up to 123,865 individuals) and to 14 for WHR_{adjBMI} (using discovery GWAS data from up to 77,167 individuals).

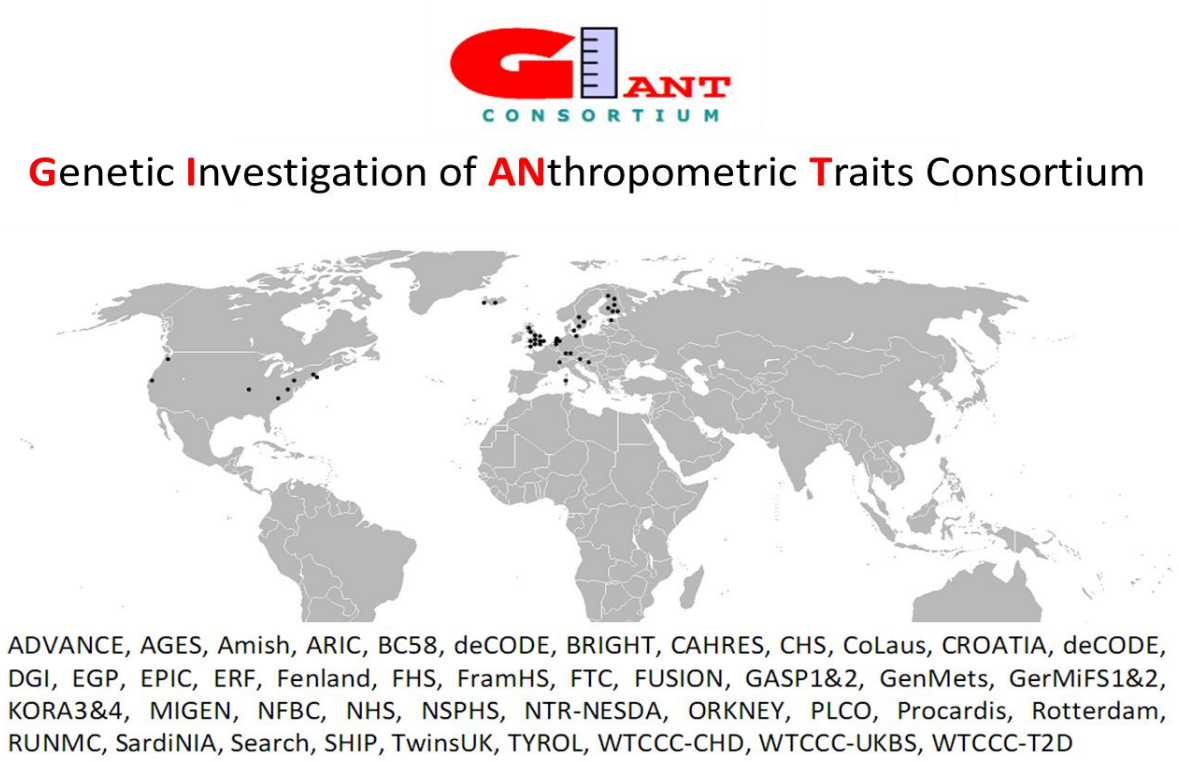


Figure 2. *GIANT consortium studies involved in the 2010 meta-analyses (for more information see the GIANT consortium website, www.broadinstitute.org/collaboration/giant).*

Due to the known sex-differences in obesity measures, the identified loci were investigated for sex-differences in consecutive follow-up analyses (using men- and women-specific GWAS results that have been provided by the study partners). No significant sex-difference was observed in any of the 32 detected BMI loci. In contrast, seven of the 14 overall associated WHR_{adjBMI} loci were found to be sex-specific, and all of the seven displayed significantly stronger effects in women than in men (Heid et al., 2010). Remarkably, these sex-differences were detected for variants that were initially identified for

overall association, an approach that might have missed other sexually dimorphic variants. Thus, a systematic genome-wide screen to identify sex-differences in genetic effects of anthropometric traits was warranted.

In addition, further GIANT projects were initiated to investigate whether other obesity risk factors modify the genetic effect on obesity traits. These include GWAMAs stratified by smoking status (non-smokers vs. current smokers), stratified by physical activity status (inactive vs. active), as well as stratified by age and sex (men≤50y vs. women≤50y vs. men>50y vs. women>50y). The latter project aims to investigate whether genetic variation contributes to the age-dependent decrease (after menopause in women) in sex-difference of body shape (Kuk, Saunders, Davidson, & Ross, 2009; Wells, 2007). In order to reflect menopause in women, age was dichotomized at 50 years of age (corresponds to mean age of menopause in women).

Unraveling such stratum-differences in genetic effects is key to improve the understanding of the genetic underpinning of obesity, key to explain some of the missing heritability, and may be key to identify novel therapeutic opportunities.

1.5 Gene-strata interaction effects in genetic association studies

Genetic effects that differ between strata (e.g., between men and women) can equivalently be denoted as gene-strata (G x S) interaction effects. G x S interaction effects are a specific form of gene-environment interaction effects that involve a dichotomous environmental (stratification) variable S (e.g., SEX coded as 0/1 for men/women). Such effects modify the genetic effect on a phenotype between strata, a circumstance that can reduce power to find the overall (strata-combined) effect in the overall GWAMA (Behrens et al., 2011). So far, many GWAS and GWAMA projects have focused on overall effects, while ignoring G x S interaction effects that might explain a substantial fraction of the missing heritability.

The following chapters introduce statistical models and available methods to account for and to identify G x S interaction effects given the GWAMA setting.

1.5.1 Modelling gene-strata interaction effects in large-scale GWAMAs

Assuming the large-scale GWAMA configuration, G x S interaction effects can either be modelled by conducting a *stratified GWAMA* or by conducting an *interaction GWAMA* (**Figure 3**). In the following, considerations are limited to linear regression models involving continuous phenotypes *Y*, additively modeled genotypes *G* and dichotomous stratification variables *S*.

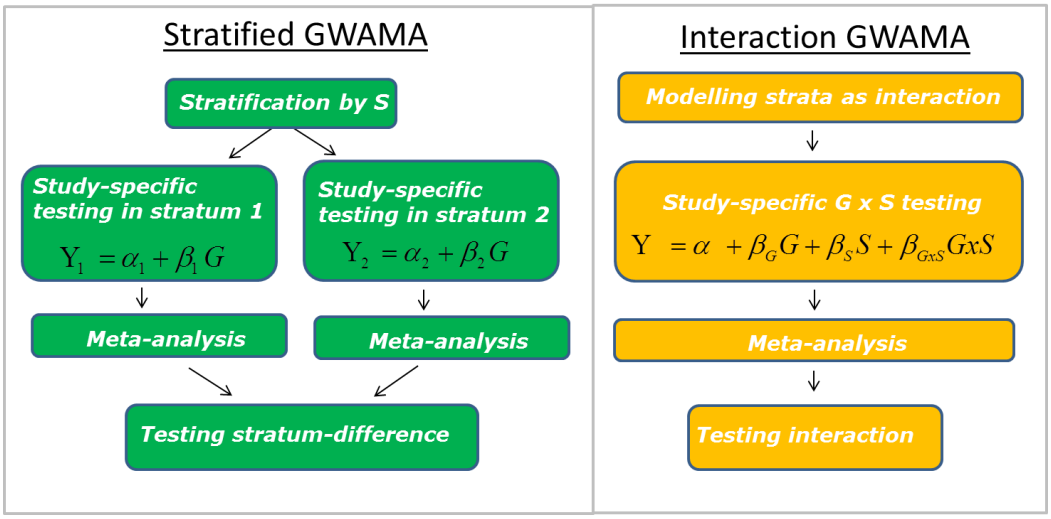


Figure 3. *Modelling G x S interaction effects in large-scale GWAMAs.*

The *interaction GWAMA model* – involving a single dichotomous environmental variable S - is given by the regression model that includes a G x S interaction term:

$$Y = \alpha + \beta_G G + \beta_S S + \beta_{GxS} GxS + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \qquad (3).$$

Here, β_G is the genetic effect on the phenotype, β_S the effect of the stratification variable on the phenotype, and β_{GxS} the G x S interaction effect on the phenotype. For each SNP, each study fits the interaction model and obtains study-specific effect and interaction estimates with standard errors. Pooled genetic effect estimates b_G and pooled G x S interaction effect estimates b_{GxS} are obtained from inverse-variance weighted meta-analyses of the respective study-specific estimates. Testing for G x S interaction effects can be accomplished by performing a t test on the pooled interaction estimates b_{GxS}.

Assuming two strata, the *stratified GWAMA model* involves two linear regression models (one for each stratum):

$$\begin{aligned} Y_1 &= \propto_1 + \beta_1 G_1 + \varepsilon_1, \quad \varepsilon_1 \sim N(0, \sigma_1^2) \\ Y_2 &= \propto_2 + \beta_2 G_2 + \varepsilon_2, \quad \varepsilon_2 \sim N(0, \sigma_2^2) \end{aligned} \qquad (4).$$

The stratification is done by a dichotomous variable that separates each study sample into two subgroups. A stratum-specific regression model is fitted for each SNP and in each study separately yielding stratum-specific effect estimates with standard errors. Pooled stratum-specific genetic effect estimates, b_1 and b_2, are obtained from stratum-specific inverse-variance weighted meta-analyses. Testing for G x S interaction can be accomplished by testing the pooled stratum-specific estimates for difference (Randall et al., 2013).

1.5.2 Methods to account for and to identify gene-strata interaction effects

When defining methods to tackle G x S interaction effects, it is extremely important to distinguish between two major aims: On the one hand one might be interested in methods to identify SNP effects while accounting for interaction; on the other hand one might be interested in detecting the interaction per se for a specific locus.

Several methods have been described that improve power to *identify SNP effects while accounting for interaction*. For example, the simple approach of stratum-specific association testing (using the pooled stratum-specific estimates gathered from a stratified GWAMA) has been shown to improve power to find stratum-sensitive variants (Behrens et al., 2011). Other methods focus on the detection of joint (main + interaction) effects (Kraft, Yen, Stram, Morrison, & Gauderman, 2007) and those methods have recently been extended to the interaction GWAMA model (Manning et al., 2011) and to the stratified GWAMA model (Aschard, Hancock, London, & Kraft, 2010). Importantly, a significant joint effect does not automatically imply significant interaction. Disentangling whether a significant joint effect is due to main effect, interaction effect, or both, has to be outlined additionally. There is some concern as to whether this can be done using the obtained main and interaction estimates from the data set that was used for discovery of joint effects.

Similarly, several methods have been described that aim at *identification of gene-environment interaction effects*. However, most of the reported methods are tailored for single studies with dichotomous disease outcomes (D. Li & Conti, 2009; Mukherjee & Chatterjee, 2008; Piegorsch, Weinberg, & Taylor, 1994). Their applicability to continuous outcomes, and to the large-scale GWAMA setting, may be limited and has not yet been shown. A structured and detailed comparison of GWAMA approaches - aiming at identification of G x S interaction effects for continuous outcomes - with regards to type 1 error and power while considering varying types of interaction effects, study designs, and statistical tests, is lacking. For example, for the GIANT sex-stratified GWAMAs for WHR_{adjBMI}, it is not yet clear what screening approach is optimal to identify loci with significant sex-difference.

1.6 Objectives

Genome-wide association meta-analyses (GWAMAs) of obesity traits have proven to successfully pinpoint associated genetic variants. For example, in 2010, a large-scale GWAMA for WHR_{adjBMI} (waist-hip ratio adjusted for BMI, as a measure of central obesity) detected significant associations at 14 genetic loci.

Interestingly, seven of the 14 loci displayed significantly stronger genetic effects in women than in men. Remarkably, these sex-differences were detected for variants that were initially selected for overall (sex-combined) association. Yet, a systematic genome-wide screen to identify variants displaying sex-difference was lacking and the dimension of sex-differences in the genetics of anthropometric traits was unknown. Sex-stratified GWAS data had already been available at that time for most studies involved in the 2010 GIANT consortium meta-analyses. However, there was uncertainty as to what screening approach should be applied ideally to identify sex-difference from the available sex-stratified GWAS. A systematic methodological evaluation of approaches with regard to type 1 error and power to identify sex-difference was lacking.

Additionally, two further GIANT projects were in planning that aimed at the identification of other stratum-differences in genetic effects for obesity traits. One project started to conduct a smoking-stratified GWAMA (to identify differences in genetic effects between current smokers and non-smokers), the other started to conduct an activity-status stratified GWAMA (to identify differences in genetic effects between physically active and inactive individuals). In contrast to the balanced sex-stratified GWAMA design (similar numbers of men and women involved) these two projects reflect an unbalanced design: For example there are fewer smokers than non-smokers available for the analyses. The impact of unequal stratum sizes on the identification of stratum-difference had not been clear.

A further project was in planning that aimed to identify age- and sex-dependent genetic effects for obesity traits. The rationale behind this analysis was that body shape changes in women upon menopause (approximately at 50 years of age) resulting in a more android body shape that is less different from men. Age- and sex-stratified GWAMAs (four strata: younger men, older men, younger women, older men; age stratified at 50 years of age) were supposed to be employed to investigate potential 3-way G x AGE x SEX interaction effects. But again, there was uncertainty as to what approach should be applied to ideally find such 3-way interaction effects from the age- and sex-stratified GWAMA results.

Finally, large-scale stratified GWAMAs are high-dimensional complex analyses. For example, GIANT involves multiple stratified GWAS results (each carrying millions of SNP-specific association test results) from hundreds of studies, for multiple genotyping platforms and imputation reference panels, for multiple anthropometric and obesity traits as well as for multiple environmental stratification variables. This involves the handling of thousands of individual GWAS result files - each with millions of SNP-specific association testing results. Ensuring validity of each single GWAS result and of the obtained GWAMA result requires extended quality control procedures and software that is able to cope with thousands of large association result files. Furthermore, software was required that provides extended statistical and graphical functionality to evaluate stratified GWAMA results.

To address the described research gaps, the four main objectives of this work were defined as follows:

1. Develop and improve stratified GWAMA approaches to identify difference in genetic effects between two strata (chapter 2).
2. Extend methods to identify 3-way G x AGE x SEX interaction effects from an age- and sex-stratified GWAMA (chapter 3).
3. Apply optimized methods to stratified GWAMA results from the GIANT consortium (chapter 4).
4. Develop software to facilitate quality control and statistical evaluation of stratified GWAMAs (chapter 5).

2 Stratified GWAMA approaches to screen for difference between two strata

Generally, *stratum-difference* is defined as the difference in genetic effects between two strata (e.g., men and women). The overarching aim of the following chapter is to provide a systematic methodological evaluation of stratified GWAMA screening approaches that are based on two strata and aim at identifying stratum-difference.

Generally, an *approach* is defined here as a combination or concatenation of multiple statistical tests (i.e., *steps*) that are applicable to stratified GWAMA outcomes and that are implemented in one or two independent data sets (i.e., *stages*). Relevant statistical tests are introduced and a systematic scheme of approaches is presented.

The performance of the approaches was compared by simulation-based estimation of type 1 error and by analytical computations of power. Varying realistic scenarios were considered and an attempt to recommend approaches - based on study design and based on type of stratum-difference - was made.

Three general types of stratum-difference were defined (**Figure 4**). Assuming an effect in one stratum (e.g., women), the effect in the other (i.e., men) may be opposite (*opposite effect direction, OED*), lacking (*single-stratum effect, SSE*) or be concordant but less pronounced (*concordant effect direction, CED*).

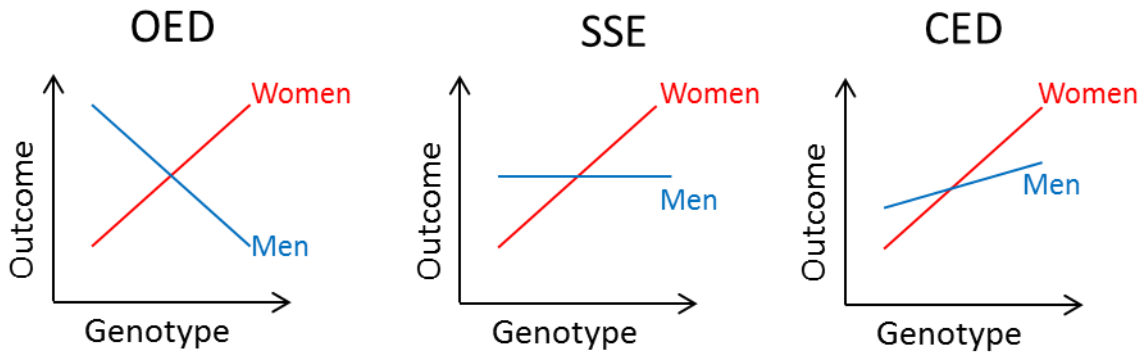


Figure 4. *Different types of stratum-specific effects on the example of sex-difference (assuming a positive effect in women).*

2.1 Materials and Methods

The following chapters describe the prerequisites of the methodological evaluation of stratified GWAMA approaches to identify stratum-difference.

After presenting general assumptions (chapter 2.1.1), relevant statistical tests are introduced (chapters 2.1.2 and 2.1.3) and then used to construct several screening approaches (chapter 2.1.4). A systematic scheme of approaches was developed that is utilized to compare performance (type 1 error and power) between approaches. Details about the simulation-based type 1 error evaluation (chapter 2.1.5), the derivation of analytical power formulae as well as about the analytical power computations (chapter 2.1.6), are presented.

2.1.1 Assumptions and definitions

In the following, considerations are based on a stratified GWAMA model involving two strata (as given by equation (4)), a continuous phenotype (e.g., WHR_{adjBMI}), a dichotomous stratification variable (e.g., sex) and additively modeled genotypes. It is assumed that the stratified GWAMAs have already been conducted so that pooled stratum-specific effect estimates and standard errors are available.

Stratum-specific continuous phenotypes are assumed to follow identical normal distributions, $Y_1 \sim N(\mu_Y, \sigma_Y^2)$ and $Y_2 \sim N(\mu_Y, \sigma_Y^2)$.

Similarly, stratum-specific additively modeled genotypes G_1 and G_2 are assumed to follow equal genotype distributions across strata implying identical minor allele frequencies (MAF) and thus identical genotype means $\mu_G = \mu_{G1} = \mu_{G2} = 2MAF$ and identical genotype variances $\sigma_G^2 = \sigma_{G1}^2 = \sigma_{G2}^2 = 2MAF(1 - MAF)$. This assumption builds upon the assumption of random mating between strata.

Importantly, the GWAMAs are assumed to only include GWAS from similar populations. Similar populations involve equal MAF across studies. This implies equal genotype distributions and homogeneous genetic effects across studies. Based on this assumption, meta-analysis of multiple study-specific SNP summary statistics yields approximatively identical results as a 'mega-analysis' of all individuals in one single large study (Evangelou & Ioannidis, 2013). Therefore, the meta-analysis concept was ignored in the following and one large study involving all individuals was assumed.

A *sex-stratified GWAMA* is defined as men- and women-specific GWAMA involving equal sex-specific sample sizes, $n_M = n_F$. For a *general stratified GWAMA*, stratum-specific sample sizes are allowed to be unequal, $n_2 = f \cdot n_1$, where f is the ratio of stratum 2 sample size to stratum 1 sample size ($f = 1$ for the sex-stratified GWAMA setting).

2.1.2 Testing for stratum-difference

To investigate stratum-difference given a stratified GWAMA model with two strata, a *difference test* can be conducted that compares the pooled stratum-specific genetic effect estimates ($H_0: b_1 = b_2$):

$$Z_{Diff} = \frac{b_1 - b_2}{\sqrt{se(b_1)^2 + se(b_2)^2}} \sim N(0,1) \mid H_0 \qquad (5).$$

Assuming the null hypothesis being true, the z statistic follows a standard normal distribution. The z test yields the difference P-Value P_{Diff}.

Under the assumption of independent samples, unrelated subjects and no latent covariate interacting with S (the stratification variable), the difference test is mathematically equivalent to testing the pooled G x S interaction estimate b_{GxS} obtained from an interaction GWAMA model (as given by equation (3)).

In order to correct for potential correlation between stratum-specific effects b_1 and b_2, an alternative of the difference test can be employed:

$$Z_{Diff} = \frac{b_1 - b_2}{\sqrt{se(b_1)^2 + se(b_2)^2 - 2r \cdot se(b_1) \cdot se(b_2)}} \sim N(0,1) \mid H_0 \qquad (6).$$

Here, r denotes the Spearman rank correlation between b_1 and b_2 that is estimated from the two stratum-specific genome-wide data sets. For example, such correlation could stem from family studies that contribute related individuals to both strata, e.g., brothers and sisters contributing to a sex-stratified GWAMA. Such relatedness would result in 'less different' effect estimates, increased type 2 error and deflated P_{Diff}. The correction should only be used with genome-wide data sets that allow for accurate estimation of the correlation.

In the following, unrelated subjects across strata are assumed and stratum-specific estimates b_1 and b_2 are assumed to be uncorrelated.

2.1.3 Statistical tests to filter stratified GWAMA data sets prior to difference testing

Often in GWAMA literature, the difference (or interaction) testing is limited to SNPs that were initially selected using other statistical tests, such as stratum-specific, overall or joint (main + interaction) association tests (Heid et al., 2010; Manning et al., 2012; Randall et al., 2013). For example, Heid and colleagues primarily screened for overall (strata-combined) associated variants for WHR_{adjBMI}, and subsequently tested the identified (overall associated) SNPs for sex-difference.

In order to construct stratified GWAMA approaches that reflect such analyses, three *filtering tests* are considered that can directly be applied to stratified GWAMA outcomes b_1 and b_2, with respective standard errors $se(b_1)$ and $se(b_2)$:

- A ***stratified test*** can be performed to infer whether any of the stratum-specific effects is associated with the phenotype ($H_0: b_1 = 0 \wedge b_2 = 0$): The stratified test employs two t tests, $T_1 = b_1/se(b_1) \sim t(n_1 - 2) \,|H_0$ and $T_2 = b_2/se(b_2) \sim t(n_2 - 2) \,|H_0$, that yield stratum-specific association P-Values, P_1 and P_2. Finally, a stratified association P-Value is defined as $P_{Strat} = 2*min(P_1,P_2)$, which is corrected for the multiple testing of two strata.
- An ***overall test*** can be performed to infer whether the overall (strata-combined) effect is associated with the phenotype ($H_0: b_{Overall} = 0$):

$$T_{Overall} = \frac{b_{Overall}}{se(b_{Overall})} \sim t(n_{Overall} - 2) \,|H_0 \qquad (7),$$

where $n_{Overall}$ is the strata-combined sample size ($n_{Overall} = n_1 + n_2$), and where $b_{Overall}$ and $se(b_{Overall})$ are the overall genetic effect estimate with standard error that are obtained from inverse-variance weighted meta-analysis of the two strata:

$$b_{Overall} = \frac{b_1/se(b_1)^2 + b_2/se(b_2)^2}{1/se(b_1)^2 + 1/se(b_2)^2}$$
$$se(b_{Overall}) = \sqrt{\frac{1}{1/se(b_1)^2 + 1/se(b_2)^2}} \qquad (8).$$

The t test yields the overall association P-Value $P_{Overall}$.

- A ***joint test*** can be performed to infer whether the joint effect of both the main effect and the interaction effect is associated with the phenotype ($H_0: b_G = 0 \wedge b_{GxS} = 0$):

$$C_{Joint} = \left(\frac{b_1}{se(b_1)}\right)^2 + \left(\frac{b_2}{se(b_2)}\right)^2 \sim \chi^2(2) \,|H_0 \qquad (9).$$

The chi-square test yields the joint-test P-Value P_{Joint}. The joint test based on stratum-specific effects originates from the joint test that simultaneously tests for main and interaction effects in an interaction model (Kraft et al., 2007). The two versions of the joint tests are identical, if the G x S interaction is modelled with dichotomized S (Aschard et al., 2010).

2.1.4 A systematic scheme of stratified GWAMA approaches to identify stratum-difference

The introduced statistical tests are used to construct several stratified GWAMA approaches, each of which aims at screening for variants with significant stratum-difference. Generally, an approach is designed with multiple steps (statistical tests). The last step of each approach is to test for difference. The stratified, the overall, or the joint test, are solely employed for filtering the genome-wide data sets prior to the difference testing. Multiple steps are either implemented within a single data set (*1-stage design*) or implemented within two independent data sets (discovery and replication data set, *2-stage design*). For the 2-stage approaches, the filtering is conducted in the discovery stage data and difference testing is

either conducted using the replication stage data only (*replication-based 2-stage* approach) or using the combined (discovery + replication) stage data (*combined 2-stage approach*).

To distinguish between approaches, a general notation is introduced: *[.]* indicates a stage and 'Test_{α}' denotes the step-specific test and the respective α-level. For example, $[\mathrm{Overall}_{\alpha_1} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ denotes a 1-stage approach that involves filtering on $P_{Overall} < \alpha_1$ and testing the selected SNPs for difference and that employs identical subjects for both steps. Alternatively, $[\mathrm{Overall}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ (or $[[\mathrm{Overall}_{\alpha_1}] \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$) denotes the respective replication-based 2-stage (or combined 2-stage) approach that involves testing the selected SNPs for difference using the replication (or combined discovery + replication) subjects.

Based on this notation, a systematic scheme of the considered approaches was developed and summarized in **Figure 5.**

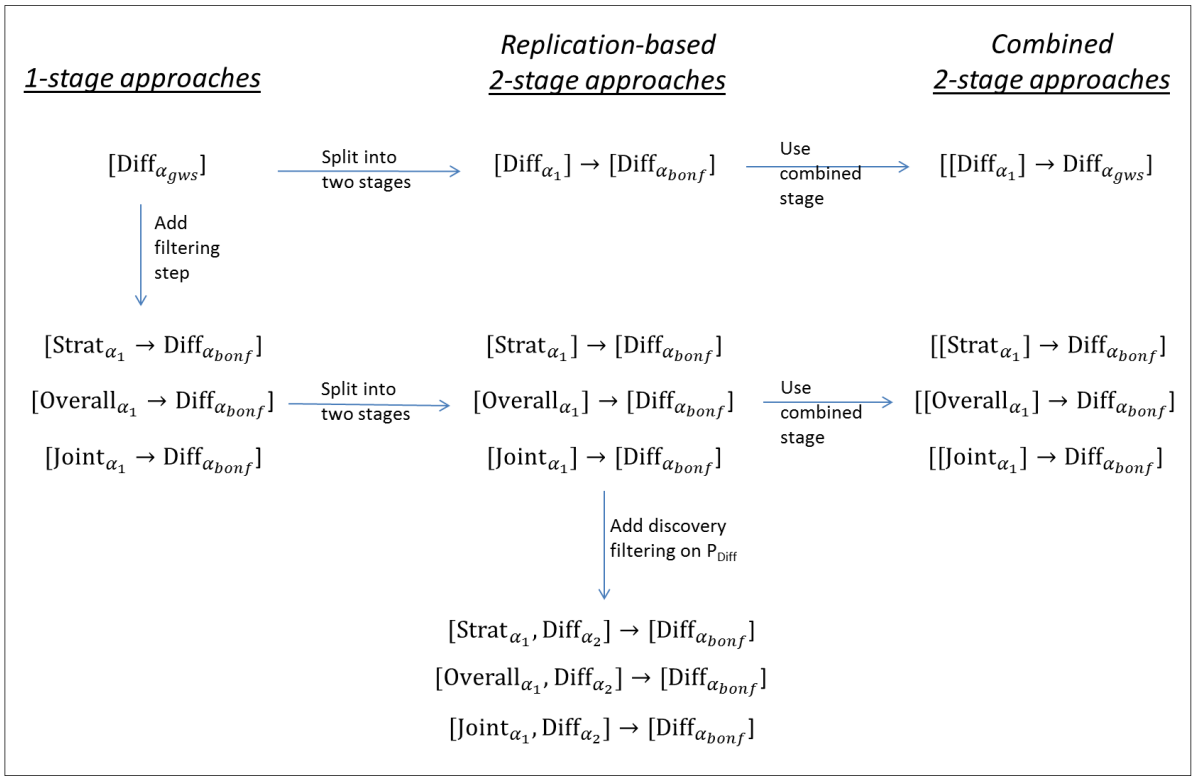


Figure 5. *Systematic scheme of stratified GWAMA approaches to identify stratum-difference (α_{gws}= 5 x 10^{-8}; α_1 = α-level for the filtering step; α_2 = α-level for the additional discovery filter on difference; α_{Bonf} = Bonferroni-corrected α-level).*

Generally, all of the considered approaches aim at using Bonferroni-corrected α-levels for the final difference test, α_{Bonf} = *0.05/M*, where *M* is the number of independent difference tests performed.

2. Screening for difference between two strata

The most intuitive approach is the 1-stage approach $[\mathrm{Diff}_{\alpha_{gws}}]$, which screens for difference at a genome-wide significance level, $\alpha_{Bonf} = \alpha_{gws} = 5 \times 10^{-8}$ (= $0.05/10^6$, Bonferroni-corrected for an approximate number of one million independent tests). The genome-wide significance level is a well-established screening threshold in GWAS of HapMap imputed data (Johnson et al., 2010).

Extending this intuitive approach to the 2-stage design, the difference test can either be implemented as a replication-based 2-stage approach $[\mathrm{Diff}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$ or as a combined 2-stage approach $[[\mathrm{Diff}_{\alpha_1}] \rightarrow \mathrm{Diff}_{\alpha_{gws}}]$. Due to employing a single statistical test (difference test), the design of the two approaches is similar to a general 2-stage GWAMA approach that screens for overall SNP association and for which the implementation into various 2-stage designs has been discussed before (Skol, Scott, Abecasis, & Boehnke, 2006). For the replication-based 2-stage approach $[\mathrm{Diff}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$, it is well known that significance for the final test can be attained using a Bonferroni-corrected α-level that is corrected for the independent number of SNPs tested (selected from the discovery data). For the combined (discovery + replication) 2-stage approach $[[\mathrm{Diff}_{\alpha_1}] \rightarrow \mathrm{Diff}_{\alpha_{gws}}]$, overlapping subjects between 'discovery stage SNP selection' and 'combined stage difference testing' are used, which is why the α-level of the final difference test has to be adjusted to the genome-wide significance level in order to yield valid type 1 error rates.

Further approaches are considered that involve initial filtering on stratified, overall or joint association tests. Their implementation into 1- and 2-stage designs has to be validated with regards to type 1 error and their impact on power to find stratum-difference has to be investigated.

For the replication-based 2-stage approaches, an additional discovery filter on difference is considered, e.g., for approach $[\mathrm{Joint}_{1e\text{-}5}\mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$, see **Figure 5**. The rationale behind this filter is that it increases power for the replication stage difference test by (i) lowering the multiple testing burden due to taking less SNPs forward, and (ii) by focusing on variants that are more likely to be truly dimorphic than variants that do not display any difference in discovery.

Practically, common stratified GWAMA projects aim at identification of any type of stratum-difference and it is likely that the optimal approach varies by type of stratum-difference. Thus, it is expected that multiple approaches have to be outlined in parallel in order to efficiently identify the various types of stratum-difference. In fact, this necessitates an additional multiple-testing correction for the final difference tests that have to be corrected for the number of screening approaches. Notably, each screening approach itself already employs a conservative Bonferroni-correction. Thus and in order to avoid overly conservative correction, such 'final' correction is ignored in the following.

2. Screening for difference between two strata

Technically, single genetic loci contain multiple correlated SNPs that are all located nearby within a specific region. To avoid overly conservative control of the final difference test, the filtered subsets of SNPs are clumped into independent regions and region-wide lead-SNPs are selected that are independent of other lead-SNPs (from other regions). Commonly used clumping criteria are LD-based (e.g., pairwise $r^2 > 0.2$, between SNPs of a specific region) or distance-based (e.g., *distance* < 500KB, pairwise between SNPs of a specific region). In the following, considerations are limited to a distance-based clumping criterion: *distance* < 500KB. The lead-SNP of a specific region is defined as the SNP with the lowest P-Value across all SNPs of the respective region. Importantly, the independent lead-SNPs are those that are put forward and are tested for difference in the final step. The total number of selected lead-SNPs is denoted as *M* in the following (as introduced before). To correct for the multiple difference testing of *M* independent lead SNPs, the α-level of the difference test undergoes a Bonferroni–correction, $\alpha_{Bonf} = 0.05/M$ (except for approaches $[\mathrm{Diff}_{\alpha_{gws}}]$ and $[[\mathrm{Diff}_{\alpha_1}] \rightarrow \mathrm{Diff}_{\alpha_{gws}}]$ that employ genome-wide significant α-levels for the final difference test).

2.1.5 Simulation-based evaluation of type 1 error

A simulation-based evaluation of type 1 error rates was performed for all of the defined approaches. Methodological details of the simulations are described in the following.

Simulated data sets were created that follow the null hypothesis of 'No difference in genetic effects between strata' ($H_0: b_1 = b_2$). More specifically, two versions of the null hypothesis were created: One assumes lack of stratum-specific effects ($H_0^{noeffect}: b_1 = b_2 = 0$), the other assumes identical (unequal zero) stratum-specific effects ($H_0^{effect}: b_1 = b_2 \neq 0$).

First, real genotypes were obtained for 1,500 men and 1,500 women from the KORA study (Wichmann, Gieger, Illig, & Group, 2005) for one well-imputed SNP: rs8138968 (MAF = 0.3).

Second, simulated sex-specific phenotypes (for the 1,500 men and 1,500 women) were created according to *Y~N(0,1)* (to reflect $H_0^{noeffect}$), and according to *Y|G=0~N(0,1)*, *Y|G=1~N($b_{80\%}$,1)* and *Y|G=2~N(2*$b_{80\%}$,1)* (to reflect H_0^{effect}). Herewith, $b_{80\%}$ corresponds to the minimum effect size detectable with 80% power by 1,500 samples (at α = 0.05), and *G=0*, *G=1* or *G=2* denote the group of individuals carrying 0, 1, or 2 minor alleles, respectively. Using G*Power, $b_{80\%}$ was estimated to be 0.111 for rs8138968 (Faul, Erdfelder, Buchner, & Lang, 2009).

Third, for each null hypothesis, the simulated phenotypes were sex-specifically tested for association with the real SNP genotypes. Men- and a women-specific genetic effect estimates, with standard errors and association P-Values, were derived.

Fourth, for each null hypothesis, the simulation and sex-specific association testing was repeated 1,000,000 times yielding 1,000,000 men- and women-specific genetic effect estimates, with standard errors and association P-Values.

Finally, for each null hypothesis, the defined 1-stage approaches were applied to the respective 1,000,000 sex-specific genetic effect estimates. Herewith, the α-level for the initial filtering was set to $\alpha_1 = 0.05$ (which corresponds to the estimation of $b_{80\%}$). For each approach, type 1 error rate (T1ER) of the final difference test was calculated by the proportion of nominally significant test results among all difference tests: $T1ER = \#(P_{Diff}<0.05)/M_{filt}$, where M_{filt} is the number of simulated data points passing the filtering (corresponding to the previously introduced *M*).

To reflect the full range of allele frequencies, the described procedure was repeated for two further SNPs: rs6002481 (MAF = 0.02), and rs6007738 (MAF = 0.5). Again, well-imputed real genotypes were taken from the KORA study. Again using G*Power, for the phenotype simulation of H_0^{effect}, $b_{80\%}$ was estimated to be 0.364 and 0.102, for rs6002481 and rs6007738, respectively.

To investigate the type 1 error of the 2-stage approaches, the 1,500 men and 1,500 women were split into two stages (yielding 750 men and 750 women for each stage), and the procedure was repeated by stage. Again using G*Power, for the phenotype simulation of H_0^{effect}, $b_{80\%}$ was estimated to be 0.514, 0.157 and 0.144, for rs6002481, rs8138968 and rs6007738, respectively. Herewith, $b_{80\%}$ reflects the minimum effect size that can be detected with 80% power by 750 samples (at $\alpha = 0.05$).

To investigate whether varying sample size between strata has a negative impact on type 1 error of the difference test, the procedure was repeated multiple times with fixed sample size in women (n_F = 1,500) and varying sample size in men (n_M = 150 to 1,500).

2.1.6 Analytical computation of power

Power comparisons aim at finding the best stratified GWAMA approach to identify stratum-difference. First, analytical power formulae were derived for the single statistical tests, i.e., for the difference, the stratified, the overall, and for the joint test. Second, power formulae for the approaches were obtained by combining the respective test-specific power formulae.

In order to model various types of stratum-difference, each formula was derived in dependence of stratum-specific explained variances R_1^2 and R_2^2. Given the genotypic variance σ_G^2, the stratum-specific phenotypic variance σ_Y^2 and the stratum-specific sample size n_i, the explained variance R_i^2 of a stratum-specific linear regression model is connected to the estimated stratum-specific genetic effect b_i (Rosner, 2006), by

$$R_i^2 = b_i^2 \frac{\sigma_G^2}{\sigma_Y^2} \quad (10).$$

For large sample sizes, it can be shown that the standard error of the genetic effect estimate is related to the explained variance (Rosner, 2006) by

$$se(b_i)^2 = \frac{\left(1 - R_i^2\right)\sigma_Y^2}{n_i \sigma_G^2} \quad (11).$$

Taking the direction of genetic effect into account, the absolute value of the stratum-specific explained variance is defined here, as

$$|R_i| = \begin{cases} -\sqrt{R_i^2}\,, \text{for } b_i \geq 0 \\ \sqrt{R_i^2}\,, \text{for } b_i < 0 \end{cases} \quad (12).$$

This notation is often referred to as 'effect size' in the following and is particularly useful to mathematically distinguish between the three general types of stratum-difference:

- $|R_1| \cdot |R_2| < 0$ denotes an OED effect
- $R_1^2 > 0 \ \wedge \ R_2^2 = 0$ denotes an SSE effect (here: effect in stratum 1)
- $R_1^2 > R_2^2 > 0 \wedge |R_1| \cdot |R_2| > 0$ denotes a CED effect (here: stronger effect in stratum 1)

To further distinguish between OED (or CED) effects, the notation OED_p (or CED_p) is introduced where *p* denotes the percentage of variance explained in stratum 2 compared to stratum 1, $p = R_2^2 \,/\, R_1^2$. For example, $OED_{25\%}$ denotes an effect that is opposite between strata and for which the variance explained in stratum 2 is a quarter of the size of the variance explained in stratum 1; $CED_{50\%}$ denotes an effect that is concordant between strata and for which the variance explained in stratum 2 is half the size of the variance explained in stratum 1.

2.1.6.1 Power formulae for the statistical tests

In the following, power formulae for the single statistical tests are presented. To allow modelling various scenarios of stage designs and various types of stratum-difference, each formula was derived in dependence of stratum-specific sample-sizes and in dependence of stratum-specific explained variances.

Difference test:

The difference test involves the z statistic Z_{Diff} (given by equation (5)). The power of the difference test is given by

$$Pwr_{Diff}(\alpha) = P\left(Z_{Diff} \leq z_{\frac{\alpha}{2}} | H_A\right) + P\left(Z_{Diff} \geq z_{1-\frac{\alpha}{2}} | H_A\right)$$
$$= \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{b_1 - b_2}{\sqrt{se(b_1)^2 - se(b_2)^2}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{b_1 - b_2}{\sqrt{se(b_1)^2 - se(b_2)^2}}\right) \quad (13),$$

where Φ denotes the cumulative standard normal distribution and z_q denotes the q-th quantile of Φ. Utilizing equations (10) and (11) yields:

$$Pwr_{Diff}(\alpha, R_1^2, R_2^2, n_1, f)$$

$$= \Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{n_1}\frac{|R_1| - |R_2|}{\sqrt{1 - R_1^2 + \frac{1}{f}(1 - R_2^2)}}\right)$$

$$+ \Phi\left(-z_{1-\frac{\alpha}{2}} + \sqrt{n_1}\frac{|R_1| - 1|R_2|}{\sqrt{1 - R_1^2 + \frac{1}{f}(1 - R_2^2)}}\right) \quad \textit{(14).}$$

Stratified test:

The stratified test involves two stratum-specific t tests that are performed simultaneously. For stratum *i*, the power of the t test is given by

$$Pwr_t(\alpha_i, n_i) = P\left(T \le t_{n_i-2,\frac{\alpha_i}{2}}|H_A\right) + P\left(T \ge t_{n_i-2,1-\frac{\alpha_i}{2}}|H_A\right)$$

$$= t_{n_i-2}\left(-t_{n_i-2,1-\frac{\alpha_i}{2}} - \frac{b_i}{se(b_i)}\right) + t_{n-2}\left(-t_{n_i-2,1-\frac{\alpha_i}{2}} + \frac{b_i}{se(b_i)}\right) \quad \textit{(15),}$$

where t_{n_i-2} is the cumulative distribution function of a t distribution with $n_i - 2$ df, $t_{n_i-2,q}$ is the q-th quantile of t_{n_i-2} and b_i and $se(b_i)$ are the genetic effect and standard error of stratum *i.* Utilizing equations (10) and (11) yields*:*

$$Pwr_t\left(\alpha_i, R_i^2, n_i\right) = t_{n_i-2}\left(-t_{n_i-2,1-\frac{\alpha_i}{2}} - \sqrt{\frac{n_i R_i^2}{1 - R_i^2}}\right) + t_{n_i-2}\left(-t_{n_i-2,1-\frac{\alpha_i}{2}} + \sqrt{\frac{n_i R_i^2}{1 - R_i^2}}\right) \quad \textit{(16).}$$

To obtain the power of the stratified test (combination of two stratum-specific t tests), the power formulae of the stratum-specific t tests have to be combined and the stratum-specific α-levels have to be corrected for the multiple testing of two stratum-specific t tests. Leveraging independence of stratum-specific t tests (due to using independent subjects), the power of the stratified test is given as follows:

$$Pwr_{Strat}(\alpha, R_1^2, R_2^2, n_1, f)$$

$$= Pwr_t\left(\frac{\alpha}{2}, R_1^2, n_1\right) + Pwr_t\left(\frac{\alpha}{2}, R_2^2, fn_1\right) - Pwr_t\left(\frac{\alpha}{2}, R_1^2, n_1\right) Pwr_t\left(\frac{\alpha}{2}, R_2^2, fn_1\right) \quad \textit{(17).}$$

Overall test:

The overall test involves the t test statistic $T = b_{Overall}/se(b_{Overall})$. The power formula for the overall test is based on using $b_{Overall}$, $se(b_{Overall})$ and $n_{Overall}$ in the power formula of the t test (given by equation (15)). Expressing the overall parameters in terms of stratum-specific parameters (see equation (8), and with $n_{Overall} = n_1 + f*n_1$) and utilizing equations (10) and (11) yields:

$$Pwr_{Overall}(\alpha, R_1^2, R_2^2, n_1, f) =$$

$$= t_{(n_1+fn_1)-2}\left(-t_{(n_1+fn_1)-2,1-\frac{\alpha}{2}} - \sqrt{n_1}\frac{\frac{|R_1|}{(1-R_1^2)} + \frac{f|R_2|}{(1-R_2^2)}}{\sqrt{\frac{1}{1-R_1^2} + \frac{f}{(1-R_2^2)}}}\right)$$

$$+ t_{(n_1+fn_1)-2}\left(-t_{(n_1+fn_1)-2,1-\frac{\alpha}{2}} + \sqrt{n_1}\frac{\frac{|R_1|}{(1-R_1^2)} + \frac{f|R_2|}{(1-R_2^2)}}{\sqrt{\frac{1}{1-R_1^2} + \frac{f}{(1-R_2^2)}}}\right) \quad \textit{(18).}$$

Joint test:

The joint test involves the chi-square test statistic C_{Joint} (see equation (9)). The power of the joint test is given by

$$Pwr_{Joint}(\alpha) = P\left(C_{Joint} \geq \chi_{2,1-\alpha} | H_A\right) = 1 - \mathrm{X}_{2,\lambda}^2\left(\chi_{2,1-\alpha}\right) \quad \textit{(19),}$$

where $\chi_{2,q}$ is the q-th quantile of a chi-square distribution with 2 df and $\mathrm{X}_{2,\lambda}^2$ is the cumulative distribution function of a non-central chi-square distribution with 2 df and with non-centrality parameter λ that can be calculated as follows: $\lambda = \left(\frac{b_1}{se(b_1)}\right)^2 + \left(\frac{b_2}{se(b_2)}\right)^2$. Utilizing equations (10) and (11) yields

$$\lambda(R_1^2, R_2^2, n_1, f) = \frac{n_1 R_1^2}{1-R_1^2} + \frac{f n_1 R_2^2}{1-R_2^2} \quad \textit{(20).}$$

Thus, the power for the joint test is given as follows:

$$Pwr_{Joint}(\alpha, R_1^2, R_2^2, n_1, f) = 1 - \mathrm{X}_{2,\lambda\left(R_1^2, R_2^2, n_1, f\right)}^2\left(\chi_{2,1-\alpha}\right). \quad \textit{(21).}$$

2.1.6.2 Power formulae for the approaches

To obtain the power formula for an approach, the derived power formulae for the single statistical tests (i.e., steps implemented in the approach) have to be combined.

Given independence of steps, the power of an approach is a product of the power of the implemented steps:

$$Pwr_{Approach} = Pwr_{Step1 \cap Step2 \cap \ldots \cap StepN} = \prod_{i=1}^{N} Pwr_{Step\ i} \quad \textit{(22).}$$

For dependent steps, the power for the intersection of dependent steps has to be obtained from simulated probability distributions.

Independence of steps can either be given through using independent data sets per step or through statistical independence of tests involved. Due to using independent data sets for filtering and difference testing, independence of steps is obvious for any replication-based approach that does not include the additional discovery filtering step on difference,

e.g., $[\text{Joint}_{1e\text{-}5}] \rightarrow [\text{Diff}_{\alpha_{\text{Bonf}}}]$. In contrast, all replication-based 2-stage approaches that include the additional discovery filtering on difference, e.g., $[\text{Joint}_{1e\text{-}5}, \text{Diff}_{0.05}] \rightarrow [\text{Diff}_{\alpha_{\text{Bonf}}}]$, require statistical independence of the two discovery steps in order to utilize equation (22) for the discovery-stage power calculation. Due to using overlapping data sets for filtering and difference testing, all 1-stage and combined 2-stage approaches require statistical independence of steps in order to utilize equation (22).

To evaluate statistical independence or dependence of steps, the simulated data sets (initially created for the type 1 error evaluation of 1-stage approaches) were utilized: Assuming the null hypothesis of 'No difference in genetic effects between strata' (as implied by the simulated data sets), applying a filter that is dependent of the difference test would yield a deflated or inflated distribution of the difference test statistic. Vice versa, if no impact on the distribution of the difference test can be observed, then the difference and the filtering test can be considered as independent. The impact of filtering tests on the distribution of the difference test statistic was investigated using QQ plots.

2.1.6.3 Details of the power computations

Analytical power formulae were derived for all considered approaches. In order to compute the power, assumptions about stratum-specific sample sizes, explained variances and α-levels have to be made. Generally, parameters were defined to reflect realistic power computation scenarios that are similar to GIANT consortium stratified GWAMA scenarios.

The total overall (strata-combined) sample size was fixed at 200,000 individuals. This is similar to the total sample sizes involved in sex-stratified or smoking-status stratified GWAMA projects of the GIANT consortium. The 2-stage approaches were defined to employ a balanced stage design, with each stage containing 100,000 individuals.

Different types of stratum-specific genetic effects were modelled based on known and realistic genetic effect sizes taken from GIANT consortium GWAMAs on WHR_{adjBMI}: A small (rs6784615 near *STAB1*), a medium (rs4684854 near *PPARG*) and a large (rs2820443 near *LYPLAL1*) genetic effect on WHR_{adjBMI} that explained 0.014%, 0.058% and 0.167% of the WHR_{adjBMI} variation in women, respectively (Heid et al., 2010; Randall et al., 2013).

The α-levels for the filtering tests were arbitrarily set to $\alpha_1 = 1 \times 10^{-5}$. For approach $[\text{Overall}_{\alpha_1} \rightarrow \text{Diff}_{\alpha_{\text{Bonf}}}]$, α_1 was further set to genome-wide significance, $\alpha_1 = 5 \times 10^{-8}$, because this reflects the common design of many overall GWAMA projects. For the replication-based 2-stage approaches that involve additional discovery filtering on difference, the α-level of the discovery difference filter was arbitrary set to nominal significance ($\alpha_2 = 0.05$).

Finally, in order to obtain the Bonferroni-corrected α-level ($\alpha_{\text{Bonf}} = 0.05/M$) for the final difference test, the number of independent lead-SNPs tested for difference (=M) had to be estimated. For the sex-stratified GWAMA scenario, M was estimated from the real sex-

stratified GIANT GENDER data set for WHR_{adjBMI} (Randall et al., 2013) (see chapter 4.1.2 for details about the data set). For the general stratified GWAMA scenario, *M* was estimated from the real GIANT SMOKING data set for WHR_{adjBMI} that comprises unbalanced stratum sizes (*f* = 3.483 or *f* = 0.287) (Justice et al, in progress) (see chapter 4.1.3 for details about the data set).

2.2 Results

The following chapters show results from the structured methodological comparison of stratified GWAMA approaches to identify stratum-difference. First, approaches were validated with regards to type 1 error (using simulation, chapter 2.2.1) and second, approaches were compared with regards to power (using analytical computations, chapter 2.1.6).

2.2.1 Simulation-based evaluation of type 1 error

To evaluate approaches with regards to type 1 error, simulation-based estimations of type 1 error rates were performed. First, type 1 error was inferred for all approaches and stage designs under the balanced sex-stratified GWAMA setting (chapter 2.2.1.1). Valid 1-stage sex-stratified GWAMA approaches were transferred to the general setting and the impact of unbalanced strata sizes was evaluated (chapter 2.2.1.2).

Prior to investigating type 1 error of the approaches, the simulated data sets were validated, i.e., were shown to truly reflect the implied null hypotheses of 'No difference in genetic effects between strata' (see **Appendix 9.1**).

2.2.1.1 Type 1 error for the sex-stratified GWAMA approaches

Generally, the type 1 error rate (T1ER) of an approach was estimated by calculating the number of nominally significant difference test results among all conducted difference tests. Assuming the null hypothesis of 'No difference in genetic effects between strata' and a 5% α-level, one would expect that 5% of difference tests reach nominal significance. Increased numbers of nominally significant test results pinpoint invalid approaches.

For the 1-stage approaches, surprisingly, filtering the data-set for overall association did not inflate the T1ER of the difference test (for any modelled scenarios, **Table 1**).

In contrast, filtering the data set for stratified or joint association strongly inflated the T1ER of the difference test (for all modelled scenarios, **Table 1**). Thus, for SNPs that are filtered on stratified or joint association, the difference cannot be established within the same

data set. To separate filtering from difference testing, such approaches require independent data sets, discovery (for filtering) and replication (for difference testing).

All of the replication-based 2-stage approaches yielded valid type 1 error rates (T1ER ranged from 4.57% to 5.44% across approaches and modelled scenarios, **Table 1**). Thus (as expected) all replication-based 2-stage approaches, including the ones applying additional discovery filtering on difference, can be considered as valid. The reason for this is that the filtering in discovery samples cannot bias the difference test statistic when this is calculated using independent replication subjects.

For the combined 2-stage approaches, all but one approach displayed an elevated T1ER for the final difference test (e.g., T1ER > 21.2% under $H_0^{noeffect}$, **Table 1**). Only the combined 2-stage approach $[[\mathrm{Overall}_{\alpha_1}]\rightarrow\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ yielded valid T1ER that ranged from 4.99 to 5.34% across modelled scenarios.

As expected, due to violating type 1 error requirements, the combined 2-stage approach $[[\mathrm{Diff}_{\alpha_1}]\rightarrow\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ cannot adopt a Bonferroni-corrected α-level for the final difference test ($\alpha_{\mathrm{Bonf}} = 0.05/M$, corrected for the number of SNPs tested). Nevertheless, adjusting the α-level to genome-wide significance ($\alpha_{\mathrm{gws}} = 5 \times 10^{-8}$) reflects a valid approach that may even be more efficient than adopting the analogous replication-based 2-stage design while obtaining significance from a Bonferroni-corrected α-level (Skol et al., 2006).

Table 1. *Simulation-based type 1 error rates (T1ER) for balanced sex-stratified GWAMA approaches to identify sex-difference.*

			$H_0^{noeffect}$			H_0^{effect}		
Approach	**SNP**	**MAF**	**#Diff-tests[a]**	**#Diff <0.05[b]**	**T1ER** [%]	**#Diff-tests[a]**	**#Diff <0.05[b]**	**T1ER** [%]
1-stage approaches								
[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	1,000,000	50,143	5.01	1,000,000	49,950	5.00
	rs8138968	0.3	1,000,000	50,115	5.01	1,000,000	49,933	4.99
	rs6007738	0.5	1,000,000	50,063	5.01	1,000,000	49,762	4.98
[$\text{Overall}_{0.05}\rightarrow\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,076	2,556	5.10	969,452	48,473	5.00
	rs8138968	0.3	50,053	2,491	4.98	975,179	48,657	4.99
	rs6007738	0.5	50,390	2,542	5.04	977,003	48,591	4.97
[$\text{Strat}_{0.05}\rightarrow\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	49,857	21,363	**42.85**	898,504	49,861	5.55
	rs8138968	0.3	49,660	21,257	**42.81**	911,957	49,843	5.47
	rs6007738	0.5	49,493	20,920	**42.27**	915,989	49,692	5.42
[$\text{Joint}_{0.05}\rightarrow\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,328	25,128	**49.93**	939,251	49,807	5.30
	rs8138968	0.3	50,215	25,025	**49.84**	949,404	49,810	5.25
	rs6007738	0.5	50,163	24,771	**49.38**	952,489	49,649	5.21
Replication-based 2-stage approaches								
[$\text{Diff}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,205	2,565	5.11	50,074	2,563	5.12
	rs8138968	0.3	50,517	2,533	5.01	50,395	2,554	5.07
	rs6007738	0.5	50,352	2,534	5.03	49,889	2,485	4.98
[$\text{Overall}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,418	2,410	4.78	964,698	49,069	5.09
	rs8138968	0.3	50,554	2,625	5.19	978,228	49,072	5.02
	rs6007738	0.5	50,983	2,571	5.04	978,496	48,673	4.97
[$\text{Strat}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	49,699	2,447	4.92	892,319	45,443	5.09
	rs8138968	0.3	49,841	2,607	5.23	918,662	46,163	5.03
	rs6007738	0.5	50,443	2,500	4.96	920,233	45,804	4.98
[$\text{Joint}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,158	2,477	4.94	931,854	47,403	5.09
	rs8138968	0.3	50,798	2,625	5.17	954,608	47,932	5.02
	rs6007738	0.5	51,098	2,528	4.95	955,261	47,515	4.97
[$\text{Overall}_{0.05}$,$\text{Diff}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	2,541	116	4.57	48,336	2,483	5.14
	rs8138968	0.3	2,443	133	5.44	49,279	2,486	5.04
	rs6007738	0.5	2,582	126	4.88	48,817	2,423	4.96
[$\text{Strat}_{0.05}$,$\text{Diff}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	21,345	1,088	5.10	49,942	2,555	5.12
	rs8138968	0.3	21,319	1,080	5.07	50,324	2,548	5.06
	rs6007738	0.5	21,488	1,049	4.88	49,831	2,479	4.97
[$\text{Joint}_{0.05}$,$\text{Diff}_{0.05}$]→[$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	24,902	1,266	5.08	49,897	2,553	5.12
	rs8138968	0.3	25,262	1,269	5.02	50,290	2,546	5.06
	rs6007738	0.5	25,261	1,229	4.87	49,794	2,479	4.98
Combined 2-stage approaches								
[[$\text{Diff}_{0.05}$]→$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,205	15,775	**31.42**	50,074	15,731	**31.42**
	rs8138968	0.3	50,517	17,225	**34.10**	50,395	17,230	**34.19**
	rs6007738	0.5	50,352	17,398	**34.55**	49,889	17,232	**34.54**
[[$\text{Overall}_{0.05}$]→$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,418	2,675	5.31	964,698	48,580	5.04
	rs8138968	0.3	50,554	2,659	5.26	978,228	49,076	5.02
	rs6007738	0.5	50,983	2,650	5.20	978,496	49,135	5.02
[[$\text{Strat}_{0.05}$]→$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	49,699	10,533	**21.19**	892,319	48,224	5.40
	rs8138968	0.3	49,841	11,331	**22.73**	918,662	49,056	5.34
	rs6007738	0.5	50,443	11,548	**22.89**	920,233	49,157	5.34
[[$\text{Joint}_{0.05}$]→$\text{Diff}_{\alpha_{diff}}$]	rs6002481	0.05	50,158	11,455	**22.84**	931,854	48,594	5.21
	rs8138968	0.3	50,798	12,353	**24.32**	954,608	49,149	5.15
	rs6007738	0.5	51,098	12,517	**24.50**	955,261	49,216	5.15

[a] Number of SNPs tested for difference; [b] Number of SNPs with nominal significant difference ($P_{Diff} < 0.05$)

2.2.1.2 Type 1 error for the unbalanced stratified GWAMA approaches

Next, the impact of unbalanced strata on the type 1 error of the final difference test was investigated. Therefore, T1ER of the 1-stage approaches $[\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ and $[\mathrm{Overall}_{\alpha_1} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ (shown to be valid under the 1-stage sex-stratified GWAMA configuration) was evaluated based on unbalanced simulated data sets.

Interestingly, employing an unbalanced strata design did not affect T1ER of the difference test (**Table 2**). Across all scenarios, the T1ER of the difference tests for the two considered 1-stage approaches remained low and ranged from 4.89% to 5.27%.

Table 2. *Simulation-based type 1 error rates (T1ER) for unbalanced stratified GWAMA approaches to identify stratum-difference. Shown is the T1ER of 1-stage approaches (validated under the sex-stratified scenario) for varying combinations of stratum-specific sample sizes.*

			$H_0^{noeffect}$			H_0^{effect}		
Approach	**n_F**	**n_M**	**#Diff-Tests[a]**	**#Diff <0.05[b]**	**T1ER** [%]	**#Diff-Tests[a]**	**#Diff <0.05[b]**	**T1ER** [%]
$[\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$	1,500	1,500	1,000,000	50,029	5.00	1,000,000	50,197	5.02
	1,500	1,250	1,000,000	50,046	5.00	1,000,000	50,394	5.04
	1,500	1,000	1,000,000	50,171	5.02	1,000,000	49,919	4.99
	1,500	750	1,000,000	50,121	5.01	1,000,000	50,257	5.03
	1,500	500	1,000,000	50,499	5.05	1,000,000	49,914	4.99
	1,500	250	1,000,000	50,875	5.09	1,000,000	50,620	5.06
$[\mathrm{Overall}_{0.05} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$	1,500	1,500	50,473	2,502	4.96	975,387	48,923	5.02
	1,500	1,250	50,199	2,532	5.04	965,813	48,652	5.04
	1,500	1,000	50,617	2,474	4.89	947,023	47,169	4.98
	1,500	750	50,428	2,531	5.02	925,791	46,560	5.03
	1,500	500	50,453	2,554	5.06	896,195	44,636	4.98
	1,500	250	50,598	2,634	5.21	855,596	43,215	5.05

[a] Number of SNPs tested for difference (i.e., number of SNPs passing the filtering steps, M_{filt})
[b] Number of SNPs with nominal significant difference ($P_{Diff} < 0.05$)

2.2.2 Analytical power comparison

To compare performance of approaches and to give recommendations for different stage designs and for different types of stratum-difference, analytical power computations were performed for various realistic scenarios.

Generally, approaches that were shown to violate type 1 error requirements were omitted from power computations. Also, the replication-based 2-stage approach $[\mathrm{Overall}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ was omitted from the power computations because of its obvious power

disadvantage compared to the combined 2-stage approach $[[\mathrm{Overall}_{\alpha_1}] \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$: Both approaches involve the same data set for the discovery filtering, but the latter comprises a larger sample size (thus larger power) for the difference testing. Finally, because of the obvious power gain through prioritizing variants with nominally significant difference at discovery stage for the replication-based 2-stage approaches $[\mathrm{Strat}_{\alpha_1}\mathrm{Diff}_{\alpha_2}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ and $[\mathrm{Joint}_{\alpha_1}\mathrm{Diff}_{\alpha_2}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ compared to $[\mathrm{Strat}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ and $[\mathrm{Joint}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$, the latter two were also omitted from the power computations.

A systematic scheme of approaches that were compared for power is shown in **Figure 6**.

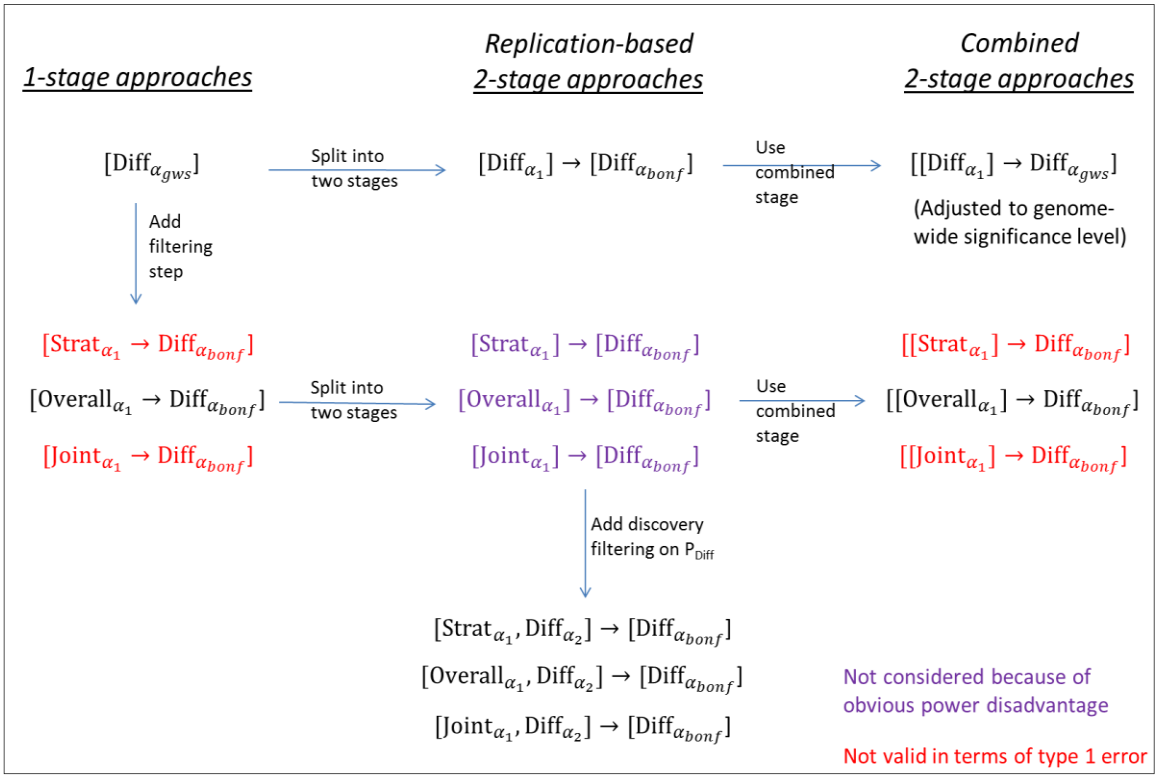


Figure 6. *Systematic scheme of stratified GWAMA approaches to identify stratum-difference considered for power comparisons (α_{gws}= 5 x 10^{-8}; α_1 = α-level for the filtering step; α_2 = α-level for the additional discovery filter on difference; α_{Bonf} = Bonferroni-corrected α-level).*

Simulation-based inference of statistical dependence between tests showed that the difference test is independent of the overall test, but dependent of the stratified test and dependent of the joint test (see **Appendix 9.2**). This is in-line with what has already been seen for the type 1 error (see chapter 2.2.1.1). As a consequence, the discovery power of the

approach $[\mathrm{Strat}_{1e\text{-}5},\mathrm{Diff}_{0.05}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and of the approach $[\mathrm{Joint}_{1e\text{-}5},\mathrm{Diff}_{0.05}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ had to be estimated from simulated probability distributions (see **Appendix 9.3**).

For all other approaches, the power could be calculated using the derived step-specific power formulae with equation (22). A list of applied power formulae including details about step-specific α-levels and sample sizes is given in **Table 3.**

Power comparisons were first performed under the sex-stratified GWAMA setting (comprising balanced stratum sizes, chapter 2.2.2.1) and then extended to the general setting (to investigate the impact of unbalanced stratum sizes, chapter 2.2.2.2).

Table 3. *Analytical power formulae for the considered 1- and 2-stage sex-stratified GWAMA approaches. To calculate the Bonferroni-corrected α-level for the final difference test (α_{Bonf} = 0.05/M) of approaches that involve filtering steps, the number of independent lead SNPs tested (=M) was estimated from the real GIANT GENDER data set (for the sex-stratified GWAMA scenario) and from the real GIANT SMOKING data set (for the general stratified GWAMA scenario).*

Design	Approach	M (GENDER)	M (SMOKING)	Analytical power formula
1-stage	$[\text{Diff}_{5e\text{-}8}]$	-	-	$Pwr_{Diff}(\text{5e-8}, R_1^2, R_2^2, n_1, f)$
	$[\text{Overall}_{1e\text{-}5} \rightarrow \text{Diff}_{\alpha_{\text{Bonf}}}]$	116	70	$Pwr_{Overall}(\text{1e-5}, R_1^2, R_2^2, n_1, f) \cdot Pwr_{Diff}(\alpha_{bonf}, R_1^2, R_2^2, n_1, f)$
	$[\text{Overall}_{5e\text{-}8} \rightarrow \text{Diff}_{\alpha_{\text{Bonf}}}]$	39	18	$Pwr_{Overall}(\text{5e-8}, R_1^2, R_2^2, n_1, f) \cdot Pwr_{Diff}(\alpha_{bonf}, R_1^2, R_2^2, n_1, f)$
2-stage replication-based	$[\text{Diff}_{1e\text{-}5}] \rightarrow [\text{Diff}_{\alpha_{\text{Bonf}}}]$	19	10	$Pwr_{Diff}\left(\text{1e-5}, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}\left(\alpha_{bonf}, R_1^2, R_2^2, \frac{n_1}{2}, f\right)$
	$[\text{Overall}_{1e\text{-}5}, \text{Diff}_{0.05}] \rightarrow [\text{Diff}_{\alpha_{\text{Bonf}}}]$	12	10	$Pwr_{Overall}\left(\text{1e-5}, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}\left(0.05, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}\left(\alpha_{bonf}, R_1^2, R_2^2, \frac{n_1}{2}, f\right)$
	$[\text{Strat}_{1e\text{-}5}, \text{Diff}_{0.05}] \rightarrow [\text{Diff}_{\alpha_{\text{Bonf}}}]$	18	20	$Pwr_{Strat \cap Diff}\left(\text{1e-5}, 0.05, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}\left(\alpha_{bonf}, R_1^2, R_2^2, \frac{n_1}{2}, f\right)$
	$[\text{Joint}_{1e\text{-}5}, \text{Diff}_{0.05}] \rightarrow [\text{Diff}_{\alpha_{\text{Bonf}}}]$	19	18	$Pwr_{Joint \cap Diff}\left(\text{1e-5}, 0.05, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}\left(\alpha_{bonf}, R_1^2, R_2^2, \frac{n_1}{2}, f\right)$
2-stage combined	$[[\text{Diff}_{1e\text{-}5}] \rightarrow \text{Diff}_{5e\text{-}8}]$	-	-	$Pwr_{Diff}\left(\text{1e-5}, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}(\text{5e-8}, R_1^2, R_2^2, n_1, f)$
	$[[\text{Overall}_{1e\text{-}5}] \rightarrow \text{Diff}_{\alpha_{\text{Bonf}}}]$	43	54	$Pwr_{Overall}\left(\text{1e-5}, R_1^2, R_2^2, \frac{n_1}{2}, f\right) \cdot Pwr_{Diff}(\alpha_{bonf}, R_1^2, R_2^2, n_1, f)$

2.2.2.1 Analytical power comparison for the sex-stratified GWAMA approaches

In the following chapter, results from power comparisons between multiple sex-stratified GWAMA approaches (to identify sex-difference) are presented. The sex-stratified GWAMA scenario assumed a total sample size of 200,000 that is equally split among strata (f = 1), $n_1 = n_2 = 100{,}000$.

To find the best approaches based on stage design, power was first compared between 1-stage approaches (chapter 2.2.2.1.1) and then compared between 2-stage approaches (chapter 2.2.2.1.2). Finally, to contrast stage designs, the recommended 1- and 2-stage approaches were compared (chapter 2.2.2.1.3).

2.2.2.1.1 Power for the 1-stage sex-stratified GWAMA approaches

First, power was compared between the 1-stage approaches $[\mathrm{Diff}_{5e\text{-}8}]$, $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Overall}_{5e\text{-}8} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ for a fixed effect size in women (defined as stratum 1; set to one of three realistic $\mathrm{WHR}_{\mathrm{adjBMI}}$ effects) and for varying effect sizes in men (defined as stratum 2, **Figure 7**). This strategy allowed for investigating several types of sexually dimorphic effects, OED, $\mathrm{OED}_{50\%}$, $\mathrm{OED}_{25\%}$, SSE, $\mathrm{CED}_{25\%}$ and $\mathrm{CED}_{50\%}$ effects.

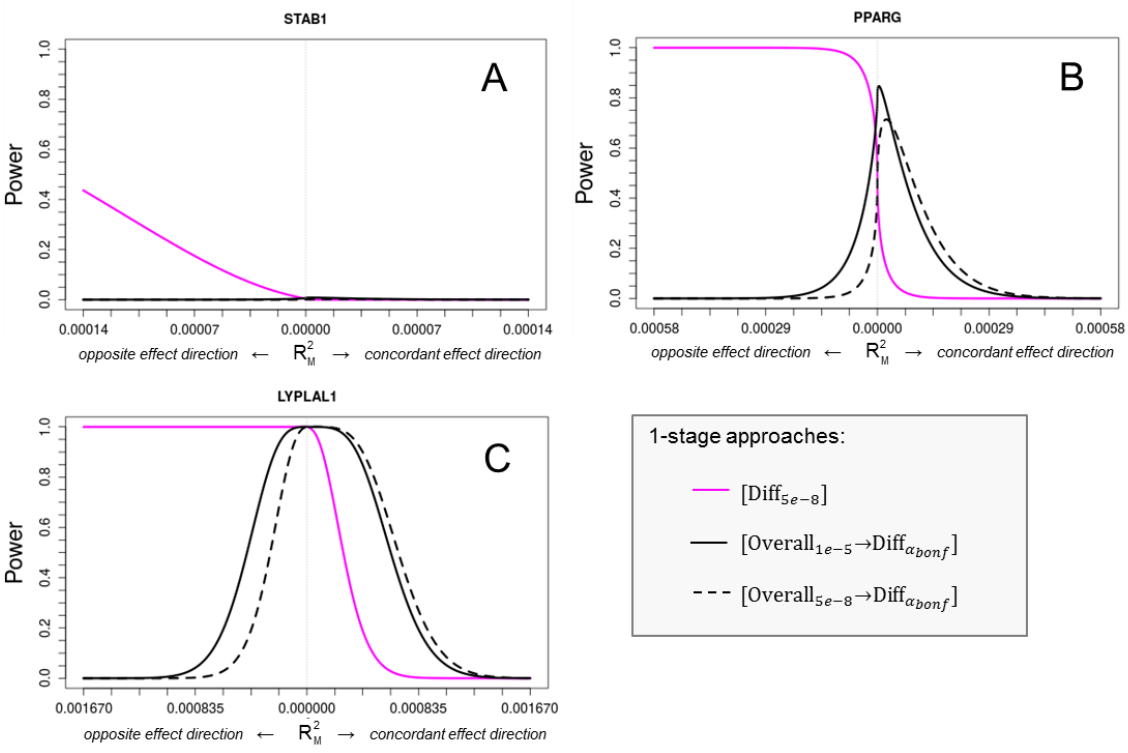


Figure 7. Power of the 1-stage approaches. *Shown is the power to detect sex-difference by the three considered 1-stage approaches. The sex-difference is modelled by varying the effect size in men (varied on x-axis) and by fixing the effect size in women that is set to a known small, medium and strong genetic effect, comparable to the known WHR_{adjBMI} effects near A: STAB1, B: PPARG and C: LYPLAL1, respectively.*

As expected, the power to detect the sex-difference generally increases with the size of the modelled effect in women, which is reflected by the increasing area under the power curves when viewing panels A, B and C (**Figure 7A-C**).

For a small modelled women effect (comparable to the genetic effect on WHR_{adjBMI} near *STAB1,* $R_F^2 = 0.014\%$) approach $[\mathrm{Diff}_{5e\text{-}8}]$ has some power for extreme OED scenarios and approaches $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Overall}_{5e\text{-}8} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ completely lack power (**Figure 7A**). Neither small SSE nor small CED types of effects are detectable by any of the approaches.

For a medium modelled women effect (comparable to the genetic effect on WHR_{adjBMI} near *PPARG,* $R_F^2 = 0.058\%$), approach $[\mathrm{Diff}_{5e\text{-}8}]$ has strong power to identify any kind of OED effect, but breaks down once the effect in men turns towards the direction of effect in women (**Figure 7B**). In contrast, approaches $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Overall}_{5e\text{-}8} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ lack power for OED scenarios, but have sufficient or at least some power to find SSE or CED effects.

For a strong modelled women effect (comparable to the genetic effect on WHR_{adjBMI} near *LYPLAL1,* $R_F^2 = 0.167\%$), the area under the power curve of approach $[\mathrm{Diff}_{5e\text{-}8}]$ broadens (**Figure 7C**). Besides the large power for OED effects, approach $[\mathrm{Diff}_{5e\text{-}8}]$ displays increased power for SSE and CED effects. The power of the two approaches $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Overall}_{5e\text{-}8} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ increased for SSE effects as well as for CED effects.

The lack of power for OED effects is expected for approaches that involve a-priori filtering on overall association. This is because the opposite sex effects are missed by the overall (sex-combined) test.

Interestingly however, the initial filtering on overall association boosts the identification of SSE and CED signals. This is particularly interesting because such effects may even be the biologically most plausible types of sexually dimorphic effects. Thus, the commonly used GWAMA approach - that is to screen for genome-wide significant overall association and to follow up identified loci for sex-difference - generally performs well to find SSE and CED type of effects.

Approach $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ has higher power for SSE effects than approach $[\mathrm{Overall}_{5e\text{-}8} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$, e.g., 80.7% compared to 46.7% power for the medium SSE effect. This suggests that a less stringent control of the overall association yields higher power for real SSE signals. In contrast, the more stringent the control on overall association, the greater is the probability to find real CED effects, e.g., 84.9% compared to 90.0% power for the large $CED_{25\%}$ effect and for $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Overall}_{5e\text{-}8} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$, respectively. Due to the larger power gain of $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ for SSE signals compared to the small

power gain of $[\text{Overall}_{5e\text{-}8} \rightarrow \text{Diff}_{\alpha_{\text{Bonf}}}]$ for CED signals, a weaker control of the overall association ($P_{Overall}$ < 1 x 10^{-5}) can generally be recommended (in order to keep down the number of recommended approaches). Only if a study specifically aims at the identification of CED loci, a stronger control of the overall association ($P_{Overall}$ < 5 x 10^{-8}) would be beneficial.

In summary, to cover the whole range of sexually dimorphic effects, screening for sex-difference genome-wide ($[\text{Diff}_{5e\text{-}8}]$, for OED) in combination with the 'liberal overall filtering' approach ($[\text{Overall}_{1e\text{-}5} \rightarrow \text{Diff}_{\alpha_{\text{Bonf}}}]$, for SSE and CED) can be recommended.

2.2.2.1.2 Power for the 2-stage sex-stratified GWAMA approaches

Next, power was compared between the six considered 2-stage approaches. Again, the power was depicted over a varying effect in men and set to a small, medium or large genetic effect in women (comparable to known effects on WHR_{adjBMI} near *STAB1*, *PPARG* and *LYPLAL1*, respectively, **Figure 8**).

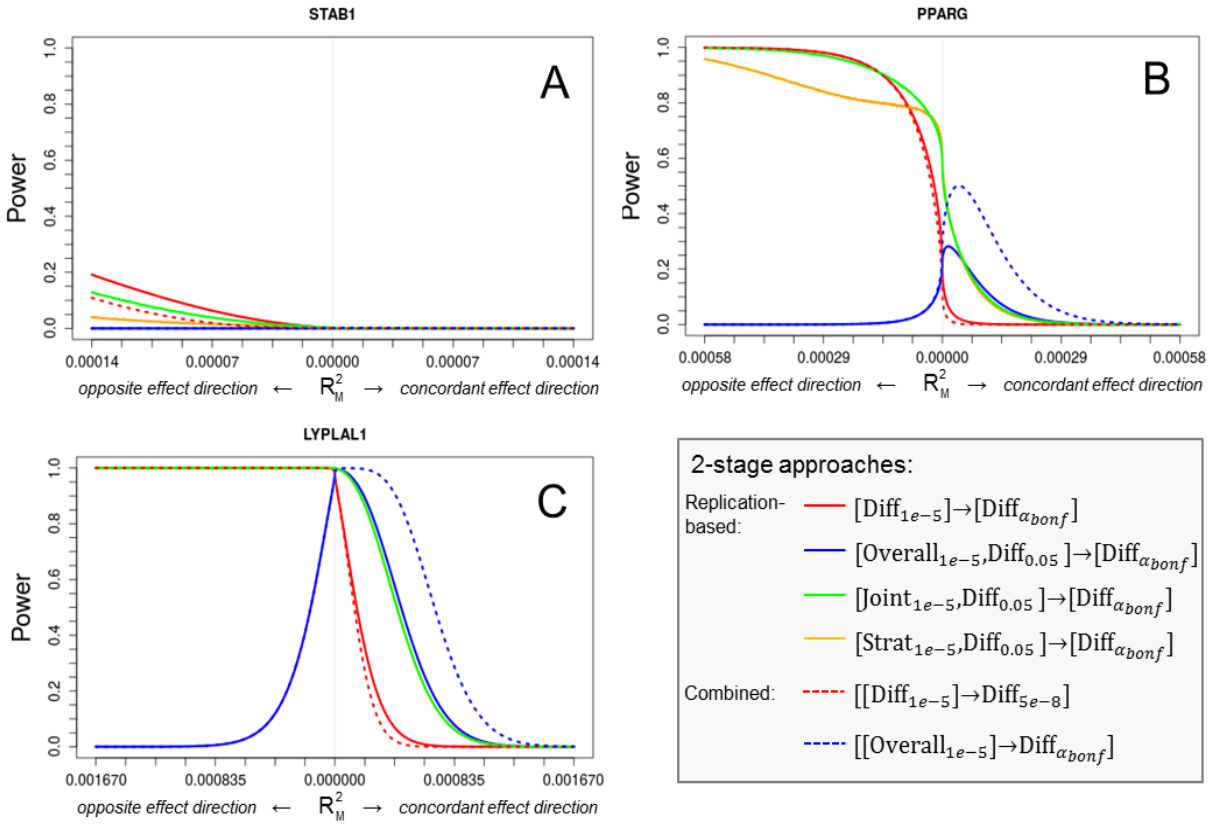


***Figure 8. Power of the 2-stage approaches**. Shown is the power to detect the sex-difference by the replication-based 2-stage (solid lines) or by the combined 2-stage (dotted lines) approaches. The sex-difference is modelled by varying the effect size in men (varied on x-axis) and by fixing the effect size in women to a known small, medium and large genetic effect, comparable to the known WHR_{adjBMI} effects near A: STAB1, B: PPARG and C: LYPLAL1, respectively. In panel C, the orange line is completely hidden by the green line.*

2. Screening for difference between two strata

The behavior of the replication-based and the combined 2-stage approaches $[\text{Diff}_{1e\text{-}5}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ and $[[\text{Diff}_{1e\text{-}5}]\rightarrow\text{Diff}_{5e\text{-}8}]$ is comparable to the 1-stage approach $[\text{Diff}_{5e\text{-}8}]$. They are well powered to identify medium OED effects and break down once the effect direction in men turns positive (**Figure 8B**).

Interestingly, the replication-based 2-stage approach $[\text{Diff}_{1e\text{-}5}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ and the combined 2-stage approach $[[\text{Diff}_{1e\text{-}5}]\rightarrow\text{Diff}_{5e\text{-}8}]$ perform almost identical. This suggests that testing sex-difference at a more relaxed α-level (α_{Bonf} = 0.05/*M* for approach $[\text{Diff}_{1e\text{-}5}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ versus $\alpha_{\text{gws}} = 5 \times 10^{-8}$ for approach $[[\text{Diff}_{1e\text{-}5}]\rightarrow\text{Diff}_{5e\text{-}8}]$) compensates for the fewer subjects used for the final sex-difference test.

The behavior of the replication-based 2-stage approach $[\text{Overall}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ and the combined 2-stage approach $[[\text{Overall}_{1e\text{-}5}]\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ is comparable to the respective 1-stage approach $[\text{Overall}_{1e\text{-}5}\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ (**Figure 8**). They lack power for medium OED effects and display increased power for SSE and CED effects. Notably, for CED signals, the power of the combined 2-stage approach $[[\text{Overall}_{1e\text{-}5}]\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ clearly surpasses the power of the replication-based 2-stage approach $[\text{Overall}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$. This suggests that the increased sample size for the final sex-difference test, i.e., using the combined discovery + replication data for approach $[[\text{Overall}_{1e\text{-}5}]\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ compared to using replication data only for approach $[\text{Overall}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$, outweighs the additional prioritization on sexually dimorphic SNPs in the discovery filtering of approach $[\text{Overall}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$.

Although the approaches involving the filtering on stratified or joint association, $[\text{Strat}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ and $[\text{Joint}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$, similarly lack power for small OED effects, they seem to be particularly advantageous for medium SSE effects (**Figure 8**). Notably, the approach $[\text{Strat}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ has lower power in the OED range compared to the approach $[\text{Joint}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$.

In summary, if a 2-stage design has to be adopted, the replication-based approach $[\text{Joint}_{1e\text{-}5},\text{Diff}_{0.05}]\rightarrow[\text{Diff}_{\alpha_{\text{Bonf}}}]$ (for OED and SSE) in combination with the combined 2-stage approach $[[\text{Overall}_{1e\text{-}5}]\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ (for CED) can be recommended.

2.2.2.1.3 Power comparison between the best 1- and 2-stage sex-stratified GWAMA approaches

As demonstrated, for a sex-stratified GWAMA project aiming at the identification of any type of sexually dimorphic effect and adopting a 1-stage design, a combination of approaches $[\text{Diff}_{5e\text{-}8}]$ (for OED) and $[\text{Overall}_{1e\text{-}5}\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ (for SSE and CED) is recommended. If a 1-stage study specifically aims at CED loci, then the 1-stage $[\text{Overall}_{5e\text{-}8}\rightarrow\text{Diff}_{\alpha_{\text{Bonf}}}]$ approach

involving a stronger control of the overall association is preferable. For a 2-stage design, the combination of the replication-based 2-stage approach [$\mathrm{Joint}_{1e\text{-}5}$,$\mathrm{Diff}_{0.05}$]→[$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] (for OED and SSE) and the combined 2-stage approach [[$\mathrm{Overall}_{1e\text{-}5}$]→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] (for CED) is recommended.

Clearly, in cases where a 1-stage design is feasible, then this is generally preferable over the 2-stage design, especially for detection of OED and SSE type of effects (**Table 4**, **Figure 9**). The reason for this is that splitting the data into two artificial stages does not exploit the full possible sample size for effect discovery. For example, the power to identify a medium SSE effect with the best 1-stage approach [$\mathrm{Overall}_{1e\text{-}5}$→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] is 80.7% compared to 57.4% for the best 2-stage approach [$\mathrm{Joint}_{1e\text{-}5}$,$\mathrm{Diff}_{0.05}$]→[$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$].

Surprisingly, for CED effects, the combined 2-stage approach [[$\mathrm{Overall}_{1e\text{-}5}$]→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] showed similar or even slightly better power (e.g., +6.6% for medium $CED_{25\%}$ effects) than the analogous 1-stage approach [$\mathrm{Overall}_{1e\text{-}5}$→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] (**Table 4**, **Figure 9E/F**). Obviously, the relaxed α-level for the sex-difference test of approach [[$\mathrm{Overall}_{1e\text{-}5}$]→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] (less loci in follow up) had a stronger positive impact on power of the difference test than using more subjects for step 1 of approach [$\mathrm{Overall}_{1e\text{-}5}$→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$], as this yielded more SNPs in follow up and required a more stringent Bonferroni-corrected α-level for the sex-difference test. Nevertheless, due to the larger power gain for SSE signals (e.g., +54.3% for medium SSE effects) of the 1-stage approach [$\mathrm{Overall}_{1e\text{-}5}$→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] compared to the 2-stage approach [[$\mathrm{Overall}_{1e\text{-}5}$]→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$], the 1-stage design can generally be recommended.

Interestingly, when depicting power over an increasing effect in women and fixing the effect in men to zero (to reflect SSE), the power curves of the 1-stage approaches [$\mathrm{Overall}_{5e\text{-}8}$→$\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}$] and [$\mathrm{Diff}_{5e\text{-}8}$] almost completely overlap (**Figure 9D**). Thus, the traditional general GWAMA approach of screening for genome-wide significant main effects and a follow up of detected loci for sex-difference is equally well powered to find an SSE effect as the genome-wide screen on sex-difference.

Notably, none of the discussed approaches has sufficient power to detect small SSE effects (power < 1%), medium $CED_{25\%}$ effects (power < 27.4%) or large $CED_{50\%}$ effects (power < 28.5%) (**Table 4, Figure 9**). More refined methods or even larger sample sizes are required to find such effects (**Figure 10**). For example, a total overall (sex-combined) sample size of at least 824,000, 460,000, or 466,000 would be required to identify small SSE, medium $CED_{25\%}$ or large $CED_{50\%}$ effects, respectively, with 80% power by the best approach.

Table 4. Power of the best approaches. *Shown is the power to identify sex-difference for various types of sexually dimorphic effects, all of which are based on a fixed effect size in women (the effect size in men is modelled accordingly). The total sample size was set to 200,000 and was assumed to be equally split among sexes and stages.*

Approach	R_F^2 [%]	**Power for varying types of effects [%]**					
		OED	$OED_{50\%}$	$OED_{25\%}$	SSE	$CED_{25\%}$	$CED_{50\%}$
$[Diff_{5e\text{-}8}]$	0.014 (*STAB1*)	43.7	17.6	6.9	0.3	0	0
	0.058 (*PPARG*)	100	100	99.6	47.4	0.3	0
	0.167 (*LYPLAL1*)	100	100	100	100	18.9	0.3
$[Overall_{1e\text{-}5} \rightarrow Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	0	0	0.1	0.7	0.4	0.2
	0.058 (*PPARG*)	0	0.2	4.2	80.7	19.8	2.5
	0.167 (*LYPLAL1*)	0	4.1	56.5	100	84.9	19.2
$[Overall_{5e\text{-}8} \rightarrow Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	0	0	0	0.1	0.2	0.1
	0.058 (*PPARG*)	0	0	0.3	46.7	27.4	4.3
	0.167 (*LYPLAL1*)	0	0.3	19.2	100	90.0	26.7
$[Joint_{1e\text{-}5}, Diff_{0.05}] \rightarrow [Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	12.9	3.9	1.5	0.1	0	0
	0.058 (*PPARG*)	99.8	97	90.3	57.4	5.8	0.6
	0.167 (*LYPLAL1*)	100	100	100	100	52.8	6.2
$[[Overall_{1e\text{-}5}] \rightarrow Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	0	0	0	0.2	0.1	0.1
	0.058 (*PPARG*)	0	0	0.6	26.7	26.4	4.7
	0.167 (*LYPLAL1*)	0	0.6	11.9	98.0	90.9	28.5

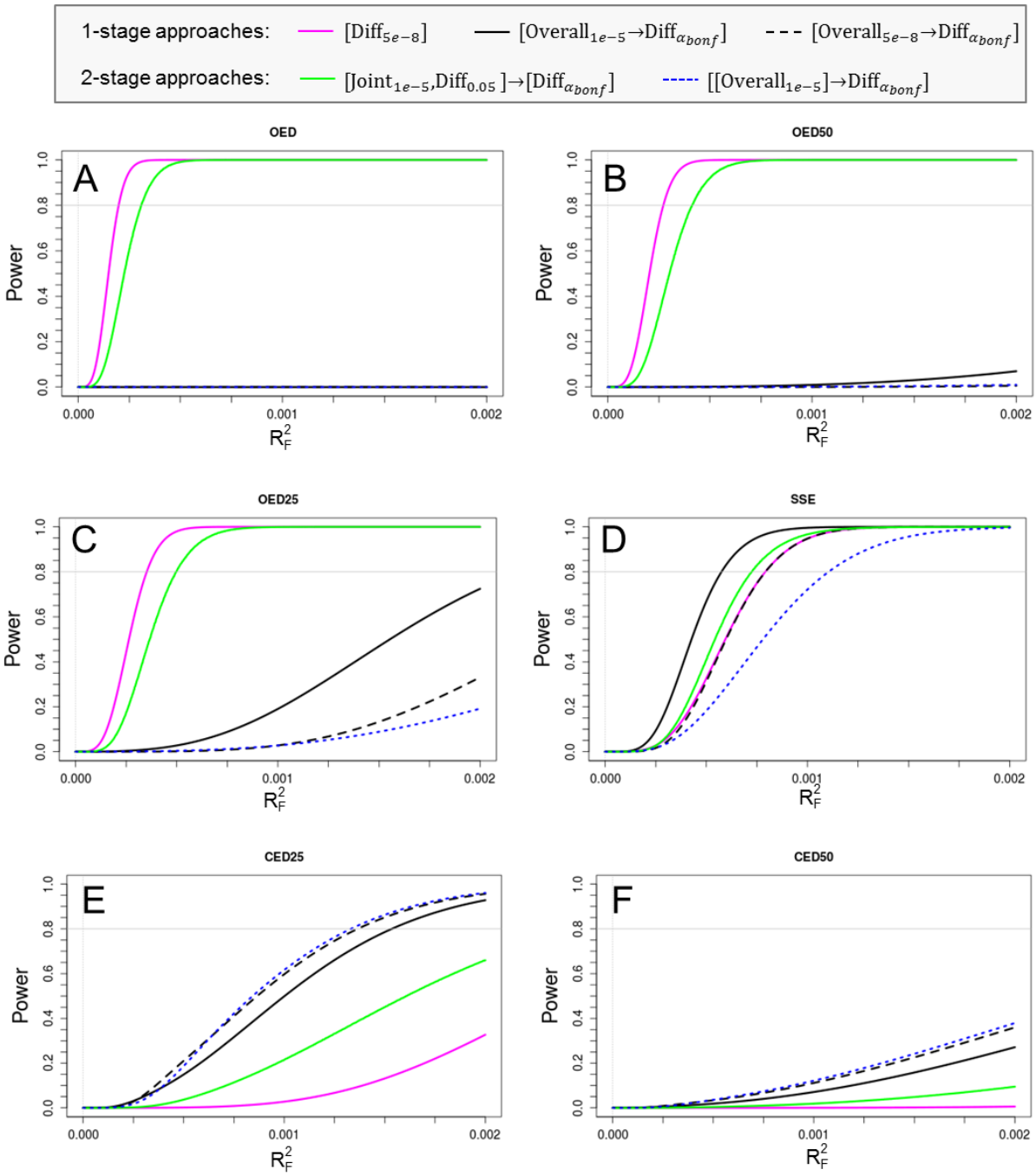


Figure 9. Power of the best approaches over explained variance in women. *Shown is the power to detect sex-difference over the effect size (explained variance R_F^2) in women and for various types of sexually dimorphic effects: A: OED, B: $OED_{50\%}$, C: $OED_{25\%}$, D: SSE, E: $CED_{25\%}$, and F: $CED_{50\%}$ (the effect size in men was modelled accordingly).*

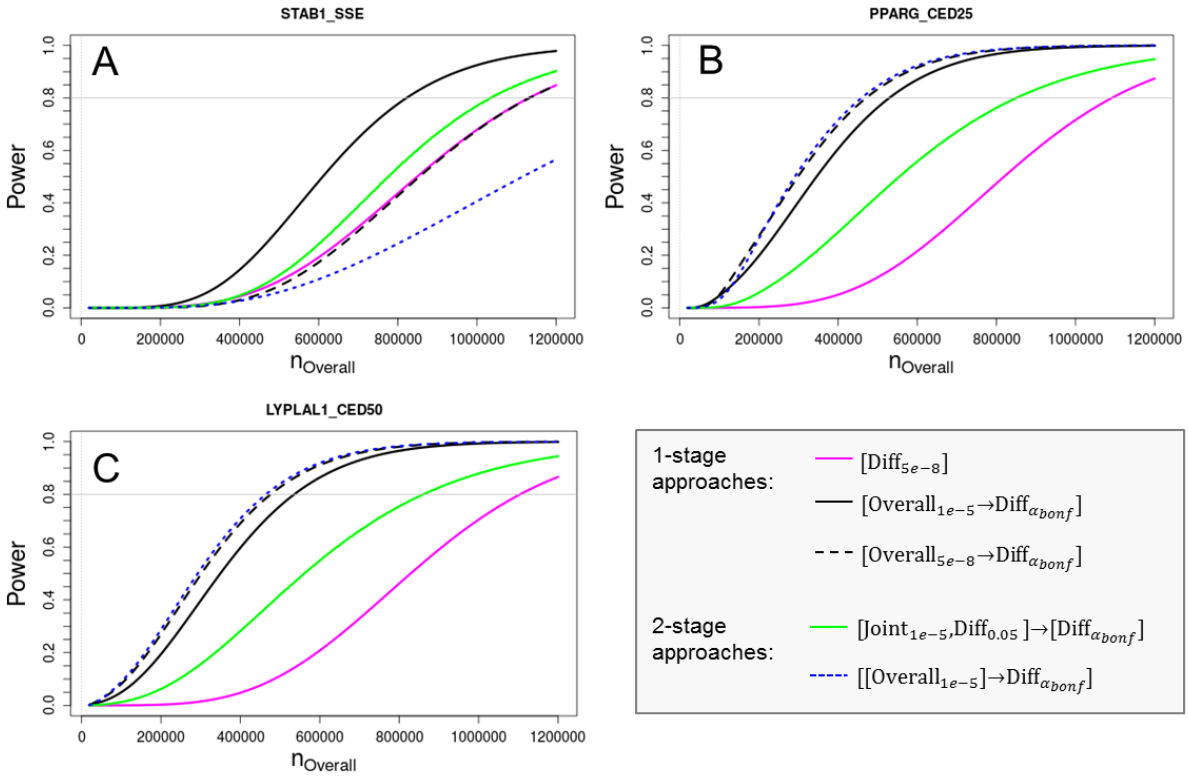


Figure 10. Power of the best approaches over total (sex-combined) sample size. *Shown is the power to detect sex-difference for A: a small SSE effect that is based on a small modelled genetic effect in women (comparable to the WHR_{adjBMI} effect near STAB1), B: a medium $CED_{25\%}$ effect that is based on a medium modelled genetic effect in women (comparable to the WHR_{adjBMI} effect near PPARG), and C: a large $CED_{50\%}$ effect that is based on a large modelled genetic effect in women (comparable to the WHR_{adjBMI} effect near LYPLAL1).*

2.2.2.2 Analytical power comparison for the unbalanced stratified GWAMA approaches

Next, the impact of unbalanced strata designs on power was investigated. Again, a total sample size of 200,000 was assumed, but the proportion of stratum-specific sample sizes *f* was varied from 0.2 to 5, which corresponds to a stratum 1 sample size that is 5-times larger (*f* = 0.2) to 5-times smaller (*f* = 5) than stratum 2.

Power of the considered 1- and 2-stage approaches was depicted over varying *f*, and for various types of stratum-difference (OED, $OED_{50\%}$, $OED_{25\%}$, SSE, $CED_{25\%}$ and $CED_{50\%}$ signals) that were modelled based on a given effect in stratum 1 (set to a medium genetic effect on WHR_{adjBMI}, comparable to the known effect in the *PPARG* locus). Results are summarized in **Figure 11** (for 1-stage approaches) and in **Figure 12** (for 2-stage approaches).

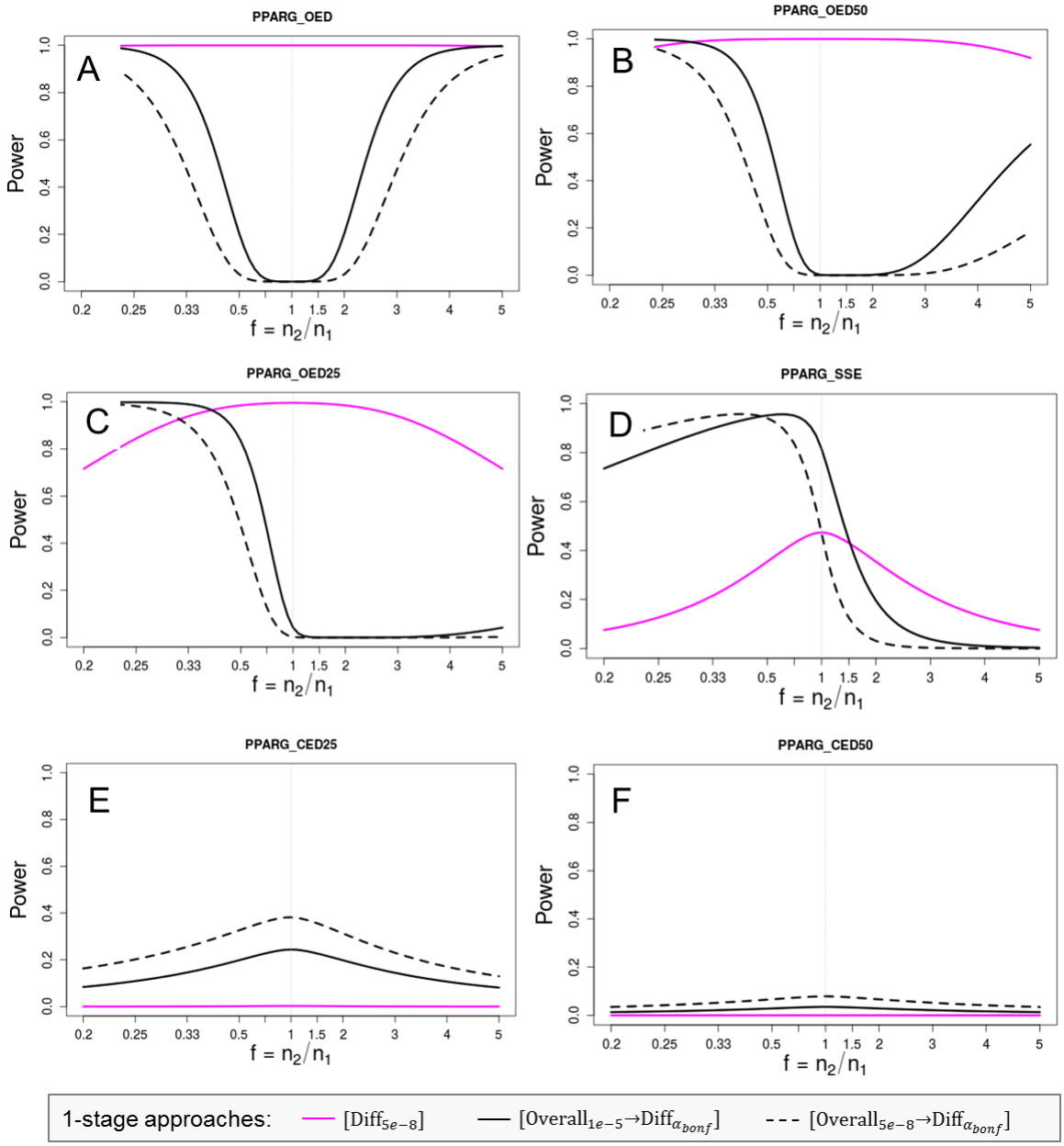


Figure 11. Power of the 1-stage approaches. *Shown is the power to identify stratum-difference depicted over the proportion of stratum-specific sample sizes f = n_2/n_1. The total sample size is 200,000 and the stratum 1 effect was fixed at a medium WHR_{adjBMI} effect (comparable to the known PPARG effect). Power is shown for varying scenarios of stratum-difference: A: OED, B: $OED_{50\%}$, C: $OED_{25\%}$, D: SSE, E: $CED_{25\%}$ and F: $CED_{50\%}$ (the effect size in men was modelled accordingly). .*

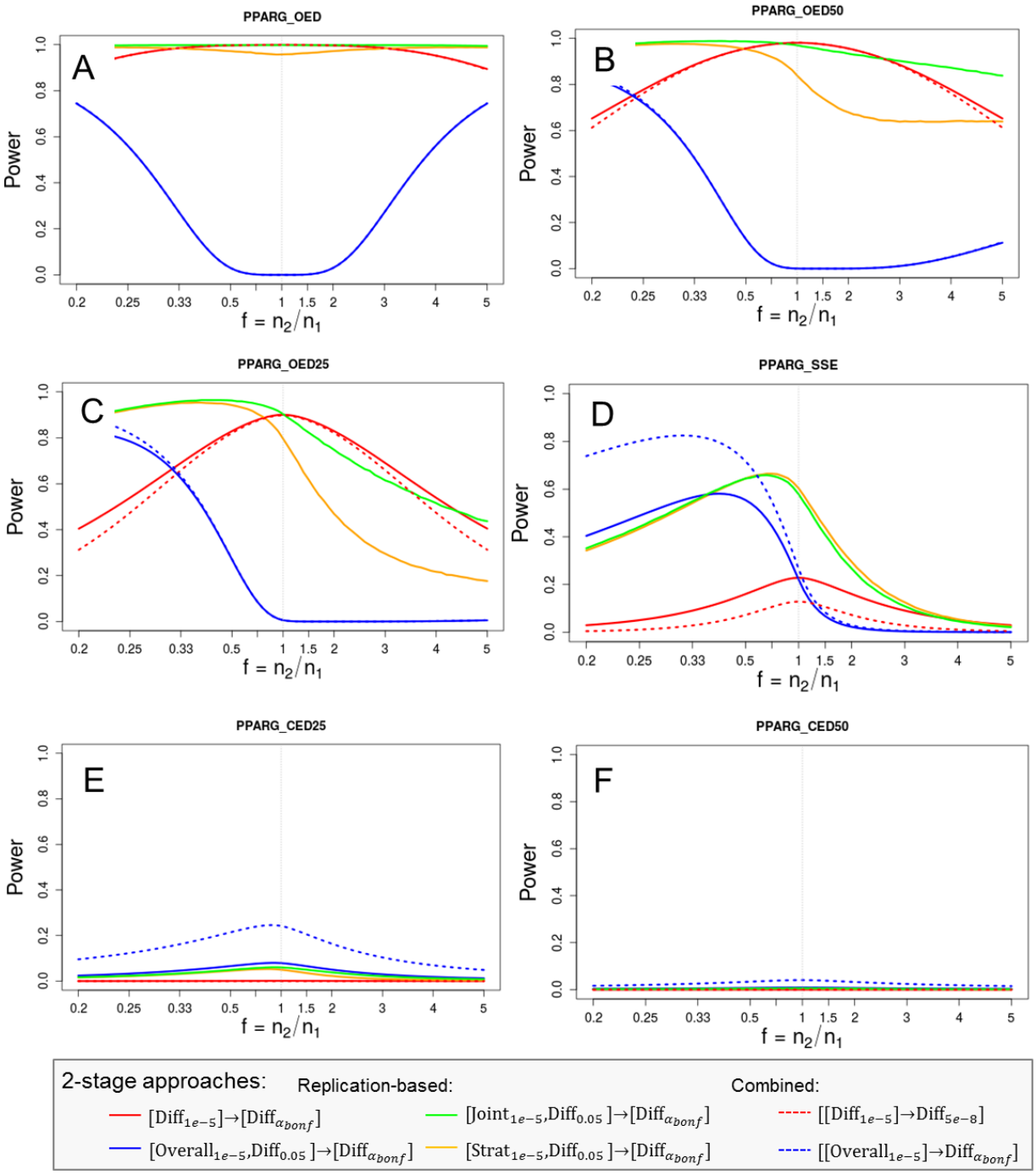


Figure 12. Power of the 2-stage approaches. *Shown is the power to detect stratum-difference depicted over the proportion of stratum-specific sample sizes f = n_2/n_1. The total sample size is 200,000 (equally split among stages). The stratum 1 effect was fixed at a medium WHR_{adjBMI} effect (comparable to the known PPARG effect). Power is shown for varying scenarios of stratum-difference: A: OED, B: $OED_{50\%}$, C: $OED_{25\%}$, D: SSE, E: $CED_{25\%}$ and F: CED_{50} (the effect in men was modelled accordingly).*

2. Screening for difference between two strata

For a general stratified GWAMA that does not aim at the identification of specific types of interaction effects, varying *f* did not change the combination of optimal approaches from the balanced strata designs: Still, the combination of approaches $[\mathrm{Diff}_{5e\text{-}8}]$ and $[\mathrm{Overall}_{1e\text{-}5}{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ (or $[\mathrm{Overall}_{5e\text{-}8}{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$) is recommended for 1-stage designs (**Figure 11**), and the combination of approaches $[\mathrm{Joint}_{\alpha_1}, \mathrm{Diff}_{\alpha_2}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[[\mathrm{Overall}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ is recommended for 2-stage designs (**Figure 12**).

However, the optimal approach to particularly identify SSE effects changes with varying *f*: For the 1-stage design and for *f* > *1*, the power of approach $[\mathrm{Diff}_{5e\text{-}8}]$ surpasses the power of the overall filtering approaches (which were previously recommended for SSE under the balanced strata design, **Figure 11D**). For the 2-stage design and for *f* < *1*, the power of the combined 2-stage approach $[[\mathrm{Overall}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ surpasses the power of the replication-based 2-stage approach $[\mathrm{Joint}_{1e\text{-}5}, \mathrm{Diff}_{0.05}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ (which was previously recommended for SSE under the balanced strata design, **Figure 12D**).

Interestingly, for all types of stratum-difference, the power of approaches $[\mathrm{Diff}_{5e\text{-}8}]$, $[\mathrm{Diff}_{1e\text{-}5}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[[\mathrm{Diff}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ (solely involving the difference test) is symmetric and at maximum at *f* = 1 (**Figure 11**, **Figure 12**). Thus, the difference test itself performs best if the two strata are balanced in size. For example, given the GIANT SMOKING scenario, the power of approach $[\mathrm{Diff}_{5e\text{-}8}]$ is identical for *f* = 0.28 (i.e., modelled effect in the larger non-smoker group) and *f* = 3.48 (i.e., modelled effect in the smaller smoker group, **Table 5**).

For SSE type of effects and for approaches that involve initial filtering steps, the maximum power is reached at *f* < 1, which corresponds to a design for which the stratum carrying the effect (stratum 1) is larger than the stratum lacking effect (stratum 2, **Figure 11**, **Figure 12**). For example, given the GIANT SMOKING scenario, it is more likely to identify SSE signals comprising the effect in non-smokers (the larger stratum) as compared to SSE signals comprising the effect in smokers (the smaller stratum) (**Table 5**).

***Table 5. Power of the best approaches for the unbalanced GIANT SMOKING scenario**. Power to identify stratum-difference is shown for varying types of effects, each assuming a modelled and fixed effect size in stratum 1 that is set to a small (STAB1), medium (PPARG) or large (LYPLAL1) known effect on WHR_{adjBMI}. For f=0.28, stratum 1 corresponds to the larger non-smoker stratum. For f=3.48, stratum 1 corresponds to the smaller smoker stratum. The effect size in stratum 1 is assumed to be given and the effect size in stratum 2 is modelled accordingly (e.g., set to 0 for the SSE effect).*

		Type of effect					
		f = 0.28			***f = 3.48***		
Approach	R_1^2 [%]	$OED_{25\%}$	SSE	$CED_{25\%}$	$OED_{25\%}$	SSE	$CED_{25\%}$
$[Diff_{5e\text{-}8}]$	0.014 (*STAB1*)	1.6	0.1	0	1.6	0.1	0
	0.058 (*PPARG*)	89.9	16.7	0.1	89.9	16.7	0.1
	0.167 (*LYPLAL1*)	100	98.5	5.0	100	98.5	5.0
$[Overall_{1e\text{-}5} \rightarrow Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	8.7	4.5	0.7	0	0	0.1
	0.058 (*PPARG*)	99.7	86.4	12.6	0.4	1.9	12.5
	0.167 (*LYPLAL1*)	100	100	66.3	8.1	36.9	66.3
$[Overall_{5e\text{-}8} \rightarrow Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	1.7	1.9	0.7	0	0	0
	0.058 (*PPARG*)	95.7	93.1	22.8	0	0.1	19.9
	0.167 (*LYPLAL1*)	100	100	79.3	0.7	8.6	79.3
$[Joint_{1e\text{-}5} Diff_{0.05}] \rightarrow [Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	1.7	0.2	0	0.2	0	0
	0.058 (*PPARG*)	94.6	49.6	2.8	56.4	7.1	1.7
	0.167 (*LYPLAL1*)	100	99.1	29.4	100	90.9	29.4
$[[Overall_{1e\text{-}5}] \rightarrow Diff_{\alpha_{Bonf}}]$	0.014 (*STAB1*)	1.3	0.9	0.2	0	0	0
	0.058 (*PPARG*)	74.3	82.1	14.1	0.1	0.3	8.5
	0.167 (*LYPLAL1*)	100	100	68.9	1.1	6.3	69.0

3 Stratified GWAMA approaches to screen for G x AGE x SEX interaction effects

Besides the differences in genetic effects between two strata, other – more complicated – scenarios may exist that involve multiple stratification variables. For example, it is well known that women's body shape changes around menopause (due to the decline in estrogen levels) from a gynoid (more fat stored around hips) to an android body shape (more fat stored around waist). Thus, after menopause, the differences in body shape between men and women are less pronounced than before menopause. To investigate whether the age-dependent change in sex-differences of body shape is influenced by genetic effects, one would particularly be interested in genetic effects that are modified by both, AGE and SEX (AGE as dichotomized age: younger vs older than 50 years, which reflects mean age of menopause).

The overarching aim of the following chapter is to provide a systematic methodological evaluation of screening approaches that aim identification of 3-way G x AGE x SEX interaction effects and are based on an age- and sex-stratified GWAMA. The proposed approaches are exemplified on this specific configuration, but can readily be applied to any other 2 x 2 strata configuration involving two dichotomous (environmental) stratification variables S_1 and S_2 and aiming at identification of 3-way G x S_1 x S_2 interaction effects.

Again, an *approach* is defined here as a combination or concatenation of multiple statistical tests (i.e., *steps*) that are applicable to age- and sex-stratified GWAMA outcomes. Relevant statistical tests are introduced and a systematic scheme of approaches is presented.

The performance of the approaches was compared by simulation-based estimation of type 1 error and by analytical computations of power. Varying realistic scenarios were considered and an attempt to recommend approaches - based on type of 3-way G x AGE x SEX interaction effect - was made.

Herewith, a range of possible combinations of stratum-specific effects are distinguished (**Figure 13**). These include biologically plausible and relevant 3-way interaction effects, such as the *1-stratum interaction* (e.g., effect is present in F≤50, but lacking in all other strata) or the *3-strata interaction* (e.g., effect is only lacking in M>50), as well as less plausible *extreme 3-way interaction* effects that involve opposite directions across AGE and SEX. Other combinations include previously discussed 2-way interaction effects, such as age-difference (i.e., G x AGE, independent of SEX) or sex-difference (i.e., G x SEX, independent of AGE).

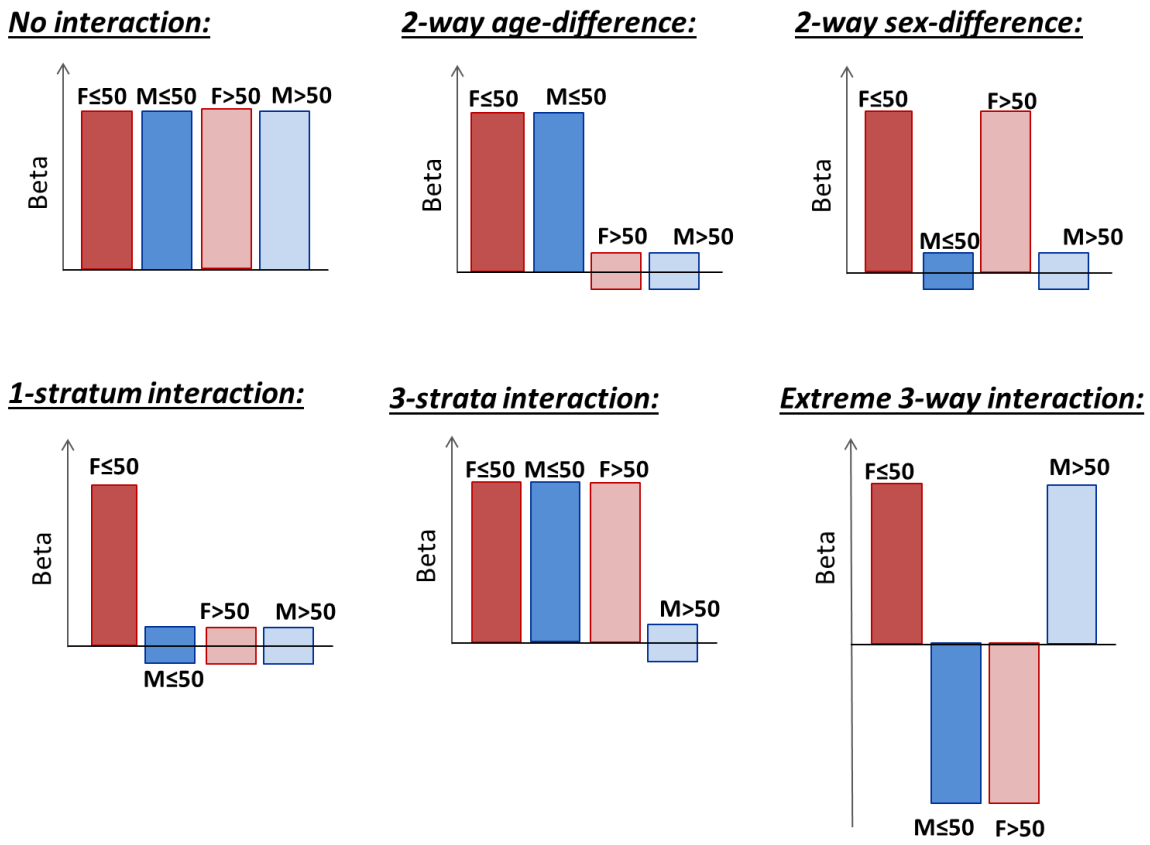


Figure 13. *Examples for possible combinations of age- and sex-specific effects.*

3.1 Materials and Methods

The following chapters describe the prerequisites of the methodological evaluation of age- and sex-stratified GWAMA approaches to identify 3-way G x AGE x SEX interaction effects.

After presenting general assumptions (chapter 3.1.1), a statistical test for difference-of-difference is introduced (chapter 3.1.2) that can be used to test for 3-way G x AGE x SEX interaction given the age- and sex-stratified GWAMA setting. Furthermore, filtering tests are introduced (chapter 3.1.3) and used to construct several screening approaches (chapter 3.1.4). A systematic scheme of approaches was developed that is utilized to compare performance (type 1 error and power) of approaches.

Methodological prerequisites of the simulation-based evaluation of type 1 error (chapter 3.1.5), the derivation of analytical power formulae as well as methodological details about the analytical power computations (chapter 3.1.6), are presented.

3.1.1 Assumptions and definitions

In the following, considerations are based on an age- and sex-stratified GWAMA model that is given by four stratum-specific linear regression models:

3. Screening for 3-way G x AGE x SEX interaction

$$\begin{aligned} Y_{M\leq 50} &= \propto_{M\leq 50} + \beta_{M\leq 50} G_{M\leq 50} + \varepsilon_{M\leq 50}, \quad \varepsilon_{M\leq 50} \sim N\left(0, \sigma^2_{M\leq 50}\right) \\ Y_{F\leq 50} &= \propto_{F\leq 50} + \beta_{F\leq 50} G_{F\leq 50} + \varepsilon_{F\leq 50}, \quad \varepsilon_{F\leq 50} \sim N\left(0, \sigma^2_{F\leq 50}\right) \\ Y_{M>50} &= \propto_{M>50} + \beta_{M>50} G_{M>50} + \varepsilon_{M>50}, \quad \varepsilon_{M>50} \sim N\left(0, \sigma^2_{M>50}\right) \\ Y_{F>50} &= \propto_{F>50} + \beta_{F>50} G_{F>50} + \varepsilon_{F>50}, \quad \varepsilon_{F>50} \sim N\left(0, \sigma^2_{F>50}\right) \end{aligned} \qquad \textit{(23).}$$

The stratification is done by two dichotomous variables (AGE and SEX, AGE dichotomized at 50 years of age, which reflects menopause in women) that separate each study sample into four subgroups: Younger men (M≤50), younger women (F≤50), older men (M>50), and older women (F>50). For each SNP, each study fits the stratum-specific regression models and obtains the stratum-specific effect estimates with standard errors. The pooled stratum-specific genetic effect estimates $b_{M\leq 50}$, $b_{F\leq 50,}$ $b_{M>50}$ and $b_{F>50}$, are obtained from stratum-specific inverse-variance weighted meta-analyses. It is assumed that the stratified GWAMAs have already been conducted so that the pooled stratum-specific effect estimates with standard errors are already available.

As for the 2-strata configuration, stratum-specific phenotypes are assumed to be continuous, and to follow equal normal distributions, $Y_i \sim N(\mu_Y, \sigma^2_Y)$ where *i* reflects M≤50, F≤50, M>50, or F>50; stratum-specific additively modeled genotypes G_i are assumed to follow equal genotype distributions; and similar populations are assumed to be involved in the meta-analyses (which allows ignoring the meta-analysis concept for the power and the type 1 error evaluations).

Finally, to avoid overly complexity, strata are assumed to be equally balanced comprising identical stratum-specific sample sizes, defined as $n = n_{M\leq 50} = n_{F\leq 50} = n_{M>50} = n_{F>50}$ (note that the total strata-combined sample size is denoted as $n_{Overall}$).

3.1.2 Testing for G x AGE x SEX interaction given the age- and sex-stratified GWAMA model

To investigate 3-way G x AGE x SEX interaction effects given the age- and sex-stratified GWAMA model, a ***difference-of-difference test*** is introduced that compares whether sex-difference differs by age (H_0: $(b_{M\leq 50} - b_{F\leq 50}) = (b_{M>50} - b_{F>50})$):

$$Z_{DiffDiff} = \frac{(b_{M\leq 50} - b_{F\leq 50}) - (b_{M>50} - b_{F>50})}{\sqrt{se(b_{M\leq 50})^2 + se(b_{F\leq 50})^2 + se(b_{M>50})^2 + se(b_{F>50})^2}} \sim N(0,1)\,|H_0 \qquad \textit{(24).}$$

Assuming the null hypothesis being true, the z statistic follows a standard normal distribution. The z test yields the difference-of-difference P-Value $P_{DiffDiff}$.

By re-arranging the formula or the null hypothesis, it can be shown that testing whether 'sex-difference differs by age' (H_0: $(b_{M\leq 50} - b_{F\leq 50}) = (b_{M>50} - b_{F>50})$) is

mathematically equivalent to testing whether 'age-difference differs by sex' ($H_0: (b_{M\leq 50} - b_{M>50}) = (b_{F\leq 50} - b_{F>50})$).

Under the assumption of independent samples, unrelated subjects and no latent covariate interacting with the AGE or SEX, the difference-of-difference test is mathematically equivalent to testing the pooled G x AGE x SEX interaction estimate $b_{GxAGExSEX}$ obtained from an *interaction GWAMA model* that is given by

$$Y = \alpha + \beta_G G + \beta_{AGE} AGE + \beta_{SEX} SEX + \beta_{GxAGE} GxAGE + \beta_{GxSEX} GxSEX + \beta_{AGExSEX} AGExSEX + \beta_{GxAGExSEX} GxAGExSEX + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \qquad (25).$$

3.1.3 Statistical tests to filter stratified GWAMA data sets prior to G x AGE x SEX interaction testing

In order to construct age- and sex-stratified GWAMA approaches that involve initial filtering steps, two types of filtering tests are distinguished: (i) Directly applicable filtering tests (chapter 3.1.3.1), and (ii) marginal filtering tests (chapter 3.1.3.2) that are indirectly applicable because they require initial meta-analysis of stratum-specific results.

3.1.3.1 Directly applicable filtering tests

Similarly to the stratified GWAMA setting involving two strata, a stratified, an overall and a joint association test can directly be applied (for filtering) to the four stratum-specific effect estimates with standard errors:

- A *stratified test* can be performed that infers whether any of the four stratum-specific effects is associated with the phenotype ($H_0: b_{M\leq 50} = 0 \wedge b_{F\leq 50} = 0 \wedge b_{M>50} = 0 \wedge b_{F>50} = 0$): The stratified test employs four stratum-specific t tests that yield stratum-specific association P-Values $P_{M\leq 50}$, $P_{F\leq 50}$, $P_{M>50,}$ *and* $P_{F>50}$. Finally, the stratified association P-Value is defined as $P_{Strat} = 4*min(P_{M\leq 50}, P_{F\leq 50,} P_{M>50,} P_{F>50})$, which is corrected for the multiple testing of four strata.
- An *overall test* can be performed to infer whether the overall (strata-combined) effect is associated with the phenotype ($H_0: b_{Overall} = 0$)*:*

  $$T_{Overall} = \frac{b_{Overall}}{se(b_{Overall})} \sim t(n_{Overall} - 2) \mid H_0 \qquad (26),$$

  where $n_{Overall}$ is the strata-combined sample size ($n_{Overall} = 4n$) and where $b_{Overall}$ and $se(b_{Overall})$ are the overall genetic effect with standard error that are obtained from inverse-variance weighted meta-analysis of the four strata:

$$b_{Overall} = \frac{\sum_i b_i/se(b_i)^2}{\sum_i 1/se(b_i)^2}, \qquad se(b_{Overall}) = \sqrt{\frac{1}{\sum_i 1/se(b_i)^2}} \qquad (27).$$

Here, *i* iteratively refers to M≤50, M>50, F≤50 and F>50. The t test yields the overall association P-Value $P_{Overall}$.

- A *joint test* can be performed to infer whether the joint (main + 2-way interaction + 3-way interaction) effect is associated with the phenotype ($H_0: b_G = 0 \wedge b_{GxAGE} = 0 \wedge b_{GxSEX} = 0 \wedge b_{GxAGExSEX} = 0$):

$$C_{Joint} = \sum_i \left(\frac{b_i}{se(b_i)}\right)^2 \sim \chi^2(4) \,|H_0 \qquad (28).$$

Here, *i* iteratively refers to M≤50, M>50, F≤50 and F>50. The chi-square test yields the joint-test P-Value P_{Joint}.

3.1.3.2 Filtering on marginal tests that require meta-analyses of age- and sex-specific results

In addition to the directly applicable filtering tests, the 2 x 2 age- and sex-stratified GWAMA scenarios offers the possibility to filter on marginal tests (**Figure 14**).

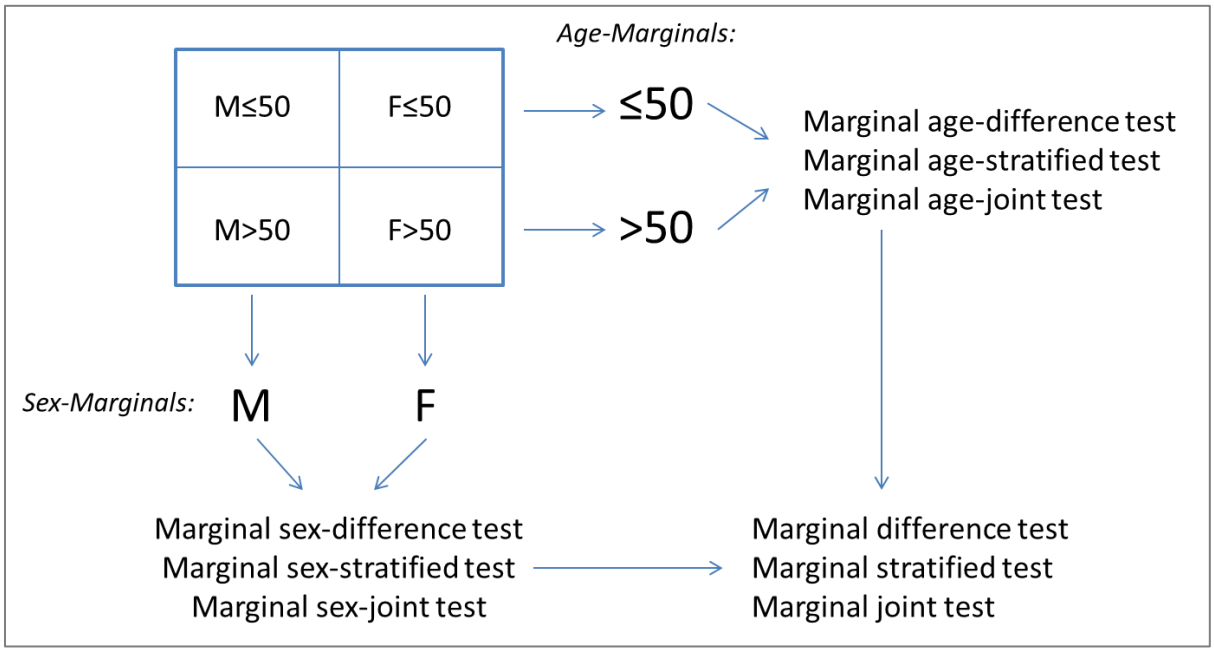


Figure 14. *Marginal tests based on age- and sex-stratified GWAMA outcomes.*

The marginal tests are based on using age-marginals (i.e., age-specific effects $b_{\leq 50}$ and $b_{>50}$, with respective standard errors) and sex-marginals (i.e., sex-specific effects b_M and b_F, with respective standard errors) that have to be obtained from pooling the respective age-

and sex-specific subgroups. For example, to obtain the genetic age-marginal effect for the younger individuals ($b_{\leq 50}$), a meta-analysis of genetic effects in younger men ($b_{M\leq 50}$) and younger women ($b_{F\leq 50}$) has to be performed.

Each marginal test is a combination of a marginal age-version and a marginal sex-version of the respective test:

- A *marginal difference test* can be performed that infers whether the genetic effect differs by age (marginal age-difference test, independent of sex) or whether the genetic effect differs by sex (marginal sex-difference test, independent of age): $H_0: b_{\leq 50} = b_{>50} \wedge b_M = b_F$. The marginal difference P-Value is defined as $P_{MarDiff} = 2*min(P_{Agediff}, P_{Sexdiff})$, which is corrected for the multiple testing of two difference tests (see chapter 2.1.2 for the definition of the difference test).
- A *marginal stratified test* can be performed that infers whether any of the age-marginal effects are associated with the phenotype (marginal age-stratified test, independent of sex) or whether any of the sex-marginal effects are associated with the phenotype (marginal sex-stratified test, independent of age): $H_0: b_{\leq 50} = 0 \wedge b_{\leq 50} = 0 \wedge b_M = 0 \wedge b_F = 0$. The marginal stratified P-Value is defined as $P_{marStrat} = 2*min(P_{Agestrat}, P_{Sexstrat})$, which is corrected for the multiple testing of two 2-strata stratified tests (see chapter 2.1.3 for the definition of the 2-strata stratified test). The marginal stratified P-Value can also be written as $P_{MarStrat} = 4*min(P_{\leq 50}, P_{>50}, P_M, P_F)$.
- A *marginal joint test* can be performed that infers whether the age-marginal effects are jointly associated with the phenotype (marginal age-joint test, independent of sex) or whether the sex-marginal effects are jointly associated with the phenotype (marginal sex-joint test, independent of age): $H_0: b_G = 0 \wedge b_{GxAGE} = 0 \wedge b_{GxSEX} = 0$. The marginal joint test P-Value is defined as $P_{MarJoint} = 2*min(P_{Agejoint}, P_{Sexjoint})$, which is corrected for the multiple testing of two 2-strata joint tests (see chapter 2.1.3 for the definition of the 2-strata joint test).

3.1.4 A systematic scheme of age- and sex-stratified GWAMA approaches to identify G x AGE x SEX interaction

The introduced statistical tests are used to construct various age- and sex-stratified GWAMA approaches, each of which aims at screening for variants with significant 3-way G x AGE x SEX interaction effects. Generally, an approach is designed with multiple steps (statistical tests). The last step of each approach is to test for difference-of-difference (as a means to test for 3-way G x AGE x SEX interaction effects). Other statistical tests are solely employed for filtering the genome-wide data sets prior to the difference-of-difference testing. Multiple steps are implemented within a single data set (*1-stage design*) and – in order to avoid overly complexity – not extended to other stage designs. Based on the previously introduced

notation (see chapter 2.1.4), a systematic scheme of considered approaches was developed and summarized in **Figure 15.**

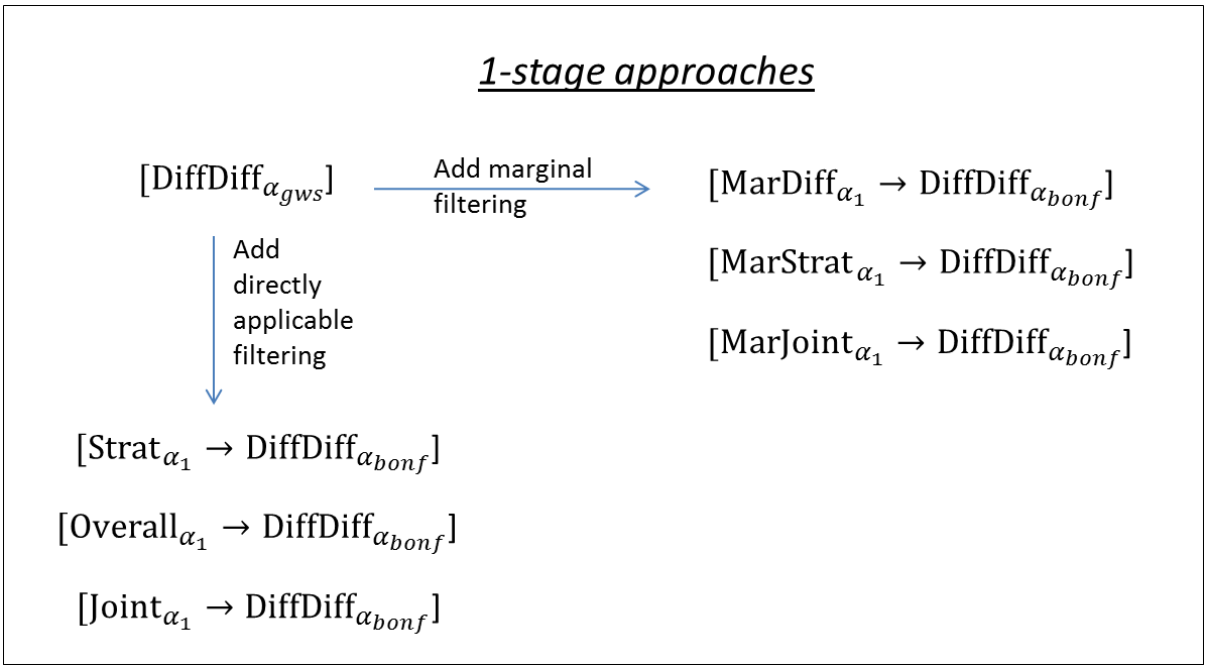


Figure 15. *Systematic scheme of 1-stage age- and sex-stratified GWAMA approaches to identify 3-way G x AGE x SEX interaction effects (α_{gws}= 5 x 10^{-8}; α_1 = α-level for the filtering step; α_{Bonf} = Bonferroni-corrected α-level).*

Generally, all of the considered approaches aim at using Bonferroni-corrected α-levels for the final difference-of-difference test: α_{Bonf} = *0.05/M*, where *M* is the number of independent difference-of-difference tests performed.

The most intuitive approach is to screen for difference-of-difference at a genome-wide significance level: $\alpha_{Bonf} = \alpha_{gws}$ = 5 x 10^{-8} (= 0.05/10^6, Bonferroni-corrected for an approximate number of one million independent tests) (Johnson et al., 2010).

Further approaches are considered that involve filtering on directly applicable or on marginal tests. Their implementation into the 1-stage design has to be validated with regards to type 1 error and the impact on power of the difference-of-difference test has to be investigated.

Practically (same as for the 2-strata configuration), common stratified GWAMA projects aim at identification of any type of 3-way interaction and it is likely that the optimal approach varies by type of interaction effect – necessitating multiple screening approaches. To avoid overly conservative correction of multiple (already conservatively controlled)

screening approaches, a 'final' multiple testing correction (for the number of outlined approaches) is not applied and ignored in the following.

Technically (same as for the 2-strata configuration), genome-wide data sets are clumped into independent regions using a distance-based criterion (*distance* < 500KB, pairwise between SNPs of a specific region) and the region-specific lead-SNP is defined as the SNP with the lowest P-Value across all SNPs of a specific region. Importantly, the independent lead-SNPs are those that are put forward and are tested for difference-of-difference in the final step. The total number of selected lead-SNPs is denoted as *M* in the following (as introduced before). To correct for the multiple difference-of-difference testing of *M* independent lead SNPs, the α-level of the difference-of-difference test undergoes a Bonferroni-correction: α_{Bonf} = *0.05/M* (except for approach $[\mathrm{DiffDiff}_{\alpha_{gws}}]$ that employs a genome-wide significant α-levels for the final difference-of-difference test).

3.1.5 Simulation based evaluation of type 1 error

A simulation-based evaluation of type 1 error rates was performed for all considered approaches. Methodological details of the simulations are described in the following.

Simulated data sets were created that follow the null hypothesis of 'No age-difference in sex-difference of genetic effects', H_0: $(b_{M\leq 50} - b_{F\leq 50}) = (b_{M>50} - b_{F>50})$. More specifically, three versions of the null hypothesis were created: One assumes lack of stratum-specific effects ($H_0^{noeffect}$: $b_{M\leq 50} = b_{F\leq 50} = b_{M>50} = b_{F>50} = 0$), one assumes identical (unequal zero) stratum-specific effects (H_0^{effect}: $b_{M\leq 50} = b_{F\leq 50} = b_{M>50} = b_{F>50} \neq 0$), and one assumes true sex-difference implying an opposite effect between men and women ($H_0^{sexdiff}$: $b_{M\leq 50} = b_{M>50} = -b_{F\leq 50} = -b_{F>50} \neq 0$).

First, real genotypes were obtained from 750 younger men, 750 younger women, 750 older men and from 750 older women, from the KORA study (Wichmann et al., 2005) for three well-imputed SNPs that cover the full allele frequency spectrum: rs6002481 (MAF = 0.02), rs8138968 (MAF = 0.3) and rs6007738 (MAF = 0.5) (same SNPs as used before for the 2-strata configuration).

Second, simulated phenotypes (for each subgroup of 750 subjects) were created according to *Y~N(0,1)* for $H_0^{noeffect}$, and according to *Y|G=0~N(0,1)*, *Y|G=1~N($b_{80\%}$,1)* and *Y|G=2~N(2*$b_{80\%}$,1)* for H_0^{effect}. For $H_0^{sexdiff}$, phenotypes for younger and older men were simulated according to Y|X=0~*N(0,1)*, Y|X=1~*N($b_{80\%}$,1)* and Y|X=2~*N(2*$b_{80\%}$,1),* and phenotypes for younger and older women were simulated according to Y|X=0~*N(0,1)*, Y|X=1~*N(-$b_{80\%}$,1)* and Y|X=2~*N(-2*$b_{80\%}$,1).* Always, $b_{80\%}$ reflects the minimum effect size detectable with 80% power by 750 samples (at α = 0.05), and *G=0*, *G=1* or *G=2* denote the group of individuals carrying 0, 1, or 2 minor alleles, respectively. Using G*Power, $b_{80\%}$ was

estimated to be 0.514, 0.157 and 0.144 for rs6002481, rs8138968 and rs6007738, respectively (Faul et al., 2009).

Third, for each null hypothesis and for each SNP, the simulated phenotypes were stratum-specifically tested for association with the real SNP genotypes. Age- and sex-stratified genetic effect estimates, with standard errors and association P-Values, were derived.

Next, for each null hypothesis and for each SNP, the simulation and stratum-specific association testing was repeated 1,000,000 times yielding 1,000,000 age- and sex-specific genetic effect estimates, with standard errors and association P-Values.

Finally, for each null hypothesis and for each SNP, the age- and sex-stratified GWAMA approaches were applied to the respective 1,000,000 stratum-specific genetic effect estimates. The α-level for the initial filtering was set to $\alpha_1 = 0.05$ (which corresponds to the estimation of $b_{80\%}$). For each approach, type 1 error rate (T1ER) of the final difference-of-difference test was calculated by the proportion of nominal significant test results among all difference-of-difference tests: $T1ER = \#(P_{DiffDiff}<0.05)/M_{filt}$ (M_{filt} being the number of data points passing the filtering, which corresponds to the previously introduced *M)*.

3.1.6 Analytical computation of power

Power comparisons aim finding the best age- and sex-stratified GWAMA approach to identify difference-of-difference. Power formulae for the single tests were derived and then combined to obtain the power formulae for the approaches.

In order to model various types of difference-of-difference, each formula was derived in dependence of the stratum-specific explained variances $R^2_{M\leq 50}$, $R^2_{F\leq 50}$, $R^2_{M>50}$ and $R^2_{F>50}$.

3.1.6.1 Power formulae for the directly applicable statistical tests

In the following, power formulae for the directly applicable statistical tests are presented.

Difference-of-difference test:

The difference test involves the z statistic $Z_{DiffDiff}$ (given by equation (24)). The power of the difference-of-difference test is given by

$$\begin{aligned} Pwr_{DiffDiff}(\alpha) &= P\left(Z_{DiffDiff} \leq z_{\frac{\alpha}{2}} | H_A\right) + P\left(Z_{DiffDiff} \geq z_{1-\frac{\alpha}{2}} | H_A\right) \\ &= \Phi\left(-z_{1-\frac{\alpha}{2}} - \frac{(b_{M\leq 50} - b_{F\leq 50}) - (b_{M>50} - b_{F>50})}{\sqrt{se(b_{M\leq 50})^2 + se(b_{F\leq 50})^2 + se(b_{M>50})^2 + se(b_{F>50})^2}}\right) \\ &+ \Phi\left(-z_{1-\frac{\alpha}{2}} + \frac{(b_{M\leq 50} - b_{F\leq 50}) - (b_{M>50} - b_{F>50})}{\sqrt{se(b_{M\leq 50})^2 + se(b_{F\leq 50})^2 + se(b_{M>50})^2 + se(b_{F>50})^2}}\right) \end{aligned} \quad (29),$$

where, Φ denotes the cumulative standard normal distribution and z_q denotes the q-th quantile of Φ. Utilizing equations (10) and (11) yields (note that *n* is the sample size of a single stratum):

$$Pwr_{DiffDiff}\left(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n\right) = \Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{n}\frac{|R_{M\leq 50}| - |R_{F\leq 50}| - |R_{M>50}| + |R_{F>50}|}{\sqrt{4 - R^2_{M\leq 50} - R^2_{M>50} - R^2_{F\leq 50} - R^2_{F>50}}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} + \sqrt{n}\frac{|R_{M\leq 50}| - |R_{F\leq 50}| - |R_{M>50}| + |R_{F>50}|}{\sqrt{4 - R^2_{M\leq 50} - R^2_{M>50} - R^2_{F\leq 50} - R^2_{F>50}}}\right) \quad (30).$$

Stratified test:

The stratified test involves four stratum-specific t tests that are performed simultaneously. The power of a single stratum-specific t test in dependence of the stratum-specific explained variance and sample size is given by equation (16). To obtain the power of the stratified test (combination of four stratum-specific t tests), the power formulae of the stratum-specific t tests have to be combined and the stratum-specific α-levels have to be corrected for the multiple testing of four stratum-specific t tests. Leveraging independence of stratum-specific t tests (due to using independent subjects), the power of the stratified test is given as follows:

$$Pwr_{Strat}\left(\alpha, (R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right) = 1 - \prod_i (1 - Pwr_t\left(\frac{\alpha}{4}, R^2_i, n\right)) \quad (31),$$

where *i* is iteratively expressed as M≤50, F≤50, M>50 and F>50 and *n* is the sample size of a single stratum.

Overall test:

The overall test involves the t test statistic $T = b_{Overall}/se(b_{Overall})$. The power formula for the overall test is based on using $b_{Overall}$, $se(b_{Overall})$ and $n_{Overall}$ in the power formula of the t test (given by equation (15)). Expressing the overall parameters in terms of stratum-specific parameters (see equation (27), and with $n_{Overall} = 4n$) and utilizing equations (10) and (11) yields:

$$Pwr_{Overall}\left(\alpha, (R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right) =$$

$$= t_{4n-2}\left(-t_{4n-2,1-\frac{\alpha}{2}} - \sqrt{n}\,\frac{\sum_i \frac{|R_i|}{(1-R_i^2)}}{\sqrt{\sum_i \frac{1}{(1-R_i^2)}}} \right)$$

$$+ t_{4n-2}\left(-t_{4n-2,1-\frac{\alpha}{2}} + \sqrt{n}\,\frac{\sum_i \frac{|R_i|}{(1-R_i^2)}}{\sqrt{\sum_i \frac{1}{(1-R_i^2)}}} \right) \quad \textit{(32)},$$

where *i* is iteratively expressed as M≤50, F≤50, M>50 and F>50.

Joint test:

The joint test involves a chi-square test statistic C_{Joint} (see equation (28)). The power of the joint test is given by

$$Pwr_{Joint}(\alpha) = P\left(C_{Joint} \geq \chi_{4,1-\alpha} | H_A\right) = 1 - \mathrm{X}^2_{4,\lambda}\left(\chi_{4,1-\alpha}\right) \quad \textit{(33)},$$

where $\chi_{4,q}$ is the q-th quantile of a chi-square distribution with 4 df and $\mathrm{X}^2_{4,\lambda}$ is the cumulative distribution function of a non-central chi-square distribution with 4 df and non-centrality parameter λ that can be calculated as follows:$\lambda = \sum_i \left(\frac{b_i}{se(b_i)}\right)^2$, where *i* is iteratively expressed as M≤50, F≤50, M>50 and F>50. Utilizing equations (10) and (11) yields

$$\lambda\left((R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right) = \sum_i \frac{nR_i^2}{1-R_i^2} \quad \textit{(34)},$$

where *i* is iteratively expressed as M≤50, F≤50, M>50 and F>50 and n is the sample size of a single stratum. Thus, the power for the joint test is given as follows:

$$Pwr_{Joint}\left(\alpha, (R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right) = 1 - \mathrm{X}^2_{4,\lambda\left((R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right)}\left(\chi_{4,1-\alpha}\right) \quad \textit{(35)}.$$

3.1.6.2 Power formulae for the marginal tests

Each of the marginal tests is a combination of an age- and a sex-version of the respective test. For example, the marginal difference test simultaneously conducts a marginal age-difference test (to compare age-marginal effects $b_{\leq 50}$ and $b_{>50}$) and a marginal sex-difference test (to compare sex-marginal effects b_M and b_F). To obtain the power formula for a marginal test (in dependence of stratum-specific explained variances of the four strata), power formulae for the respective age- and the sex-version of the test have to be derived, and then have to be combined. The derivations of power formulae are exemplified on the age-versions and can easily be transferred to the respective sex-version.

Marginal age-difference test:

The marginal age-difference test involves the z statistic $Z_{Agediff}$ that corresponds to a basic stratum-difference test statistic (as given by equation (5)). The power formula for the marginal age-difference test is based on using the age-marginal effects $b_{\leq 50}$, $b_{>50}$ with $se(b_{\leq 50})$ and $se(b_{>50})$ in the power formula of the difference test (equation (13)). Expressing the age-marginal parameters in terms of stratum-specific parameters (see equation (8), and with $n_{\leq 50} = n_{>50} = 2n$) and utilizing equations (10) and (11) yields:

$$Pwr_{Agediff}\left(\alpha, (R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right)$$

$$= \Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{n}\,\frac{\frac{\sum_j \frac{|R_j|}{(1-R_j^2)}}{\sum_j \frac{1}{(1-R_j^2)}} - \frac{\sum_k \frac{|R_k|}{(1-R_k^2)}}{\sum_k \frac{1}{(1-R_k^2)}}}{\sqrt{\sum_j \frac{1}{(1-R_j^2)} + \sum_k \frac{1}{(1-R_k^2)}}}\right)$$

$$+ \Phi\left(-z_{1-\frac{\alpha}{2}} + \sqrt{n}\,\frac{\frac{\sum_j \frac{|R_j|}{(1-R_j^2)}}{\sum_j \frac{1}{(1-R_j^2)}} - \frac{\sum_k \frac{|R_k|}{(1-R_k^2)}}{\sum_k \frac{1}{(1-R_k^2)}}}{\sqrt{\sum_j \frac{1}{(1-R_j^2)} + \sum_k \frac{1}{(1-R_k^2)}}}\right) \quad \textit{(36)},$$

where *j* is iteratively expressed as M≤50 and F≤50, and *k* is iteratively expressed as M>50 and F>50. The formula can equally be used to derive the power of the *marginal sex-difference test*, in what case, *j* is iteratively expressed as M≤50 and M>50, and *k* is iteratively expressed as F≤50 and F>50.

Marginal age-stratified test:

The marginal age-stratified test involves two age-marginal t test statistics $T_{\leq 50} = \frac{b_{\leq 50}}{se(b_{\leq 50})}$ and $T_{>50} = \frac{b_{>50}}{se(b_{>50})}$. Since each of the age-marginal effects is obtained from pooling two respective stratum-specific effects (e.g., $b_{\leq 50}$ is obtained from meta-analysis of $b_{M\leq 50}$ and $b_{F\leq 50}$), the power of a single age-marginal t test in dependence of respective stratum-specific explained variances is similar to the power of the 2-strata overall test (which is also a meta-analysis of two strata): $Pwr_{\leq 50}\left(\alpha, (R^2_{M\leq 50}, R^2_{F\leq 50}), n\right) = Pwr_{Overall}\left(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, n, 0.5\right)$; and $Pwr_{>50}\left(\alpha, (R^2_{M>50}, R^2_{F>50}), n\right) = Pwr_{Overall}\left(\alpha, R^2_{M>50}, R^2_{F>50}, n, 0.5\right)$. To obtain the power of the marginal age-stratified test (combination of the two age-marginal t tests), the power formulae of the two age-marginal t tests have to be combined and the α-levels have to be corrected for the multiple testing of two t tests. Leveraging independence of the age-marginal t tests (due

to using independent subjects), the power of the marginal age-stratified test is given as follows:

$$Pwr_{Agestrat}(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) = Pwr_{Overall}\left(\frac{\alpha}{2}, R^2_{M\leq 50}, R^2_{F\leq 50}, n, 0.5\right) + Pwr_{Overall}\left(\frac{\alpha}{2}, R^2_{M>50}, R^2_{F>50}, n, 0.5\right) - Pwr_{Overall}\left(\frac{\alpha}{2}, R^2_{M\leq 50}, R^2_{F\leq 50}, n, 0.5\right) Pwr_{Overall}\left(\frac{\alpha}{2}, R^2_{M>50}, R^2_{F>50}, n, 0.5\right) \quad (37).$$

The formula can equally be used to derive the power of the *marginal sex-stratified test*, in what case the subgroups for the overall power have to be defined by sex instead of by age group.

Marginal age-joint test:

The marginal age-joint test involves the chi-square statistic $C_{Agejoint}$ that corresponds to a basic 2-stata joint test statistic (given by equation (9)). The power formula for the marginal age-joint test is based on using the age-marginal effects $b_{\leq 50}$, $b_{>50}$ with $se(b_{\leq 50})$ and $se(b_{>50})$ in the power formula of the 2-strata joint test (equation (19)). Expressing the age-marginal parameters in terms of stratum-specific parameters (see equation (8), and with $n_{\leq 50} = n_{>50} = 2n$) and utilizing equations (10) and (11) yields:

$$Pwr_{Agejoint}\left(\alpha, (R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right) = 1 - \mathrm{X}^2_{2,\lambda\left((R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right)}\left(\chi_{2,1-\alpha}\right) \quad (38)$$

, with non-centrality parameter λ that is given by

$$\lambda\left((R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}), n\right) = n\left(\frac{\sum_j \frac{|R_j|}{(1-R_j^2)}}{\sqrt{\sum_j \frac{1}{(1-R_j^2)}}}\right)^2 + n\left(\frac{\sum_k \frac{|R_k|}{(1-R_k^2)}}{\sqrt{\sum_k \frac{1}{(1-R_k^2)}}}\right)^2 \quad (39),$$

where *j* is iteratively expressed as M≤50 and F≤50, and *k* is iteratively expressed as M>50 and F>50. The formula can equally be used to derive the power of the *marginal sex-joint test*, in what case, *j* is iteratively expressed as M≤50 and M>50, and *k* is iteratively expressed as F≤50 and F>50.

Combining marginal age- and marginal sex-specific power formulae:

The power for the marginal difference, the marginal stratified, and the marginal joint test is derived by combining the age- and the sex-version of the respective test using an OR relationship, while correcting the α-levels of the age- and the sex-version for the multiple testing of two versions:

$$Pwr_{MarDiff}\left(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n\right) = Pwr_{Agediff\cup Sexdiff}\left(\frac{\alpha}{2}, \dots\right)$$
$$= Pwr_{Agediff}\left(\frac{\alpha}{2}, \dots\right) + Pwr_{Sexdiff}\left(\frac{\alpha}{2}, \dots\right) - Pwr_{Agediff\cap Sexdiff}\left(\frac{\alpha}{2}, \dots\right) \quad (40)$$

$$Pwr_{MarStrat}\left(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n\right) = Pwr_{Agestrat\cup Sexstrat}\left(\frac{\alpha}{2}, \dots\right)$$
$$= Pwr_{Agestrat}\left(\frac{\alpha}{2}, \dots\right) + Pwr_{Sexstrat}\left(\frac{\alpha}{2}, \dots\right) - Pwr_{Agestrat\cap Sexstrat}\left(\frac{\alpha}{2}, \dots\right) \quad (41)$$

$$Pwr_{MarJoint}\left(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n\right) = Pwr_{Agejoint\cup Sexjoint}\left(\frac{\alpha}{2}, \dots\right)$$
$$= Pwr_{Agejoint}\left(\frac{\alpha}{2}, \dots\right) + Pwr_{Sexjoint}\left(\frac{\alpha}{2}, \dots\right) - Pwr_{Agejoint\cap Sexjoint}\left(\frac{\alpha}{2}, \dots\right) \quad (42)$$

The power can be computed analytically if the age- and the sex-version of the test are statistically independent. For example, assuming independence of the marginal age-difference and the marginal sex-difference test, the power of the intersection can be computed as follows: $Pwr_{Agediff\cap Sexdiff} = Pwr_{Agediff} \cdot Pwr_{Sexdiff}$. The simulated data sets (initially created for the type 1 error evaluation of approaches) were utilized to evaluate whether the respective age- and sex-versions of the marginal tests are dependent or independent. In case of dependence, the power of the intersection has to be obtained from simulated probability distributions.

3.1.6.3 Power formulae for the approaches

To obtain the power formula for an approach, the derived power formulae for the single statistical tests (i.e., steps implemented in the approach) have to be combined.

As for the 2-strata configuration, for independent steps, the power of an approach is a product of the power of the implemented steps (see equation (22)). However, for dependent steps, the power for the intersection of dependent steps has to be obtained from simulated probability distributions. Again, the simulated data sets (initially created for the type 1 error evaluation of approaches) were utilized to evaluate statistical dependence of steps.

3.1.6.4 Details of the power comparisons

Analytical power formulae were derived for the considered approaches. In order to compute the power, assumptions about stratum-specific sample sizes, explained variances and α-levels have to be made. Generally, the parameters were defined to reflect realistic power computation scenarios that are similar to the GIANT consortium age- and sex-stratified GWAMA scenario.

A total balanced sample size of 200,000 was assumed to be equally distributed among the four age- and sex-specific strata.

Different types of age- and sex-specific effects were modelled based on known and realistic genetic effect sizes taken from GIANT consortium GWAMAs on WHR_{adjBMI}: A small (rs6784615 near *STAB1*), a medium (rs4684854 near *PPARG*) and a large (rs2820443 near *LYPLAL1*) genetic effect on WHR_{adjBMI} that explained 0.014%, 0.058% and 0.167% of the WHR_{adjBMI} variation in women, respectively (Heid et al., 2010; Randall et al., 2013).

The α-levels for the filtering tests were arbitrarily set to $\alpha_1 = 1 \times 10^{-5}$. For each approach that involves filtering, the number of independent lead-SNPs passing the filtering (=*M*) had to be estimated from the real GIANT AGE x SEX data set (see chapter 4.1.4 for details) (Winkler et al, in revision). This was required to obtain the Bonferroni-corrected α-level for the final difference-of-difference test ($\alpha_{Bonf} = 0.05/M$).

3.2 Results

The following chapters show results from the structured methodological comparison of age- and sex-stratified GWAMA approaches to identify 3-way G x AGE x SEX interaction effects. First, approaches were validated with regards to type 1 error (using simulation, chapter 3.1.5) and second, approaches were compared with regards to power (using analytical computations, chapter 3.2.2).

3.2.1 Simulation-based evaluation of type 1 error

To evaluate approaches with regards to type 1 error, simulation-based estimations of type 1 error rates were performed.

As for the 2-strata configuration, prior to investigating type 1 error of the approaches, the simulated data sets were validated, i.e., shown to truly reflect the implied null hypotheses of 'No age-difference in sex-difference of genetic effects' (see **Appendix 9.4**).

Type 1 error rate (T1ER) of an approach was estimated by calculating the number of nominally significant difference-of-difference test results among all conducted difference-of-difference tests. Assuming the null hypothesis of 'No age-difference in sex-difference of genetic effects' and a 5% α-level one would expect that 5% of difference-of-difference tests reach nominal significance. Increased numbers of nominally significant test outcomes pinpoint invalid approaches.

As for the 2-strata configuration, filtering on overall association did not inflate T1ER, but the stratified and the joint test increased the T1ER of the difference-of-difference test (for all modelled scenarios, **Table 6**).

3. Screening for 3-way G x AGE x SEX interaction

Surprisingly, approaches that involve filtering on marginal tests yielded valid T1ER for the difference-of-difference test (for all of the modelled scenarios, **Table 6**). Obviously, comparing marginal effects (marginal age-effects, independent of sex; or marginal sex-effects, independent of age) cannot negatively impact T1ER of the difference-of-difference test. The impact of marginal filtering tests on power of the difference-of-difference test has to be investigated.

***Table 6**. Simulation-based type 1 error rates (T1ER) for the age- and sex-stratified GWAMA approaches.*

			$H_0^{noeffect}$			H_0^{effect}			$H_0^{sexdiff}$		
Approach	**SNP**	**MAF**	**#DiffDiff-Tests[a]**	**#DiffDiff <0.05[b]**	**T1ER** [%]	**#DiffDiff-Tests[a]**	**#DiffDiff <0.05[b]**	**T1ER** [%]	**#DiffDiff-Tests[a]**	**#DiffDiff <0.05[b]**	**T1ER** [%]
$[\mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	1,000,000	50,044	5.00	1,000,000	50,464	5.05	1,000,000	50,258	5.03
	rs8138968	0.3	1,000,000	49,996	5.00	1,000,000	49,872	4.99	1,000,000	49,920	4.99
	rs6007738	0.5	1,000,000	49,986	5.00	1,000,000	50,044	5.00	1,000,000	50,398	5.04
$[\mathrm{Overall}_{0.05} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	50,596	2,623	5.18	999,795	50,448	5.05	82,331	4,119	5.00
	rs8138968	0.3	50,499	2,522	4.99	999,848	49,869	4.99	53,353	2,634	4.94
	rs6007738	0.5	50,578	2,604	5.15	999,863	50,036	5.00	53,192	2,736	5.14
$[\mathrm{Strat}_{0.05} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	50,119	13,193	**26.32**	976,618	50,416	5.16	973,883	50,215	5.16
	rs8138968	0.3	49,934	13,140	**26.31**	977,357	49,845	5.10	977,206	49,894	5.11
	rs6007738	0.5	49,673	12,979	**26.13**	978,424	50,029	5.11	975,798	50,374	5.16
$[\mathrm{Joint}_{0.05} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	50,924	15,831	**31.09**	997,193	50,451	5.06	996,886	50,250	5.04
	rs8138968	0.3	50,693	15,780	**31.13**	997,792	49,870	5.00	997,763	49,910	5.00
	rs6007738	0.5	50,714	15,648	**30.86**	997,981	50,044	5.01	997,601	50,395	5.05
$[\mathrm{MarDiff}_{0.05} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	49,414	2,570	5.20	49,568	2,499	5.04	999,288	50,221	5.03
	rs8138968	0.3	50,115	2,595	5.18	49,884	2,538	5.09	999,549	49,891	4.99
	rs6007738	0.5	49,756	2,555	5.14	49,765	2,640	5.30	999,531	50,380	5.04
$[\mathrm{MarStrat}_{0.05} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	45,080	2,298	5.10	996,997	50,304	5.05	992,370	49,846	5.02
	rs8138968	0.3	45,384	2,348	5.17	997,447	49,759	4.99	994,216	49,590	4.99
	rs6007738	0.5	45,216	2,299	5.08	997,642	49,923	5.00	993,903	50,127	5.04
$[\mathrm{MarJoint}_{0.05} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$	rs6002481	0.05	41,773	2,118	5.07	998,569	50,390	5.05	997,897	50,147	5.03
	rs8138968	0.3	41,986	2,195	5.23	998,900	49,818	4.99	998,462	49,831	4.99
	rs6007738	0.5	41,687	2,098	5.03	999,026	49,994	5.00	998,424	50,317	5.04

[a] Number of SNPs tested for difference-of-difference (i.e., number of SNPs passing the filtering steps, M_{filt})
[b] Number of SNPs with nominal significant difference-of-difference ($P_{DiffDiff} < 0.05$)

3.2.2 Analytical power comparison

To compare performance of approaches and to give recommendations for specific types of 3-way G x AGE x SEX interaction effects, analytical power computations were performed for various realistic scenarios.

Approaches that were shown to violate type 1 error requirements were omitted from power computations. A systematic scheme of approaches that were compared for power is shown in **Figure 16**.

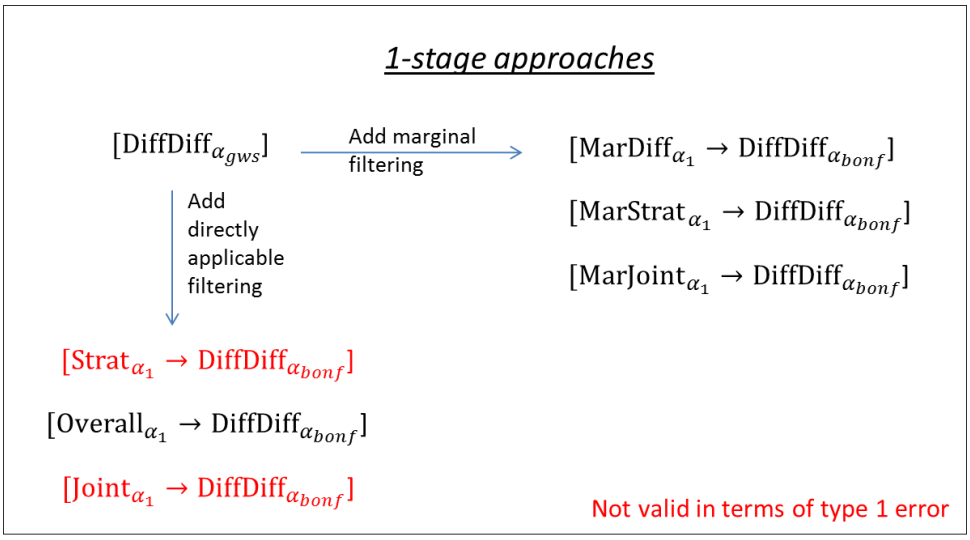


Figure 16. *Systematic scheme of age- and sex-stratified GWAMA approaches to identify 3-way G x AGE x SEX interaction effects that are considered for the power comparisons (α_{gws}= 5 x 10^{-8}; α_1 = α-level for the filtering step; α_{Bonf} = Bonferroni-corrected α-level).*

Simulation-based inference of statistical dependence between tests showed independence (i) between the difference-of-difference and the overall test, and (ii) between the marginal age-difference and the marginal sex-difference test (see **Appendix 9.5**). As a consequence, equation (22) could be used to calculate power of approaches [Overall$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$] and [MarDiff$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]. In contrast, the marginal age-stratified and the marginal sex-stratified test, as well as the marginal age-joint and the marginal sex-joint test were shown to be dependent. Thus, simulated probability distributions had to be utilized to estimate the power of the intersection of tests $Pwr_{AgeStrat \cap SexStrat}$ and $Pwr_{AgeJoint \cap SexJoint}$ (see **Appendix 9.6**).

A list of power formulae including details about step-specific α-levels and sample sizes is shown in **Table 7.**

Table 7. *Analytical power formulae for the considered age- and sex-stratified GWAMA approaches. To calculate the Bonferroni-corrected α-level for the final difference-of-difference test (α_{Bonf} = 0.05/M) of approaches that involve filtering steps, the number of independent lead SNPs tested (=M) was estimated from the real GIANT AGExSEX data set.*

Approach	**M (AGExSEX)**	**Analytical power formula**
[DiffDiff$_{5e\text{-}8}$]	-	$Pwr_{DiffDiff}(\text{5e-8}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n)$
[Overall$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	135	$Pwr_{Overall}(\text{1e-5}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) \cdot Pwr_{DiffDiff}(\alpha_{bonf}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n)$
[MarDiff$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	22	$Pwr_{MarDiff}(\text{1e-5}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) \cdot Pwr_{DiffDiff}(\alpha_{bonf}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n)$ with: $Pwr_{MarDiff}(\alpha, \dots) = Pwr_{AgeDiff}\left(\frac{\alpha}{2}, \dots\right) + Pwr_{SexDiff}\left(\frac{\alpha}{2}, \dots\right) - Pwr_{AgeDiff}(\frac{\alpha}{2}, \dots) \cdot Pwr_{SexDiff}(\frac{\alpha}{2}, \dots)$
[MarStrat$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	95	$Pwr_{MarStrat}(\text{1e-5}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) \cdot Pwr_{DiffDiff}(\alpha_{bonf}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n)$ with: $Pwr_{MarStrat}(\alpha, \dots) = Pwr_{AgeStrat}\left(\frac{\alpha}{2}, \dots\right) + Pwr_{SexStrat}\left(\frac{\alpha}{2}, \dots\right) - Pwr_{AgeStrat \cap SexStrat}(\frac{\alpha}{2}, \dots)$
[MarJoint$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	120	$Pwr_{MarJoint}(\text{1e-5}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) \cdot Pwr_{DiffDiff}(\alpha_{bonf}, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n)$ with: $Pwr_{MarJoint}(\alpha, \dots) = Pwr_{AgeJoint}\left(\frac{\alpha}{2}, \dots\right) + Pwr_{SexJoint}\left(\frac{\alpha}{2}, \dots\right) - Pwr_{AgeJoint \cap SexJoint}(\frac{\alpha}{2}, \dots)$

3. Screening for 3-way G x AGE x SEX interaction

First, power was compared for the biologically most plausible 1-stratum and for 3-strata interaction effects. To accomplish this, power for the five considered approaches was depicted over an increasing modelled effect size: To reflect the 1-stratum interaction, the modeled effect was set to be present in the F≤50 group and lacking in all other strata, and to reflect the 3-strata interaction, the modeled effect was set to be equally present in the F≤50, M≤50 and F>50 groups and lacking in M>50 (**Figure 17**).

Surprisingly, the marginal filtering approaches $[\mathrm{MarStrat}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ performed best for both types of interaction effects. In comparison, approach $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ displayed similar power for 3-strata interaction effects, but lower power for 1-stratum interaction effects.

Notably, only 1-stratum and 3-strata interaction effects involving large stratum-specific genetic effect sizes (comparable to the genetic effect on WHR_{adjBMI} near *LYPLAL1,* $R^2 = 0.167\%$) can be detected efficiently by the given setting (e.g., power = 85.1% to find large 1-stratum interaction with approach $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$, **Table 8**).

For a medium modelled effect size (comparable to the genetic effect on WHR_{adjBMI} near *PPARG,* $R^2 = 0.058\%$), none of the approaches displayed sufficient power to identify plausible 3-way interaction effects (power < 22%, **Table 8**).

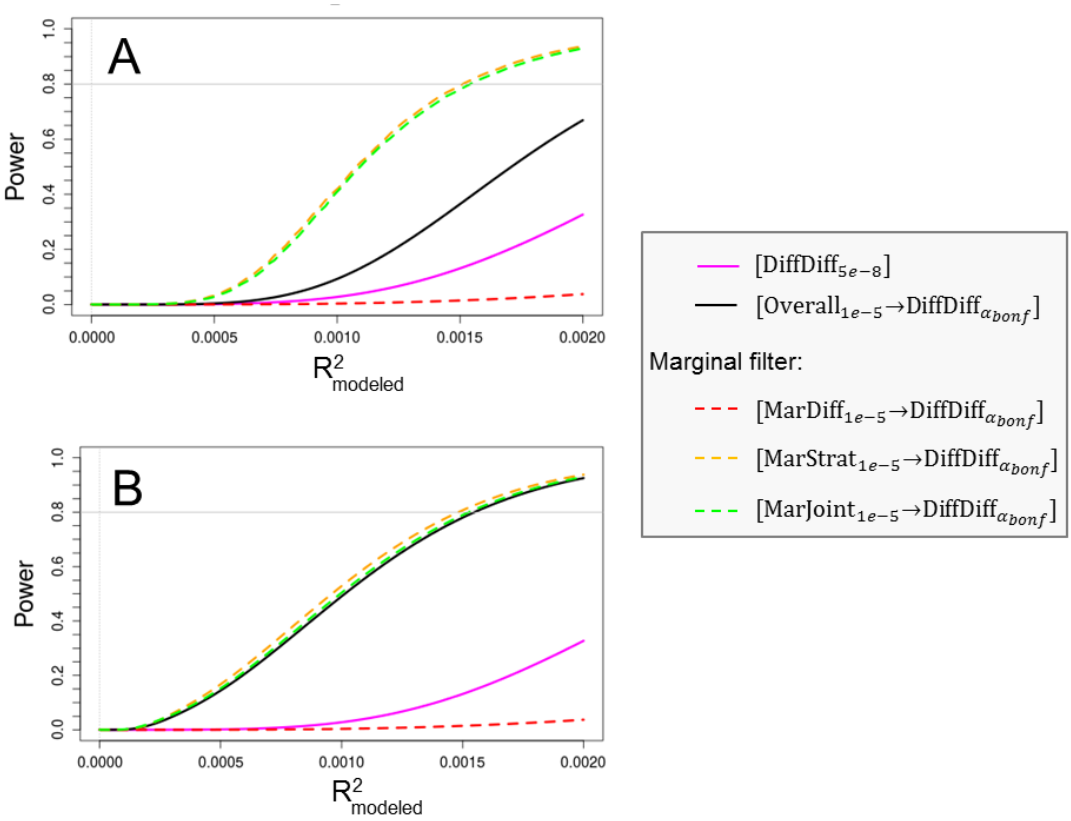


Figure 17. *Power to detect plausible 3-way interaction effects. A: Power to detect 1-stratum interaction over modeled effect size in F≤50 and assuming no effect in all other strata. B: Power to detect 3-strata interaction over modeled and identical effect sizes in F≤50, F>50 and M≤50 and assuming no effect in M>50.*

Table 8. *Power to detect plausible 3-way interaction effects (total sample size = 200,000).*

Approach	Modeled R^2 [%]	Power	
		1-stratum interaction	3-strata interaction
[DiffDiff$_{\alpha_{gws}}$]	0.014 (*STAB1*)	0	0
	0.058 (*PPARG*)	0.3	0.3
	0.167 (*LYPLAL1*)	18.9	19.0
[Overall$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	0.014 (*STAB1*)	0	0.4
	0.058 (*PPARG*)	0.8	19.3
	0.167 (*LYPLAL1*)	47.5	84.4
[MarStrat$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	0.014 (*STAB1*)	0	0.4
	0.058 (*PPARG*)	6.3	22.0
	0.167 (*LYPLAL1*)	85.8	86.6
[MarJoint$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	0.014 (*STAB1*)	0	0.6
	0.058 (*PPARG*)	6.0	20.2
	0.167 (*LYPLAL1*)	85.1	85.2
[MarDiff$_{1e\text{-}5}$→DiffDiff$_{\alpha_{Bonf}}$]	0.014 (*STAB1*)	0	0
	0.058 (*PPARG*)	0	0
	0.167 (*LYPLAL1*)	2.1	2.1

3. Screening for 3-way G x AGE x SEX interaction

To further investigate the characteristics of approaches in greater detail and to illustrate their performance for other 3-way interaction effects, power was depicted in heatplots (**Figure 18**). For each heatplot, the effect size in F≤50 was set to a realistic genetic effect size and the effect size in M>50 was set to zero. The effect sizes in M≤50 and F>50 were varied on the x- and y-axis, respectively. This strategy allowed to depict power for a range of interesting types of interaction effects (see **Figure 18A** for a map of interesting heatplot regions).

The figure shows that for a medium modelled genetic effect (comparable to the genetic effect on WHR_{adjBMI} in the *PPARG* region), the power of approaches $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{MarDiff}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ is low for any type of 3-way interaction effect.

The genome-wide screening approach $[\mathrm{DiffDiff}_{\alpha_{\mathrm{gws}}}]$ has good power to identify extreme 3-way interaction effects (see 3rd quadrant of **Figure 18B**) but lacks power for any other types.

The behavior of approaches $[\mathrm{MarStrat}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ is similar across all types of 3-way interaction effects (**Figure 18E/F)**. However, approach $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ shows slightly better power in the 2nd, 3rd and 4th quadrant. For example, the power to identify an effect that is positive in F≤50, negative in M≤50 and lacking in the other two strata, is 67.6% for approach $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ and 34.8% for approach $[\mathrm{MarStrat}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$.

In summary, the power computations found approach $[\mathrm{DiffDiff}_{\alpha_{\mathrm{gws}}}]$ to be optimal for extreme 3-way interaction effects (see 3rd quadrant of the heatplots) and approach $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ to be optimal for plausible 1-stratum and 3-strata interaction effects (see regions a) and b) in the heatplots) as well as for the less extreme 3-way interaction effects (see 2nd and 4th quadrant in the heatplots).

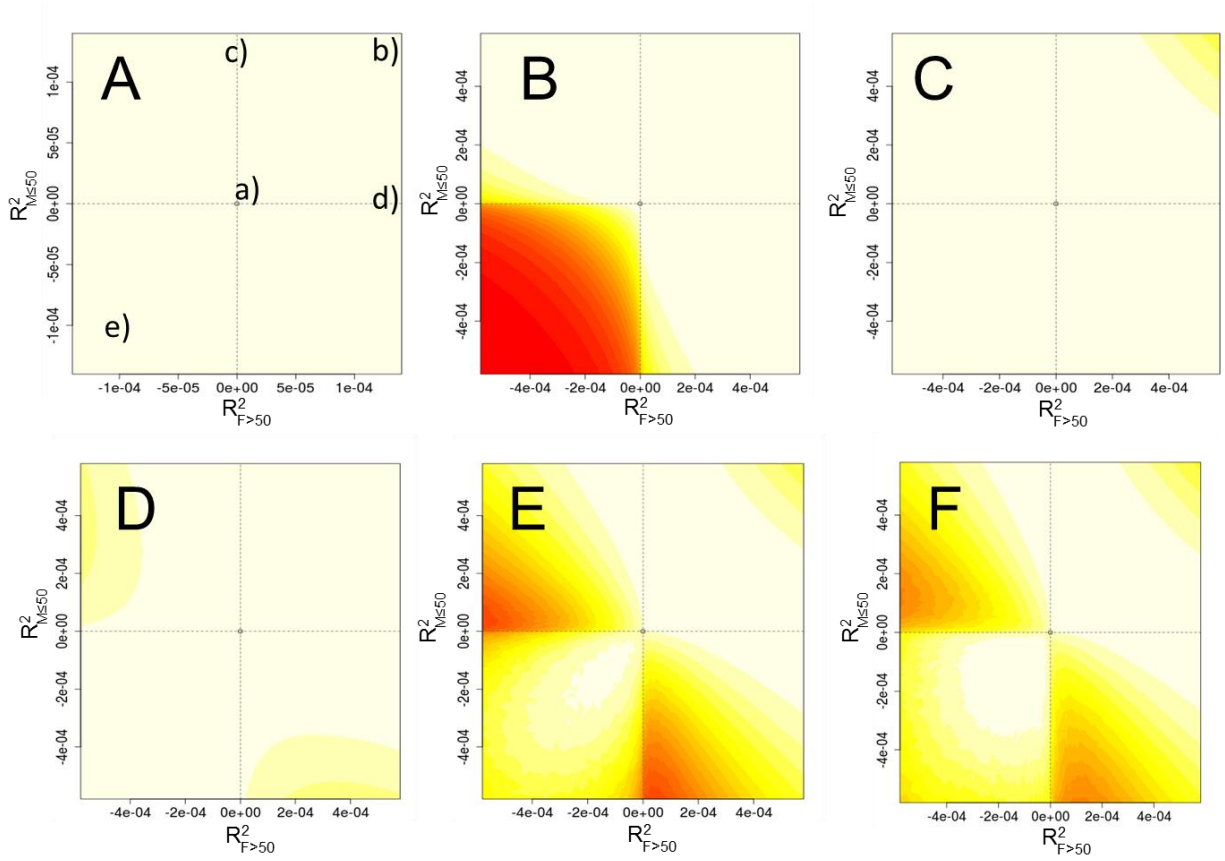


Figure 18. Power heatplots. *Shown is the power to identify 3-way interaction for the five considered age- and sex-stratified GWAMA approaches. Power is drawn for a fixed effect size in F≤50 (set to the medium genetic effect on WHR_{adjBMI}, near PPARG), a fixed effect size in M>50 (set to zero) and over varying effect sizes in M≤50 (y-axis) and F>50 (x-axis). A: Regions of interest: a) 1-stratum interaction, b) 3-strata interaction, c) 2-way age-differential SSE (effect in younger, none in older), d) 2-way sexually dimorphic SSE (effect in women, none in men), and e) extreme 3-way interaction. Power is displayed for B: $[DiffDiff_{\alpha_{gws}}]$, C: $[Overall_{1e\text{-}5} \rightarrow DiffDiff_{\alpha_{Bonf}}]$. D: $[MarDiff_{1e\text{-}5} \rightarrow DiffDiff_{\alpha_{Bonf}}]$, E: $[MarJoint_{1e\text{-}5} \rightarrow DiffDiff_{\alpha_{Bonf}}]$, and F: $[MarStrat_{1e\text{-}5} \rightarrow DiffDiff_{\alpha_{Bonf}}]$.*

4 Application to stratified GWAMAs for obesity traits

In order to explore the performance of approaches using real data and to systematically screen for potential stratum-difference in the genetic effects of central obesity, the best approaches were applied to real stratified GWAMA data sets for WHR_{adjBMI} from the GIANT consortium.

First, in order to identify sex-difference in genetic effects for WHR_{adjBMI}, the best approaches were applied to the balanced sex-stratified GIANT GENDER project data set for WHR_{adjBMI} (Randall et al., 2013). Results were compared between best approaches and related to the reported results as well as to the originally applied approaches (Randall et al., 2013).

Second, to explore approaches using a real unbalanced data set and to identify difference in genetic effects for WHR_{adjBMI} between smokers and non-smokers, the best approaches were applied to the smoking-status stratified GIANT SMOKING project data set for WHR_{adjBMI} (Justice et al, in progress).

Third, to identify 3-way G x AGE x SEX interaction effects for WHR_{adjBMI} between, best 3-way approaches were applied to the age- and sex-stratified GIANT AGE x SEX project data set for WHR_{adjBMI} (Winkler et al, in revision).

4.1 Materials and Methods

Except using different stratification variables (i.e., sex for GIANT GENDER, smoking-status for GIANT SMOKING, and dichotomized age and sex for GIANT AGE x SEX), similar statistical methods were employed for the conduct of the stratified GWAMAs (see chapter 4.1.1). Descriptions of the project-specific data sets as well as an overview on the applied approaches are given in chapter 4.1.2 (for GIANT GENDER), chapter 4.1.3 (for GIANT SMOKING) and chapter 4.1.4 (for GIANT AGE x SEX).

4.1.1 Statistical analysis for the stratified GWAMA projects of the GIANT consortium

Generally, each project involved three consecutive steps: Study-specific GWAS testing, central quality-control of GWAS summary results and central GWAMAs.

First, study-specific stratified GWAS were conducted. In order to obtain comparable summary GWAS results across studies and strata, an analysis plan was developed by the

consortium analysts and sent out to the study analysts. The plan detailed instructions about phenotype transformations, genotype models and GWAS testing.

Each study analyst stratified the study data by the respective stratification variable, i.e., by sex, by smoking-status or by age and sex (age dichotomized at the age of 50 years). For studies with a case-control design, a further stratification by cases and controls was applied. For each stratum separately, waist-hip ratio (WHR) phenotypes were adjusted for age, age^2 and BMI. To accomplish this, residuals were calculated from a non-linear regression model for WHR that included age, age^2 and BMI. Finally, to obtain homogenously distributed phenotypes across strata and across studies, the stratum-specific residuals were inverse-normal transformed yielding a standard-normal distributed phenotype, defined as WHR_{adjBMI}.

Next, for each stratum, additively modeled SNP genotypes were tested for association with WHR_{adjBMI} via linear regression using MACH2QTL (Y. Li, Willer, Ding, Scheet, & Abecasis, 2010), SNPTEST (Marchini, Howie, Myers, McVean, & Donnelly, 2007), ProbABEL (Aulchenko, Struchalin, & van Duijn, 2010), GenABEL (Aulchenko, Ripke, Isaacs, & van Duijn, 2007), Merlin (Abecasis & Wigginton, 2005), PLINK (Purcell et al., 2007) or QUICKTEST (Kutalik et al., 2011). The obtained stratum-specific GWAS summary results were provided to the consortium analysts via upload to an ftp server.

Second, the collected study-specific GWAS results were validated centrally using an extensive quality control (QC) procedure (Winkler, Day, et al., 2014). The QC involved general file checks, such as checks for issues with phenotype transformations or issues with allele frequencies, as well as exclusion of low quality SNPs, such as exclusion of SNPs with poor imputation quality or with low minor allele count (see chapter 5.2 for details).

Third, for each stratum separately, an inverse-variance weighted meta-analysis assuming a fixed effect model was performed using metal (Willer, Li, & Abecasis, 2010). In order to ensure proper results, each meta-analysis was conducted by two analysts simultaneously, followed by a comparison of results. Again, results were validated using an extensive meta-analysis quality control (QC) procedure (see chapter 5.2 and (Winkler, Day, et al., 2014) for details).

4.1.2 Utilizing the GIANT GENDER data for WHR_{adjBMI} to screen for stratum-difference under a balanced design

The GIANT GENDER data for WHR_{adjBMI} reflects a balanced 2-stage sex-stratified GWAMA results data set that was created according to the methods described in chapter 4.1.1 (Randall et al., 2013).

The discovery stage included sex-stratified GWAS results from 46 HapMap imputed studies (each analyzing up to 2.8M SNPs) and comprised up to 60,586 men and 73,137

women of European descent. Randall and colleagues selected SNPs from discovery stage and followed them up in additional 48 independent replication studies that comprised up to 62,395 men and 74,657 women. The replication data set contained both HapMap imputed studies as well as studies that were genotyped using the custom MetaboChip array (analyzing up to ~195K SNPs)(Voight et al., 2012). The two separate stages of the GIANT GENDER data were directly applicable to the considered 2-stage approaches. To additionally make the GIANT GENDER data applicable to 1-stage approaches, sex-specific inverse-variance weighted meta-analyses of the GIANT GENDER discovery and replication data sets were conducted using metal (Willer et al., 2010).

The best 1-stage approaches, $[\mathrm{Diff}_{5e\text{-}8}]$ and $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ as well as the best 2-stage approaches $[\mathrm{Joint}_{1e\text{-}5}, \mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[[\mathrm{Overall}_{1e\text{-}5}] \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ were applied to the GIANT GENDER data for WHR_{adjBMI}. For each approach, a list of independent lead SNPs with significant sex-differences was extracted according to the technical procedure described in chapter 2.1.4. Results were compared between the best approaches and related to the power computation results.

Moreover, results of the best approaches were related to originally applied approaches and to the reported results (Randall et al., 2013). Since the original systematic screen for sex-difference conducted by Randall and colleagues considered nine anthropometric traits in parallel, the results were not directly comparable to the here applied approaches that focus on WHR_{adjBMI}. Thus (in order to improve comparability), the Randall-like 2-stage approaches $[\mathrm{Strat}_{\alpha_1}, \mathrm{Diff}_{\alpha_2}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ and $[\mathrm{Diff}_{\alpha_1}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ were applied to the WHR_{adjBMI} data set and results were compared with the best approaches.

4.1.3 Utilizing the GIANT SMOKING data for WHR_{adjBMI} to screen for stratum-difference under an unbalanced design

The GIANT SMOKING data for WHR_{adjBMI} reflects an unbalanced smoking-status stratified GWAMA results data set that was created according to the methods described in chapter 4.1.1.

In total, results from 88 HapMap imputed or typed MetaboChip GWA studies were included that comprised up to 37,300 smokers and up to 129,926 non-smokers of European descent. This reflected an unbalanced stratified GWAMA design that implies f = 3.48 (proportion of non-smokers to smokers, i.e., smokers as stratum 1) or f = 0.28 (proportion of smokers to non-smokers, i.e., non-smokers as stratum 1). Due to adopting both, 1- and 2-stage approaches in the GIANT SMOKING project analyses, the data sets were directly applicable to all of the here considered approaches.

Here, the best 1-stage approaches, $[\mathrm{Diff}_{5e\text{-}8}]$ and $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{Bonf}}]$ as well as the best 2-stage approaches $[\mathrm{Joint}_{1e\text{-}5}, \mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$ and $[[\mathrm{Overall}_{1e\text{-}5}] \rightarrow \mathrm{Diff}_{\alpha_{Bonf}}]$ were applied to the GIANT SMOKING data for WHR_{adjBMI}. For each approach, a list of independent lead SNPs with significant differences in genetic effects between smokers and non-smokers was extracted according to the technical procedure described in chapter 2.1.4. Results were compared between the best approaches and related to the power computation results.

4.1.4 Utilizing the GIANT AGE x SEX data for WHR_{adjBMI} to screen for G x AGE x SEX interaction

The GIANT AGE x SEX data for WHR_{adjBMI} reflects an age- and sex-stratified GWAMA results data set that was created according to the methods described in chapter 4.1.1.

For WHR_{adjBMI}, age- and sex-stratified GWAS results from 84 HapMap imputed or typed MetaboChip studies were analyzed that comprise up to 216,654 individuals of European descent. Due to adopting a 1-stage design in the GIANT AGE x SEX project analyses, the data set was directly applicable to the here considered 1-stage approaches.

The best approaches, $[\mathrm{DiffDiff}_{5e\text{-}8}]$ and $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{Bonf}}]$ were applied to the GIANT AGE x SEX data for WHR_{adjBMI}. For each approach, a list of independent lead SNPs with significant 3-way G x AGE x SEX interaction effects was extracted according to the technical procedure described in chapter 3.1.4. Results were compared between the considered approaches and related to the power computation results.

4.2 Results

To explore difference in genetic effects on WHR_{adjBMI} between men and women (chapter 4.2.1), between smokers and non-smokers (chapter 4.2.2), as well as to identify 3-way G x AGE x SEX interaction effects for WHR_{adjBMI} (chapter 4.2.3), selected approaches were applied to real stratified GWAMA results data for WHR_{adjBMI} from the GIANT consortium.

4.2.1 Identification of sex-differences in genetic effects for WHR_{adjBMI}

In order to screen for genetic effects for WHR_{adjBMI} that are significantly different between men and women, the best 1-stage approaches $[\mathrm{Diff}_{5e\text{-}8}]$ (for OED) and $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{Bonf}}]$ (for SSE and CED, chapter 4.2.1.1) as well as the best 2-stage approaches $[\mathrm{Joint}_{1e\text{-}5}, \mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$ (for OED and SSE) and $[[\mathrm{Overall}_{1e\text{-}5}] \rightarrow \mathrm{Diff}_{\alpha_{Bonf}}]$ (for CED, chapter 4.2.1.2) were applied to the GIANT GENDER data set for WHR_{adjBMI}.

Furthermore, results were compared between stage designs (chapter 4.2.1.3) and compared to approaches $[\mathrm{Strat}_{1e\text{-}5}, \mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$ and $[\mathrm{Diff}_{1e\text{-}5}] \rightarrow [\mathrm{Diff}_{\alpha_{Bonf}}]$ that are

similar to the originally employed approaches by Randall and colleagues (chapter 4.1.4) (Randall et al., 2013).

4.2.1.1 The best 1-stage approaches

In total, the best 1-stage approaches $[\mathrm{Diff}_{5e\text{-}8}]$ and $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ identified 10 independent loci with significant sex-differences in the genetic effects for WHR_{adjBMI} (**Table 9**). Nine of the 10 loci showed a stronger effect in women: Seven being women SSE signals that completely lack association in men (near *SLC30A10, COBLL1, PPARG, PLXND1, TNFAIP8, VEGFA, NKX3-1*) and two being women CED signals that show an effect in women accompanied by a less pronounced (but nominally significant and concordant) effect in men (near *ADAMTS9, ITPR2*). One of the 10 loci displayed significant effects in both men and women, yet with opposite effect directions (near *LRRC69*).

Consistent with the results from the power computations, the two CED signals were only identified by approach $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and the OED signal was only detected by approach $[\mathrm{Diff}_{5e\text{-}8}]$. While four of the seven women SSE signals were found by both 1-stage approaches, the other three SSE signals were only detected by $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$. This is again consistent with the power comparison results that demonstrated that approach $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ is advantageous over $[\mathrm{Diff}_{5e\text{-}8}]$ for SSE signals.

The here identified loci include all of the six sexually dimorphic WHR_{adjBMI} loci reported by Randall and colleagues as well as four novel sexually dimorphic loci (near *PLXND1, NKX3-1, LRRC69, ITPR2*) that were missed by Randall and colleagues (Randall et al., 2013). This can be attributed to the more optimal approaches but also to the different SNP selection method applied by Randall and colleagues that involved a False-Discovery-Rate (FDR) approach to correct for the multiple testing of nine considered anthropometric and obesity phenotypes (thus results were not directly comparable).

Table 9. *Ten loci with sexually dimorphic effects for WHR_{adjBMI} detected by the best 1-stage approaches [$Diff_{5e-8}$] (Diff) and [$Overall_{1e-5} \rightarrow Diff_{\alpha_{Bonf}}$] (OverallDiff). To indicate loci detected by both approaches but with different lead SNPs, independent loci are distinguished by alternating shading. The table is sorted by chromosome and position.*

								Men			Women		
Gene[a]	**SNP**	**Approach**	**Type**	**EA**[b]	**EAF**[c]	**$P_{Sexdiff}$**	**$P_{Overall}$**	*Beta*	*P*	*N*	*Beta*	*P*	*N*
SLC30A10	chr1:217820132	Diff OverallDiff	SSE	T	0.72	1.2E-16	1.2E-23	0.005	0.37	76,626	0.064	4.6E-37	98,352
COBLL1	chr2:165236870[d]	OverallDiff	SSE	G	0.60	6.5E-16	2.7E-17	3.0E-04	0.95	75,573	0.054	1.8E-31	97,142
	chr2:165247907[d]	Diff	SSE	T	0.59	1.0E-16	1.4E-14	-0.003	0.61	76,594	0.052	2.0E-29	98,322
PPARG	chr3:12463882	OverallDiff	SSE	C	0.43	2.2E-06	2.1E-09	0.004	0.41	74,653	0.037	4.2E-14	96,473
*PLXND1**	chr3:130816923[e]	OverallDiff	SSE	A	0.79	3.2E-07	1.4E-09	0.003	0.68	74,655	0.044	7.5E-15	96,056
	chr3:130822305[e]	Diff	SSE	T	0.79	5.3E-09	1.5E-07	-0.007	0.33	44,701	0.048	2.4E-14	73,628
ADAMTS9	chr3:64676186	OverallDiff	CED	A	0.70	4.1E-05	4.9E-20	0.018	9.4E-04	75,590	0.047	5.3E-21	97,150
TNFAIP8	chr5:118757185	OverallDiff	SSE	C	0.71	5.2E-06	2.6E-08	0.003	0.55	75,584	0.037	7.9E-13	97,136
VEGFA	chr6:43872529	Diff OverallDiff	SSE	T	0.47	7.1E-12	1.6E-22	0.010	0.05	75,704	0.060	2.4E-31	97,269
*NKX3-1**	chr8:23659269	OverallDiff	SSE	A	0.77	6.5E-07	3.8E-07	-5.0E-04	0.94	75,598	0.038	3.4E-12	97,157
*LRRC69**	chr8:92217371	Diff	OED	T	0.67	4.1E-08	0.95	-0.026	4.4E-05	44,792	0.019	5.4E-04	73,744
*ITPR2**	chr12:26361436	OverallDiff	CED	T	0.25	1.4E-04	2.1E-13	0.013	0.02	74,269	0.042	4.4E-15	95,750

[a] Nearest gene; [b] Effect allele: Chosen to reflect the WHR_{adjBMI} increasing allele in women; [c] Effect allele frequency calculated from the sex-combined sample; [d] r^2 between *COBLL1* lead SNPs is 0.94; [e] r^2 between *PLXND1* lead SNPs is 1.00. * Novel locus compared to Randall and colleagues.

4.2.1.2 The best 2-stage approaches

In total, the best 2-stage approaches $[\text{Joint}_{1e\text{-}5}, \text{Diff}_{0.05}] \rightarrow [\text{Diff}_{\alpha_{Bonf}}]$ and $[[\text{Overall}_{1e\text{-}5}] \rightarrow \text{Diff}_{\alpha_{Bonf}}]$ identified seven loci with significant sex-differences in the genetic effects on WHR_{adjBMI} (**Table 10**).

Of the seven loci, four loci were identified by both approaches (near *SLC30A10, COBLL1, ADAMTS9, VEGFA*) and three have only been detected by approach $[[\text{Overall}_{1e\text{-}5}] \rightarrow \text{Diff}_{\alpha_{Bonf}}]$ (near *PLXND1, ITPR2, HOXC13*). Two of the latter three can be categorized as CED effects, which supports the power analysis result that recommended approach $[[\text{Overall}_{1e\text{-}5}] \rightarrow \text{Diff}_{\alpha_{Bonf}}]$ for CED effects. The remaining five loci can be classified as SSE effects, with one locus being a borderline SSE/CED locus (near *ADAMTS9)*. No OED loci were identified.

While four of the seven identified loci overlap with results reported by Randall and colleagues, three novel loci were identified by approach $[[\text{Overall}_{1e\text{-}5}] \rightarrow \text{Diff}_{\alpha_{Bonf}}]$ (near *PLXND1, ITPR2, HOXC13*) (Randall et al., 2013). Notably, two of the loci identified by Randall and colleagues were completely missed by the here conducted analysis. The main reason for this may be that Randall and colleagues applied a different SNP selection method that involved a FDR approach to correct for the multiple testing of nine considered anthropometric and obesity phenotypes (thus results were not directly comparable).

Table 10. *Seven sexually dimorphic loci for WHR_{adjBMI} detected by the best 2-stage approaches $[Joint_{1e\text{-}5}, Diff_{0.05}] \rightarrow [Diff_{\alpha_{Bonf}}]$ (Joint2Diff), and $[[Overall_{1e\text{-}5}] \rightarrow Diff_{\alpha_{Bonf}}]$ (Overall2Diff). Independent loci are distinguished by alternating shading. The table is sorted by chromosome and position.*

											MEN			WOMEN	
Gene[a]	**SNP**	**Approach**	**Type**	**EA**[b]	**EAF**[c]	**Stage**	**$P_{Sexdiff}$**	**$P_{Overall}$**	**$P_{Sexjoint}$**	**Beta**	**P**	**N**	**Beta**	**P**	**N**
SLC30A10	chr1:217815441[d]	Joint2Diff	SSE	C	0.72	Disc	1.2E-07	3.4E-12	4.5E-17	0.009	0.23	34,599	0.064	8.5E-18	42,732
						Repl	1.1E-09	2.6E-11	2.9E-18	-2.0E-04	0.98	40,935	0.064	1.4E-19	53,799
						Comb	3.1E-16	4.2E-22	9.3E-35	0.005	0.39	75,534	0.064	1.7E-35	96,530
	chr1:217820132[d]	Overall2Diff	SSE	T	0.72	Disc	1.2E-07	1.2E-12	1.7E-17	0.010	0.19	34,601	0.064	3.7E-18	42,735
						Repl	3.2E-10	5.0E-12	1.8E-19	-6.0E-04	0.94	42,025	0.064	1.8E-20	55,617
						Comb	1.2E-16	1.2E-23	1.1E-36	0.005	0.37	76,626	0.064	4.6E-37	98,352
COBLL1	chr2:165221337[e]	Overall2Diff	SSE	T	0.59	Disc	3.0E-06	1.1E-10	2.8E-14	0.009	0.20	34,579	0.053	6.3E-15	42,708
						Repl	1.9E-11	2.3E-07	4.4E-16	-0.012	0.12	40,091	0.054	1.3E-16	53,347
						Comb	6.5E-16	3.2E-16	1.2E-28	-7.0E-04	0.90	74,670	0.053	6.6E-30	96,055
	chr2:165265564[e]	Joint2Diff	SSE	C	0.62	Disc	1.5E-07	6.5E-08	8.6E-13	0.001	0.89	34,572	0.052	9.1E-14	42,712
						Repl	2.3E-10	1.1E-06	2.2E-14	-0.011	0.15	40,998	0.052	4.9E-15	54,425
						Comb	1.5E-16	4.1E-13	3.5E-26	-0.005	0.37	75,569	0.052	3.8E-27	97,136
ADAMTS9	chr3:64679931[f]	Overall2Diff	CED	T	0.70	Disc	0.04	4.0E-08	4.0E-08	0.018	0.02	34,601	0.038	1.1E-07	42,735
						Repl	7.4E-04	3.1E-13	1.1E-14	0.018	0.02	40,995	0.053	1.8E-14	54,422
						Comb	8.6E-05	8.2E-20	6.3E-22	0.018	7.9E-04	75,595	0.046	3.5E-20	97,156
	chr3:64686944[f]	Joint2Diff	SSE	C	0.75	Disc	3.3E-04	2.0E-05	2.4E-07	0.004	0.62	34,589	0.042	3.8E-08	42,708
						Repl	2.1E-03	7.4E-12	6.4E-13	0.019	0.03	40,997	0.052	1.1E-12	54,424
						Comb	1.2E-06	3.6E-15	5.1E-19	0.011	0.06	75,586	0.048	2.9E-19	97,132
*PLXND1**	chr3:130788609	Overall2Diff	SSE	T	0.90	Disc	8.2E-06	1.7E-06	8.1E-10	0.003	0.83	34,520	0.074	9.9E-11	42,670
						Repl	0.25	3.0E-03	6.4E-03	0.014	0.27	42,004	0.032	2.9E-03	55,613
						Comb	7.8E-05	3.6E-08	1.6E-10	0.008	0.36	76,524	0.052	3.8E-11	98,283
VEGFA	chr6:43872529	Joint2Diff Overall2Diff	SSE	T	0.47	Disc	3.0E-05	1.5E-10	2.9E-13	0.012	0.11	34,594	0.057	1.1E-13	42,727
						Repl	6.2E-08	2.4E-13	1.4E-18	0.009	0.22	41,110	0.062	2.8E-19	54,542
						Comb	7.1E-12	1.6E-22	4.1E-31	0.010	0.05	75,704	0.060	2.4E-31	97,269
*ITPR2**	chr12:26354349	Overall2Diff	CED	C	0.25	Disc	9.4E-04	8.3E-08	3.1E-09	0.011	0.14	34,601	0.046	1.2E-09	42,735
						Repl	0.08	2.2E-06	3.0E-06	0.015	0.07	40,916	0.034	2.3E-06	54,373
						Comb	3.2E-04	1.2E-12	2.4E-14	0.013	0.02	75,517	0.040	2.9E-14	97,108
*HOXC13**	chr12:52628951	Overall2Diff	CED	A	0.24	Disc	6.3E-04	1.8E-07	4.6E-09	0.011	0.18	34,599	0.048	1.4E-09	42,735
						Repl	0.09	1.4E-06	2.1E-06	0.016	0.05	41,112	0.035	2.0E-06	54,536
						Comb	2.9E-04	1.2E-12	2.1E-14	0.013	0.02	75,711	0.041	3.0E-14	97,271

[a] Nearest gene; [b] Effect allele: WHR_{adjBMI} increasing allele in women; [c] Effect allele frequency from the sex-combined sample; [d] r^2 between *SLC30A10* lead SNPs is 0.96; [e] r^2 between *COBLL1* lead SNPs is 0.69; [f] r^2 between *ADAMTS9* lead SNPs is 0.74. * Novel locus compared to Randall and colleagues.

4.2.1.3 Comparison between the 1- and 2-stage approaches

As expected, the 1-stage approaches outperformed the 2-stage approaches (10 versus seven identified loci).

Consistent with the expectation that the power gain of the 1-stage approach $[\mathrm{Overall}_{1e\text{-}5}{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ for SSE loci outweighs the small power loss for CED loci compared to the 2-stage approach $[[\mathrm{Overall}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$, three SSE loci were detected by the 1-stage approach $[\mathrm{Overall}_{1e\text{-}5}{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ (and missed by any 2-stage approach), but only one CED locus (near *ITPR2)* was detected by the 2-stage approach $[[\mathrm{Overall}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ (and missed by any 1-stage approach).

The general recommendation for adopting a 1-stage design is further supported by the fact that the only identified OED locus (near *LRRC69*) was completely missed by the 2-stage approaches.

4.2.1.4 Comparison with re-analyzed GIANT GENDER project approaches

In order to improve comparability between the here applied best 1- and 2-stage approaches (focused on WHR_{adjBMI} and employed Bonferroni-corrected α-levels) and the originally applied approaches by Randall and colleagues (considered nine anthropometric and obesity traits in parallel and employed a FDR approach to correct for the multiple testing), Randall-like approaches were applied to the GIANT GENDER data set for WHR_{adjBMI}. More specifically, the replication-based 2-stage approaches $[\mathrm{Strat}_{1e\text{-}5},\mathrm{Diff}_{0.05}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Diff}_{1e\text{-}5}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ were applied because these are similar to the originally employed screening approaches (Randall et al., 2013).

In total, the two Randall-like approaches identified four independent loci with significant sex-difference (**Table 11**). All four loci were similarly identified by the best 1- and 2-stage approaches, which however, additionally identified another six and three loci, respectively. More specifically, the Randall-like approach $[\mathrm{Strat}_{1e\text{-}5},\mathrm{Diff}_{0.05}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ identified the exact same set of loci as approach $[\mathrm{Joint}_{1e\text{-}5},\mathrm{Diff}_{0.05}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$, but missed the three loci that were only identified by the combined 2-stage approach $[[\mathrm{Overall}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$. This is consistent with the power analyses that demonstrated equal power of approaches $[\mathrm{Strat}_{1e\text{-}5},\mathrm{Diff}_{0.05}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Joint}_{1e\text{-}5},\mathrm{Diff}_{0.05}]{\rightarrow}[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ for SSE and CED effects, but increased power of approach $[[\mathrm{Overall}_{1e\text{-}5}]{\rightarrow}\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ for CED effects.

Table 11. *Four sexually dimorphic loci for WHR_{adjBMI} detected by the re-analyzed Randall-like 2-stage approaches $[Strat_{1e\text{-}5}, Diff_{0.05}] \rightarrow [Diff_{\alpha_{Bonf}}]$ (Strat2Diff) and $[Diff_{1e\text{-}5}] \rightarrow [Diff_{\alpha_{Bonf}}]$ (Diff2Diff). Independent loci are distinguished by alternating shading. The table is sorted by chromosome and position.*

										MEN			WOMEN	
Gene[a]	**SNP**	**Approach**	**Type**	**EA[b]**	**EAF[c]**	**Stage**	**$P_{Sexdiff}$**	**$P_{Sexstrat}$**	**Beta**	**P**	**N**	**Beta**	**P**	**N**
SLC30A10	chr1:217815441	Strat2Diff	SSE	C	0.72	Disc	1.2E-07	1.7E-17	0.009	0.23	34,599	0.064	8.5E-18	42,732
		Diff2Diff				Repl	1.1E-09	2.8E-19	-2.0E-04	0.98	40,935	0.064	1.4E-19	53,799
COBLL1	chr2:165265564	Strat2Diff	SSE	C	0.62	Disc	1.5E-07	1.8E-13	0.001	0.89	34,572	0.052	9.1E-14	42,712
		Diff2Diff				Repl	2.3E-10	9.8E-15	-0.011	0.15	40,998	0.052	4.9E-15	54,425
ADAMTS9	chr3:64686944	Strat2Diff	SSE	C	0.75	Disc	3.3E-04	7.6E-08	0.004	0.62	34,589	0.042	3.8E-08	42,708
						Repl	2.1E-03	2.1E-12	0.019	0.03	40,997	0.052	1.1E-12	54,424
VEGFA	chr6:43872529	Strat2Diff	SSE	T	0.47	Disc	3.0E-05	2.2E-13	0.012	0.11	34,594	0.057	1.1E-13	42,727
						Repl	6.2E-08	5.5E-19	0.009	0.22	41,110	0.062	2.8E-19	54,542

[a] Nearest gene; [b] Effect allele: WHR_{adjBMI} increasing allele in women; [c] Effect allele frequency from the sex-combined sample

4.2.2 Identification of differences between smokers and non-smokers in genetic effects for WHR_{adjBMI}

In order to screen for genetic effects for WHR_{adjBMI} that are significantly different between smokers and non-smokers, the best 1-stage approaches $[\mathrm{Diff}_{5e\text{-}8}]$ and $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ as well as the best 2-stage approaches $[\mathrm{Joint}_{1e\text{-}5}, \mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[[\mathrm{Overall}_{1e\text{-}5}] \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ were applied to the unbalanced GIANT SMOKING data set for WHR_{adjBMI}, and results were compared.

In total, three loci were identified that display significant differences in genetic effects between non-smokers and smokers (near *SOX11*, *MYT1L* and *EYA4*) (**Figure 19**). All of the three loci were identified by the 1-stage approach $[\mathrm{Overall}_{1e\text{-}5} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$. The 2-stage approach $[[\mathrm{Overall}_{1e\text{-}5}] \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ missed the locus near *EYA4.* Approaches $[\mathrm{Diff}_{5e\text{-}8}]$ and $[\mathrm{Joint}_{1e\text{-}5}, \mathrm{Diff}_{0.05}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ did not identify any loci.

All of the identified loci are categorized as non-smoker-specific SSE loci displaying an effect in non-smokers while lacking effect in smokers. This is consistent with the power comparisons that suggested that power to detect SSE effects carrying the effect in the larger subgroup (non-smokers, $f < 1$) is larger than power to detect SSE effects carrying the effect in the smaller subgroup (smokers, $f > 1$). Clearly, larger sample sizes – in particular more smokers - are needed to increase power to detect subtle smoker-specific SSE effects.

The identification of SSE signals by the approaches that include a-priori filtering on overall association further supports the power computations results.

Same as for the smoker-specific SSE effects, larger sample sizes are needed to identify realistic CED effects. However, given the large power to identify OED effects (power was shown to be comparable to the power of SSE $f < 1$ signals), the lack of $OED_{25\%}$ loci suggests that such loci may simply not exist for the WHR_{adjBMI}.

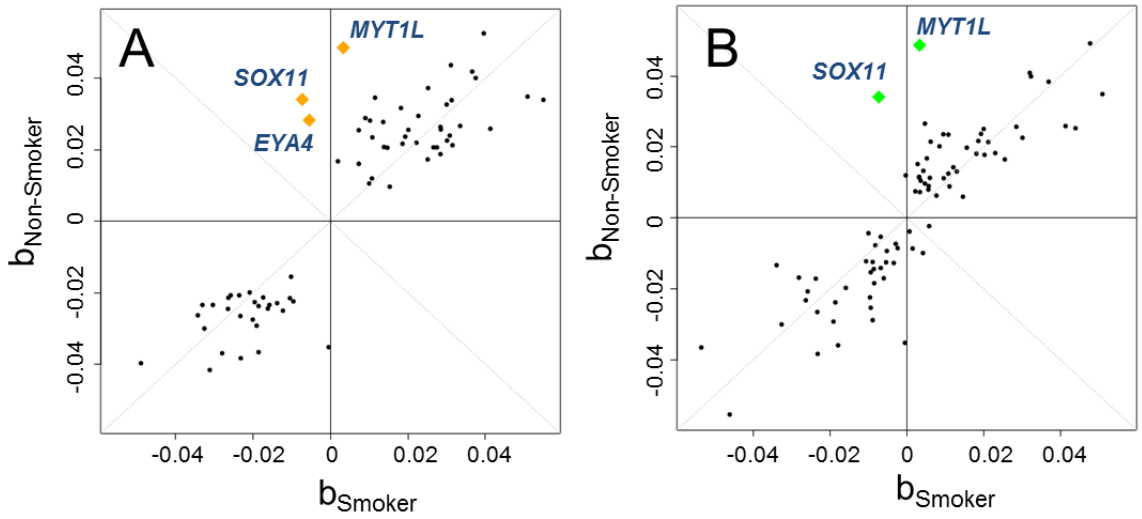


Figure 19. Betas in non-smokers compared to betas in smokers. *The plots contrast betas (pooled genetic effects) between non-smokers (y-axes) and smokers (x-axes). A: 1-stage approaches: Shown are independent lead SNPs for loci that survived the 1-stage filtering criterion $P_{Overall} < 1 \times 10^{-5}$. The orange highlighted loci displayed significant difference between smokers and non-smokers and were identified by approach $[Overall_{1e\text{-}5} \rightarrow Diff_{\alpha_{Bonf}}]$. B: 2-stage approaches: Shown are discovery + replication data set results for independent lead SNPs for loci that survived the discovery filtering criterion $P_{Overall} < 1 \times 10^{-5}$. The green loci displayed significant difference between smokers and non-smokers and were identified by approach $[[Overall_{1e\text{-}5}] \rightarrow Diff_{\alpha_{Bonf}}]$.*

4.2.3 Identification of G x AGE x SEX interaction for WHR_{adjBMI}

In order to screen for 3-way G x AGE x SEX interaction effects for WHR_{adjBMI}, the best approaches $[\mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ and $[\mathrm{MarJoint}_{1e\text{-}5} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ were applied to the GIANT AGE x SEX data set for WHR_{adjBMI}, and results were compared.

No significant 3-way G x AGE x SEX interaction effects were identified (**Figure 20**). Given the large power to find extreme 3-way interaction effects involving medium genetic effect sizes (comparable to the known genetic effect on WHR_{adjBMI} near *PPARG*, power > 80%) suggests that such effects may simply not exist and supports the prior believe that such extreme interaction effects are implausible.

However, lack of biologically plausible 3-way interaction effects may be due to low power. With the current GIANT AGE x SEX setting, only 1-stratum and 3-strata interaction effects could have been identified that involve large genetic effect sizes. Thus, larger sample sizes would be needed to efficiently identify plausible 3-way interaction effects that involve medium or even smaller effect sizes. For example, identification of a medium 1-stratum interaction effect (comparable to the known *PPARG* effect that is present in one stratum and lacking in the other three) with 80% power would require a total sample size of 536,000; and

identification of a medium 3-strata interaction effect (comparable to the known *PPARG* effect that is present in three and lacking in one stratum) with 80% power would require a total sample size of 528,000.

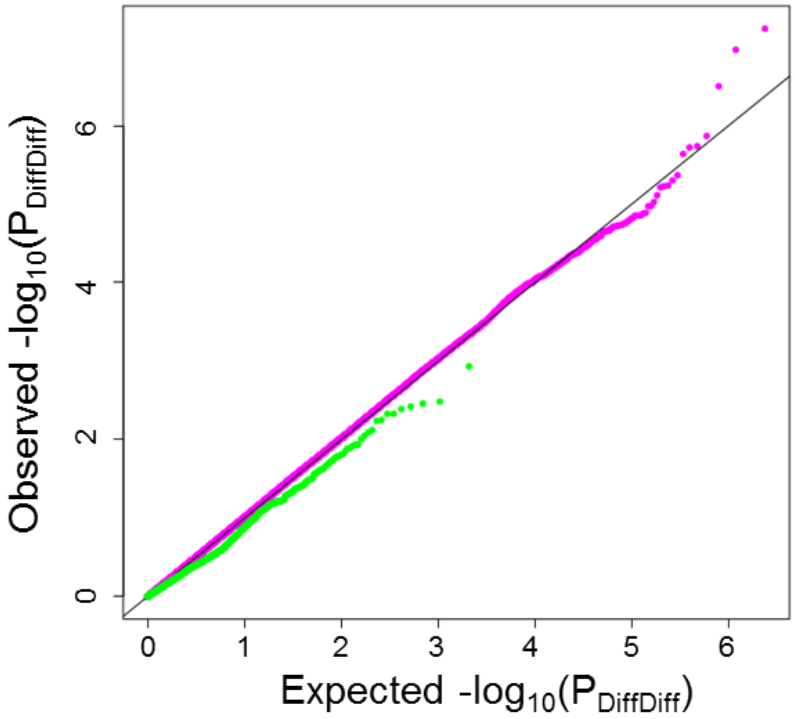


Figure 20. *QQ plot depicting distribution of difference-of-difference P-Values ($P_{DiffDiff}$) for all SNPs genome-wide (magenta) and for a subset of SNPs that passed filtering on marginal joint association ($P_{MarJoint} < 1 \times 10^{-5}$) (green). No significant difference-of-difference P-Values were observed.*

5 The Easy R packages

Large-scale GWAMA consortia nowadays gather data from hundreds of single GWAS studies. Often, consortia are evaluating a multitude of phenotypes in parallel and specific requirements, such as stratified analyses, further increase the overall complexity.

For example, recent stratified GWAMAs from the GIANT consortium involved multiple obesity traits (e.g., BMI, waist circumference and waist-hip ratio), multiple stratification variables (e.g., smoking, physical activity, age and sex) and included up to 320,000 individuals from hundreds of single GWA studies (Justice et al, in progress; Winkler et al, in revision). In total, more than 3,000 single GWAS result files were evaluated, each of which contained ~2.8M rows of SNP-specific association results. Proper conduct of the stratified GWAMAs necessitated an extensive quality control (QC) procedure and software that was able to deal with the large number of large GWAS result files.

In addition, a further software tool was required to evaluate stratified GWAMA results, i.e., to apply the here defined screening approaches, to extract significant results and to generate graphical summaries.

Two R packages were developed that facilitate conduct, QC and evaluation of stratified GWAMAs (**Figure 21**). The R package *EasyQC* was originally developed for QC of large numbers of general large scale GWAS and GWAMA data sets, but can readily be applied to stratified GWAS and stratified GWAMA data sets (Winkler, Day, et al., 2014). The R package *EasyStrata* is tailored for the statistical and graphical evaluation of stratified GWAMA results data (Winkler, Kutalik, et al., 2014).

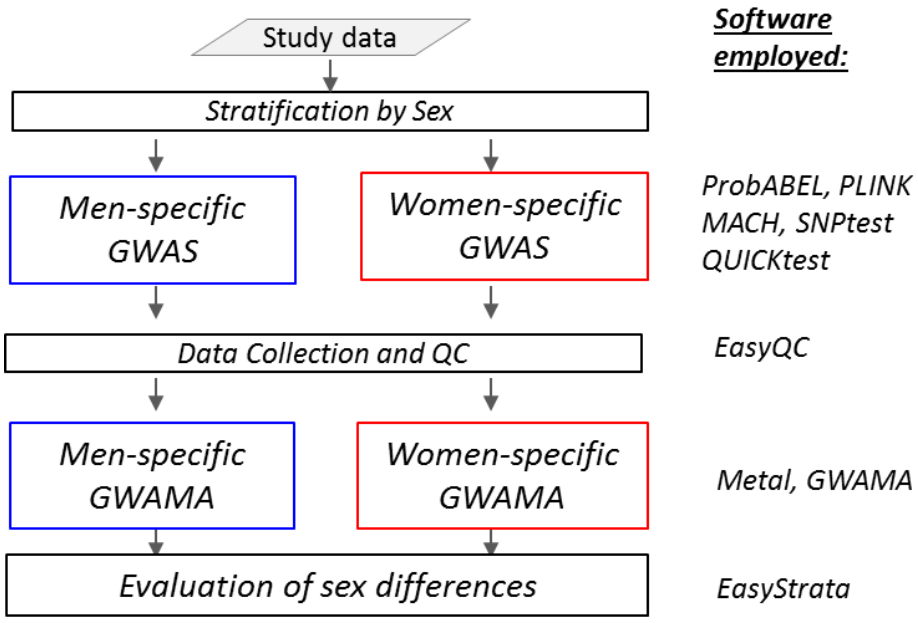


Figure 21. *Integration of EasyQC and EasyStrata into a sex-stratified GWAMA pipeline. EasyQC is applied for the QC of the study-specific GWAS results and EasyStrata is applied to evaluate the sex-stratified GWAMA results, e.g., to apply the screening approach, to extract independent loci with significant sex-difference effects or to create graphical presentations of results.*

5.1 The Easy framework

Both packages were implemented into the *Easy* framework. The framework only exports a single function, either *EasyQC()* or *EasyStrata()*. To start the program, the function is called from the R command line with a single parameter that is the path to a customized ecf-file, e.g., *EasyStrata("/path2ecf/pipeline.ecf")*. The ecf-file is a plain text file that consists of two parts: a configuration-section at the beginning that defines data input and output, and a scripting interface that defines the commands to be executed. The scripting interface allows aligning multiple statistical, graphical and general data handling functions in a flexible fashion. The ecf-file concept and the *Easy* framework simplify data-handling and allow the user to easily develop customized QC or evaluation pipelines. Customized ecf-pipelines to facilitate GWAMA QC and evaluation of stratified GWAMA results were developed and can be downloaded from the website www.genepi-regensburg.de/software. An example *EasyStrata* ecf-pipeline to evaluate sex-stratified GWAMA results is shown in **Appendix 9.7.**

5.1.1 Object oriented programming

In order to use a structured object oriented programming technique, each of the exported *EasyQC* or *EasyStrata* commands was integrated as separate class. Class definitions were

similarly structured and provide variables and functions specifically for the respective command. Each command class contains functions to (i) to translate the scripting interface parameters into class variables; (ii) conduct validity checks, e.g., to ensure that the stated parameters are applicable with the command and with the current input data; and (iii) to execute the actual command-specific algorithm. This object-oriented programming technique allows the developer to easily update the package by new functionality or commands that are automatically exported to the scripting interface.

5.1.2 Big data

To efficiently handle large data-sets, the *Easy* framework utilizes the *data.table* package that provides state-of-the-art functions to handle big data in R. Objects of type *data.table* are an extension to the commonly used *data.frames,* but provide a much faster and more efficient aggregation and combination of data. The *data.table* concept was implemented in a specific data type called *GWADATA* that was developed to supply general data handling and data manipulation functions and that is utilized throughout the whole package implementation.

5.1.3 Requirements and performance

Since the *Easy* framework inherits functionality from the graphical R packages *Cairo*, *plotrix* and *data.table*, these packages have to be available prior to installation of *EasyQC* or *EasyStrata*. Extracting independent loci using LD-based thresholds requires the software *PLINK* to be installed (Purcell et al., 2007).

For GWAS or stratified GWAMA results based on HapMap imputed data (~2.8M SNPs) (International HapMap et al., 2010), it is recommended to have at least 4GB of Random-Access Memory (RAM) available (see **Table 12** for an evaluation of runtime and memory requirements for a standard evaluation pipeline). While the package was primarily developed for HapMap imputed data sets, it is also applicable to data sets that are based on 1000 Genomes imputation (~40M SNPs), which are increasingly being implemented in meta-analysis projects (Genomes Project et al., 2010). To execute standard evaluation pipelines with 1000 Genomes imputed data sets, at least ~20GB of RAM have to be available (**Table 12**).

Table 12. *Runtime and memory requirements for a standard EasyStrata evaluation. The table is based on Supplementary Table 3 from (Winkler, Kutalik, et al., 2014).*

GWAMA data set	***#SNPs***	***Maximum RAM allocated****	***Computational Time****
Typed MetaboChip	185,648	256 MB	10 s
HapMap Imputed	2,515,652	3.0 GB	3 min 4 s
1000 Genomes Imputed	19,395,227	19.8 GB	24 min 39 s

* The presented values represent average values, gathered from 10 identical *EasyStrata* runs using the pipeline shown in **Appendix 9.7**, R version 2.15.3 on a Linux Server with a 3.07GHz Intel Xeon CPU and 96GB RAM.

5.2 EasyQC

GWAMA consortium analysts often have to deal with thousands of individual big GWAS result data sets, a circumstance that necessitates extensive quality control (QC) and requires software that is able to handle such large numbers of big GWAS result data sets.

To accomplish this, a powerful tool called *EasyQC* was developed that facilitates QC of multiple study-specific GWAS results as well as of meta-analyzed GWAMA results. A typical workflow of GWAMA conduct and QC was developed and is illustrated in **Figure 22**.

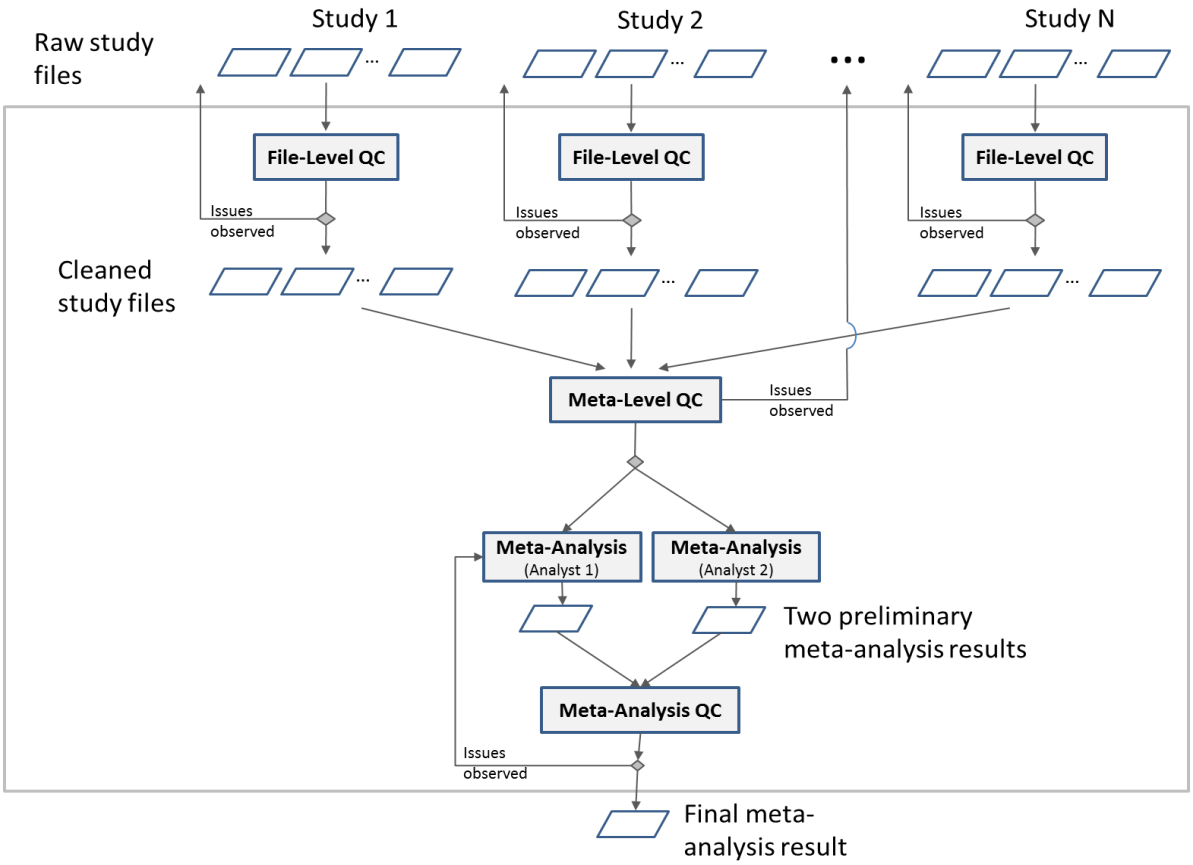


Figure 22. *Workflow of conduct and QC of a typical GWAMA. The figure is based on Figure 1 from Winkler et al (Winkler, Day, et al., 2014).*

5. The Easy R packages

Assume study-specific GWAS results have been collected centrally by a team of consortium analysts. The typical GWAMA conduct and QC can be divided into four major steps that are outlined consecutively: File-level QC, Meta-level QC, Meta-analysis and Meta-analysis QC.

At *file-level QC*, each study-specific GWAS result file is *cleaned* and *checked*. File cleaning involves multiple steps such as harmonization of column names, file formats, or marker names or exclusion of low quality SNPs: For example, in GIANT consortium GWAMAs, SNPs were excluded due to low minor allele count (mac ≤ 6, mac defined as 2*MAF*N), low imputation quality (< 0.3, <0.4 or <0.8, respectively for MACH, IMPUTE or PLINK imputed data sets) or due to low genotype quality (call-rate < 0.95 or out of Hardy-Weinberg equilibrium: $P_{HWE} < 10^{-5}$). File checking involves sanity checks, such as checking overall descriptive statistics of the results or the number of SNP exclusions. The task is typically outlined for each study separately and aims at identification of study-specific issues and at exclusion of poor quality data.

At *meta-level QC*, summary statistics are compared across studies. The aim of this step is to identify study-specific issues that have yet been undetected by file-level QC. Meta-level QC involves checking for a range of analytical issues, such as issues with trait transformations (checked by creating so-called SE-N plots that contrast median standard errors with sample sizes across studies), issues with allele frequencies or strand (checked by plotting the study allele frequencies against reference frequencies), or issues with population stratification (checked by creating QQ plots and by calculating genomic-control (GC) inflation factors; note that the actual GC correction is applied during meta-analysis (Devlin & Roeder, 1999)).

Meta-analysis denotes the stage at which the actual meta-analysis is conducted. To limit potential errors with regards to study in- or exclusion or other failures in the meta-analysis scripts, this task is typically outlined by two analysts independently and in parallel. In order to correct for potential population stratification issues, a GC correction is applied to each single GWAS result during meta-analysis (this can be accomplished by the *metal* function GENOMICCONTROL ON).

Finally, *meta-analysis QC* stands for checking the meta-analysis results. This involves the comparison of results between the two meta-analysis performed by the different analysts; and the QC of the meta-analysis result itself. For example, meta-analysis result QC involves checking for issues with between-study heterogeneity or checking the meta-analysis GC inflation factor.

Customized *EasyQC* pipelines have been developed to accomplish file-level, meta-level and meta-analysis QC and can be downloaded from www.genepi-regensburg.de/software.

5.3 EasyStrata

After the conduct of the actual stratified GWAMA itself, analysts still have to deal with big GWAMA data sets that contain pooled association results for millions of SNPs and for multiple strata. Application of the here discussed statistical tests or approaches, and extraction of significant results, requires customized software.

To facilitate this, *EasyStrata* was developed as a pipelining tool that combines state-of-the-art statistical and graphical methods to evaluate large-scale stratified GWAMA results data.

The prerequisite of an *EasyStrata* evaluation is that the stratified GWAMAs itself have already been conducted by meta-analytical software programs such as metal (Willer et al., 2010) or GWAMA (Magi & Morris, 2010) (**Figure 21**). Although *EasyStrata* was primarily developed for evaluation of stratified GWAMAs based on continuous trait GWAS that involve linear regression models on the study-level, it also provides an extended applicability for other GWAMA configurations (see **Appendix 9.8**).

A multitude of *EasyStrata* evaluation pipelines for stratified GWAMA results, e.g. to graphically display and to extract independent significant sexually dimorphic loci from sex-stratified GWAMA outcomes, have been developed and can be accessed from www.genepi-regensburg.de/software. An example *EasyStrata* ecf-pipeline to evaluate sex-stratified GWAMA results is shown in **Appendix 9.7.**

5.3.1 Statistical Functionality

EasyStrata has implemented handling of (i) stratum-specific inverse-variance weighted GWAMA outcomes, i.e., for each SNP the stratum-specific pooled genetic estimate with standard error (Cox & Hinkley, 1979), and (ii) stratum-specific sample size weighted Z-score based GWAMA outcomes, i.e., for each SNP the stratum-specific pooled Z-score and accumulated sample size (Stouffer, 1949). A summary of implemented statistics tailored for evaluation of stratified GWAMA outcomes is given in **Table 13**. Other statistical methods that are capable to general GWAMA settings, such as multiple testing or genomic control correction, are described in **Appendix 9.8.**

Table 13. *EasyStrata functions for the statistical analysis of stratified GWAMA results. The table is based on Supplementary Table 2 from Winkler et al (Winkler, Kutalik, et al., 2014). The given probability distributions assume that the null hypothesis of the respective test is true.*

Function	**#Strata**	**Test-Statistics**		**References**
		Inverse-variance weighted GWAMA; $w_i = 1/SE_i^2$	*Sample size weighted z-score based GWAMA;* $w_i = \sqrt{N_i}$*	
METAANALYSIS	*i=1…m*	$\beta_{Overall} = \frac{\sum_i \beta_i w_i}{\sum_i w_i}$; $SE_{Overall} = \sqrt{\frac{1}{\sum_i w_i}}$ $Z_{Overall} = \frac{\beta_{Overall}}{SE_{Overall}} \sim N(0,1)$	$N_{Overall} = \sum_i N_i$, $Z_{Overall} = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}} \sim N(0,1)$	(Cox & Hinkley, 1979; Stouffer, 1949)
JOINTTEST	*i=1…m*	$C_{Joint} = \sum_i (\beta_i^2 w_i) \sim \chi^2_m$	$C_{Joint} = \sum_i Z_i^2 \sim \chi^2_m$	(Aschard et al., 2010)
CALCPDIFF	*2*	$Z_{Diff} = \frac{\beta_1 - \beta_2}{\sqrt{SE_1^2 + SE_2^2}} \sim N(0,1)$ or $Z_{Diff} = \frac{\beta_1 - \beta_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \sim N(0,1)$ with r = correlation between β_1 and β_2 §	$Z_{Diff} = \frac{Z_1/\sqrt{N_1} - Z_2/\sqrt{N_2}}{\sqrt{1/N_1 + 1/N_2}} \sim N(0,1)$ or $Z_{Diff} = \frac{Z_1/\sqrt{N_1} - Z_2/\sqrt{N_2}}{\sqrt{1/N_1 + 1/N_2 - 2r/\sqrt{N_1N_2}}} \sim N(0,1)$ with r = correlation between $\frac{Z_1}{\sqrt{N_1}}$ and $\frac{Z_2}{\sqrt{N_2}}$ §	(Magi, Lindgren, & Morris, 2010; Randall et al., 2013)
CALCPDIFFDIFF	*2x2*	$Z_{DiffDiff} = \frac{(\beta_{11} - \beta_{12}) - (\beta_{21} - \beta_{22})}{\sqrt{SE_{11}^2 + SE_{12}^2 + SE_{21}^2 + SE_{22}^2}} \sim N(0,1)$	$Z_{Diff} = \frac{(Z_{11}/\sqrt{N_{11}} - Z_{12}/\sqrt{N_{12}}) - (Z_{21}/\sqrt{N_{21}} - Z_{22}/\sqrt{N_{22}})}{\sqrt{1/N_{11} + 1/N_{12} + 1/N_{21} + 1/N_{22}}} \sim N(0,1)$	(Winkler et al, in revision)
CALCPHET	*i=1…m*	$C_{Het} = \sum_i [(\beta_i - \beta_{Overall})^2 w_i] \sim \chi^2_{m-1}$	$C_{Het} = \sum_i [(Z_i/\sqrt{N_i} - Z_{Overall}/\sqrt{N_{Overall}})^2 {w_i}^2] \sim \chi^2_{m-1}$	(Cochran, 1954)

*for dichotomous traits, with unequal numbers of controls and cases, we suggest using the total effective sample size $N_{eff} = 4/(1/N_{cases} + 1/N_{controls})$, which can be calculated in EasyStrata by the ADDCOL function: "ADDCOL --rcdAddCol 4/(1/NCASES+ 1/NCONTROLS) --colOut Neff", given columns NCASES and NCONTROLS in the input data set. § calculated as the Spearman rank correlation coefficient across all SNPs

5.3.2 Graphical Functionality

To facilitate extended graphical presentation of stratified GWAMA results data, commonly used graphical functions like Quantile-Quantile-(QQ), scatter- and Manhattan-plots were implemented.

A particular functionality of *EasyStrata* is tailored for comparison of stratified GWAMA results within single plots. Standard Manhattan plots display SNP-specific association P-Values over respective chromosomal base-positions. *EasyStrata* introduces so-called Miami plots that contrast two Manhattan plots in a single graph (**Figure 23**). This is particularly useful for a locus-wise comparison of two stratified GWAMA results.

A general comparison of the association strength between multiple strata can be obtained by the *EasyStrata* functionality to draw multiple QQ plot curves into a single graph.

Scatterplots are one of the most simple and practical tools to investigate the relationship between two variables. *EasyStrata* can create scatterplots comparing two types of statistics and allows adding further dimensions by applying user-defined colors, symbols and symbol-sizes. For example, this can be helpful when comparing effect sizes from two strata.

EasyStrata provides many other helpful extended graphical features, such as locus highlighting, exclusion of less significant SNPs to increase plotting speed or breaking up axes scales to properly display highly significant results (see **Appendix 9.8**). These features may also be helpful for the interpretation of general GWAMA data.

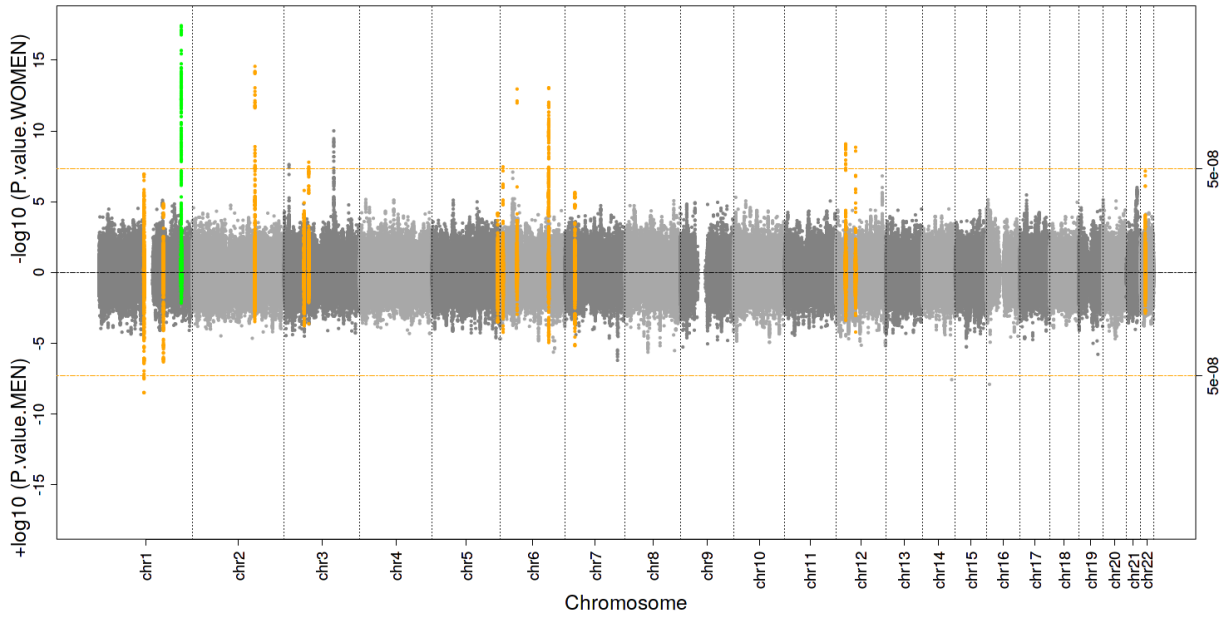


Figure 23. *Example Miami-plot contrasting women- and men-specific GWAMA results for WHR_{adjBMI}. The plot shows discovery stage results from the GIANT GENDER project (Randall et al., 2013). The coloring highlights WHR_{adjBMI} loci that were previously established for overall association by Lindgren and colleagues (green)(Lindgren et al., 2009) and by Heid and colleagues (orange)(Heid et al., 2010). Since identical samples have been utilized in the sex-combined screen conducted by Heid and colleagues as compared to the sex-difference screen conducted by Randall and colleagues, the plot nicely shows that two additional genome-wide significant loci on chromosome 3 could have been identified by Heid and colleagues, if an additional sex-stratified screen would have been conducted.*

6 Discussion

6.1 Summary of main results

In the following chapters a summary of results is presented for each of the initial objectives. Based on a systematic methodological evaluation, optimized stratified GWAMA methods to screen for stratum-difference in genetic effects between two strata were developed (summary in chapter 6.1.1), considerations were extended to stratified GWAMA methods to screen for 3-way G x AGE x SEX interaction effects (summary in chapter 6.1.2), recommended methods were applied to real GIANT consortium stratified GWAMA data sets for WHR_{adjBMI} (summary in chapter 6.1.3) and software tools were developed that facilitate conduct and evaluation of large-scale stratified GWAMAs (summary in chapter 6.1.4).

6.1.1 Optimal stratified GWAMA approaches to screen for stratum-difference

First, several stratified GWAMA approaches to identify difference in the genetic effects between two strata were defined. A systematic scheme of approaches was developed that incorporated different statistical filtering methods.

Simulation-based evaluations of type 1 error rates as well as analytical comparisons of power were performed and optimal approaches were recommended for specific types of stratum-difference and for various stage designs - including 1-stage and 2-stage designs as well as balanced strata (equal stratum sizes) and unbalanced strata (unequal stratum sizes) designs (**Table 14**).

Table 14. *Optimal stratified GWAMA approaches to identify stratum-difference.*

Type of stratum-difference[a]	**Stage Design**	
	1-stage	2-stage
Opposite effect direction (OED)	$[\mathrm{Diff}_{5e\text{-}8}]$	$[\mathrm{Joint}_{1e\text{-}5},\mathrm{Diff}_{0.05}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$
Single-stratum effect (SSE)	$[\mathrm{Overall}_{1e\text{-}5}\rightarrow\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ *for f > 1.5*[b]*:* $[\mathrm{Diff}_{5e\text{-}8}]$	$[\mathrm{Joint}_{1e\text{-}5},\mathrm{Diff}_{0.05}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ *for f < 0.66*[b]*:* $[[\mathrm{Overall}_{1e\text{-}5}]\rightarrow\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$
Concordant effect direction (CED)	$[\mathrm{Overall}_{5e\text{-}8}\rightarrow\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$	$[[\mathrm{Overall}_{1e\text{-}5}]\rightarrow\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$

[a] assuming a given effect in stratum 1; [b] proportion of stratum 2 sample size to stratum 1 sample size

For stratified GWAMAs with a balanced 1-stage design, (i) the intuitive approach of testing for difference at a genome-wide significant level is optimal for OED effects, and (ii) testing for difference with a priori filtering for overall association is optimal for SSE and CED effects (**Table 14**).

For example, assuming a total sample size of 200,000, the power to identify the difference between a medium-sized genetic effect on WHR_{adjBMI} in women and no effect in men was 80.7% and 47.4%, with and without filtering for overall association, respectively. Interestingly, filtering for weaker overall association ($P_{Overall} < 1 \times 10^{-5}$) was beneficial for SSE effects, while a more stringent threshold ($P_{Overall} < 5 \times 10^{-8}$) was beneficial for CED effects. However, because the power gain of the weaker control for SSE effects was larger than the power loss for CED effects, generally a weaker control of the overall association is recommended.

Remarkably, simulations demonstrated inflated type 1 error for the difference test for SNPs that were filtered for stratified or joint association. As a consequence, filtering for stratified or joint association necessitated an independent data set to perform the difference test and can only be implemented in a 2-stage stratified GWAMA design.

For stratified GWAMAs with a balanced 2-stage design, (i) testing for difference using the replication stage data while focusing on SNPs that were filtered for joint association as well as for nominally significant difference in the discovery stage data is optimal for OED and SSE effects, and (ii) testing for difference using the combined (discovery + replication) stage data while focusing on SNPs that were filtered for overall association in the discovery stage data is optimal for CED effects (**Table 14**).

When comparing stage designs, the 1-stage approaches generally displayed larger power to identify stratum-difference than the 2-stage approaches.

For unbalanced strata designs, the optimal approach did only change for SSE effects (**Table 14**). Interestingly, the power comparisons also highlighted that it is generally more likely to identify stratum-difference that exhibits a stronger effect in the larger stratum. For example, for an analysis that contains more non-smokers than smokers, the power to detect stratum-difference with effect in non-smokers (and a less pronounced or no effect in smokers) is higher than the power to detect stratum-difference with effect in smokers (and a less pronounced or no effect in non-smokers).

6.1.2 Optimal stratified GWAMA approaches to screen for G x AGE x SEX interaction

Methodological considerations were extended to age- and sex-stratified GWAMAs that are based on two dichotomous stratification variables (SEX: men vs women; AGE: younger vs older than 50 years of age, which reflects mean age of menopause in women). A difference-

of-difference test was introduced that can be used to test for 3-way G x AGE x SEX interaction effects given an age- and sex-stratified GWAMA model. Several approaches to identify 3-way G x AGE x SEX interaction effects under a balanced 1-stage design were defined and a systematic scheme of approaches was developed.

As for the 2-strata configuration, simulations demonstrated inflated type 1 error rates for the difference-of-difference test for SNPs that were filtered for stratified or joint association, but valid type 1 error rates for SNPs that were filtered for overall (strata-combined) association. Surprisingly, filtering data for marginal tests (using age-marginal and sex-marginal effects, obtained from meta-analysis of respective strata), also displayed valid type 1 error rates for the difference-of-difference test, when this is calculated for the filtered SNPs and within the filtering data set.

Analytical power formulae were derived and power of the validated age- and sex-stratified GWAMA approaches was compared for varying realistic scenarios. Power computations demonstrated that testing for difference-of-difference at a genome-wide significance level is optimal for extreme 3-way interaction effects. Surprisingly, filtering for marginal joint association and testing the filtered SNPs for difference-of-difference resulted in highest power for biologically plausible 3-way interaction effects that imply an effect in one stratum (e.g., younger women) and no effect in the other three, or an identical effect in three strata and no effect in one stratum.

6.1.3 Application to stratified GWAMAs for obesity traits

The optimal approaches were applied to real GIANT consortium sex-stratified, smoking-status stratified and age- and sex-stratified GWAMA results for WHR_{adjBMI}.

Ten loci were identified that displayed significant sex-differences in genetic effects on WHR_{adjBMI}. Nine of the ten loci showed a significant effect on WHR_{adjBMI} in women and no or less pronounced (but concordant) effect in men, and one locus displayed an opposite effect on WHR_{adjBMI} between men and women.

Three loci showed significantly stronger genetic effects on WHR_{adjBMI} in non-smokers as compared to smokers. Given the larger sample size in non-smokers, this was consistent with the power computation result that concluded that it is more likely to identify stratum-difference with larger effect in the larger stratum.

No loci displayed significant 3-way G x AGE x SEX interaction effects, which may however be attributable to low power of the study to identify biological plausible 1-stratum or 3-strata interaction effects involving medium genetic effects.

Generally, the application of approaches underscored the power computation results and demonstrated that significant differences in genetic effects for WHR_{adjBMI} exist between men and women, as well as between smokers and non-smokers.

6.1.4 Software for conduct and evaluation of stratified GWAMAs

Finally, two software packages, *EasyQC* and *EasyStrata*, were developed that facilitate (i) conduct and quality control of stratified GWAMAs, and (ii) statistical evaluation and graphical presentation of stratified GWAMA results as well as application of the here defined statistical screening approaches to identify significant stratum-differences. Both packages can easily be integrated into the workflow of large scale stratified GWAMA projects and are increasingly being utilized by various GWAMA consortia.

6.2 Stratified GWAMA methods to tackle gene-environment interaction effects

The here discussed stratified GWAMA approaches aim at the identification of stratum-difference and were specifically evaluated for continuous outcomes Y, additively modelled genotypes and for the meta-analytical framework. Testing for stratum-difference under a stratified model is mathematically equivalent to testing gene-strata (G x S) interaction effects under an interaction model (under the assumption of unrelated subjects and no covariate interacting with G or S). G x S interaction effects involve a dichotomous environmental (stratification) variable S and – as such - are a specific form of general gene-environment (G x E) interaction effects (that may also involve continuous or categorical E).

In the literature, reported statistical methods can be distinguished into those that (i) aim at identifying SNP effects while accounting for potential G x E interaction, or (ii) aim at identifying G x E interaction as such (Hutter et al., 2013). In the following, these methods and their applicability to the stratified GWAMA setting (involving continuous Y and dichotomous S) are discussed (chapter 6.2.1 and chapter 6.2.2). Moreover, advantages and disadvantages of stratified compared to interaction modelling are discussed (chapter 6.2.3).

6.2.1 Reported methods to account for G x E interaction effects

Over the past years a number of methods have been developed that aim at the identification of novel SNP effects while accounting for potential G x E interaction effects (Aschard et al., 2010; Dai et al., 2012; Kraft et al., 2007; Manning et al., 2011) (and including own work (Behrens et al., 2011). A summary can be found in **Table 15**.

Table 15. *Reported methods for gene-environment interaction effects.*

Reference	Type of test	Outcome Y	Exposure E	Step 1 filter	Applicability to meta-analysis summary statistics shown	Applicability to continuous Y shown
Testing for SNP effects while accounting for GxE interaction effects						
(Kraft et al., 2007)	joint	bin / cont	bin / cont	NA	Yes	Yes
(Aschard et al., 2010)	stratified joint meta-analysis	bin / cont	bin	NA	Yes	Yes
(Manning et al., 2011)	joint meta-analysis	bin / cont	bin / cont	NA	Yes	Yes
(Behrens et al., 2011)	stratified	bin / cont	bin	NA	Yes	Yes
(Dai et al., 2012)	simultaneous marginal + interaction	bin	bin	NA	No	No
Testing for G x E interaction effects						
(Piegorsch et al., 1994)	Case-only	bin	categorical	NA	No	No
(Kooperberg & Leblanc, 2008)	2-step	bin	SNP	$P_{Overall}$	No	Yes
(Mukherjee et al., 2008)	Empirical Bayes	bin	bin	NA	No	No
(D. Li & Conti, 2009)	Bayes model averaging	bin	bin	NA	No	No
(Murcray, Lewinger, & Gauderman, 2009)	2-step	bin	bin	P_{G-E}	No	No
(Murcray, Lewinger, Conti, Thomas, & Gauderman, 2011)	2-step	bin	bin / cont	$P_{Overall}$ + P_{G-E}	No	No
(Hsu et al., 2012)	2-step	bin	bin	$P_{Overall}$ + P_{G-E}	No	No
(Gauderman, Zhang, Morrison, & Lewinger, 2013)	2-step	bin	bin	$P_{Overall}$ + P_{G-E}	No	No
This work	2-step	cont	bin	$P_{Overall}$ / P_{joint} / P_{Strat} / P_{Diff}	Yes	Yes

bin = binary; cont = continuous; 2-step denotes consecutive filtering and interaction tests; NA = not applicable

The most straightforward way to account for potential G x S interaction effects is to conduct stratum-specific analyses and to screen for associated variants within each stratum separately, called the stratified screen in the following (Behrens et al., 2011).

Another method to generally account for potential G x E interaction effects is the joint test screen that simultaneously tests the main and the interaction effect. While the original version of the joint test had solely been developed for single-study G x E interaction models (Kraft et al., 2007), recent versions have integrated the joint test into the interaction GWAMA setting (Manning et al., 2011) as well as into the stratified GWAMA setting (Aschard et al., 2010).

Both, stratified and joint test screens have commonly and successfully been employed to identify associated variants for many different traits (Hamza et al., 2011; Hancock et al., 2012; Manning et al., 2012; Simino, Sung, Kume, Schwander, & Rao, 2013; Taylor et al., 2013; Winham et al., 2014). However, to distinguish whether the highlighted genetic effects are significantly different between strata, or whether the highlighted loci were detected for main or for interaction effects, one would need to additionally test the highlighted variants for stratum-difference or for interaction. Neither the stratified nor the joint test can be used to establish the interaction as such.

In this work, stratified and joint tests have been considered as methods to filter data sets prior to difference testing. Via simulation, it has been demonstrated that SNPs that were filtered for stratified or joint association cannot be tested for difference (or for G x S interaction) within the same data set. Thus, the many GWAMA projects employing stratified or joint test approaches require independent data sets to test the discovered variants for difference or interaction.

6.2.2 Reported methods to detect G x E interaction effects

Similarly, a multitude of methods specifically tailored to identify G x E interaction effects were released over the past years (**Table 15**).

These include case-only approaches (Piegorsch et al., 1994), data adaptive methods (D. Li & Conti, 2009; Mukherjee et al., 2008) as well as hedge methods that involve initial filtering steps on marginal gene-disease association, gene-environmental exposure correlation, or both (Gauderman et al., 2013; Hsu et al., 2012; Kooperberg & Leblanc, 2008; Murcray et al., 2011; Murcray et al., 2009).

All of the reported methods were developed for single-study analyses, for interaction models and for dichotomous disease outcomes. It is not clear at all whether or how the presented methods can be transferred into the meta-analysis framework, into the stratified setting or whether they can be extended to continuous outcomes. While some of the reported methods may indeed be applicable to continuous outcomes (still to be demonstrated), others

were specifically developed for disease outcomes (e.g., those that involve case-only approaches) and are not extendible to continuous outcomes.

Notably, no methods have been reported that consider 3-way G x E_1 x E_2 interaction effects so far.

This work is the first structured methodological evaluation of approaches to identify G x S interaction effects, focusing on continuous outcomes, stratified modelling and on an implementation into the meta-analytical framework. Moreover this is the first work to consider 3-way G x S_1 x S_2 interaction effects at all.

6.2.3 Relating stratified modelling to interaction modelling

The stratified GWAMA approach to identify gene-strata (G x S) interaction (measured as stratum-difference) has some important advantages and disadvantages compared to GWAMAs based on study-specific G x S interaction modeling.

A disadvantage of the here discussed stratified GWAMA approaches to identify G x S effects is that they are limited to dichotomous (environmental) stratification variables S, whereas the interaction modelling as such is extendible to categorical or continuous environmental variables E. However, any continuous E can generally be dichotomized into two strata (by applying a specific threshold that separates E into two groups). It depends on the respective scenario as to whether the continuous or the dichotomized version is to be preferred. For example, women's body shape changes after menopause (due to hormonal changes) and for some G x E interaction effects, fitting the interaction model with dichotomized age (younger vs older than 50 years of age) might be beneficial over fitting continuous age.

For categorical E (>2 categories), the stratified GWAMA per se would be straightforward (one meta-analysis per category). However more refined methods, such as trend tests for ordinal categorical variables, would have to be evaluated using the pooled category-specific summary effect estimates. Such methods are not yet available for genome-wide summary statistics and have not been considered so far, thus being a research topic of future efforts

Another disadvantage of the stratified model is that it cannot model family structures that may be present across strata. For example, a sex-stratified analysis may not properly control for the relatedness across sexes when brothers and sisters would be analyzed separately. For the stratum-difference test, a covariance correction has been introduced that is supposed to control for such relatedness (Randall et al., 2013). However a structured methodological evaluation to investigate this in detail has not yet been performed and similar corrections for other employed tests (e.g., overall or joint tests) are not yet available.

One advantage of the stratified modelling - compared to interaction modelling - is that it provides better control for confounding and can avoid model misspecification. Typically in genetic association studies, additional covariables C, such as AGE, are added to the regression models in order to avoid confounding and to increase power to find the genetic association. Many published G x E interaction studies added co-variables to the interaction model yet missed the modelling of potential gene-covariable (G x C) and environment-covariable (E x C) interaction effects. Missing these interaction terms leads to model misspecification and can create spurious G x E interaction effects (Keller, 2014). The reason for missing these interaction terms is not merely lack of knowledge, but mostly the fact that commonly used software programs (to facilitate genome-wide G x E interaction studies) lack the functionality to properly account for G x C and E x C interaction effects. For example the software program PLINK cannot model E x C terms (Purcell et al., 2007) and ProbABEL cannot account for G x C or E x C effects at all (Aulchenko et al., 2010). In contrast, the stratified model is not affected by such model misspecification issues because it provides proper control of confounding by adding covariables to each of the stratified 'main effect' models separately.

A big advantage of stratified modelling is larger power for a very practical reason: Especially in the context of large scale GWAMA projects that combine data from hundreds of studies, the stratified GWAMA will tend to comprise a larger total sample size – thus resulting in larger power. One simple reason for this is that some studies can be expected to lack individuals for a specific stratum. Women-only studies may contribute to the women-specific meta-analysis, but a G x SEX interaction model cannot be fitted due to the lack of men in the study. For example, the GIANT consortium sex-stratified GWAMA from 2013 involved >7,000 men from five men-only studies and >30,000 women from six women-only studies that would have been missed by the interaction modelling (Randall et al., 2013).

Another reason for the larger sample size is that the stratified model generally reflects a straightforward model that can be implemented on the study-level without requiring specialized GWAS tools - other than those that are commonly applied for main effect GWAS and that are familiar to study analysts. For example in the GIANT analyses to study gene-smoking interaction, the sample size of the smoking-stratified GWAMA was ~10% larger than the sample size of the GWAMA based on G x SMOKING interaction modelling, just because some of the studies were not able to contribute genome-wide interaction model results due to increased computational burdens (Justice et al, in progress).

Finally, a further advantage of the stratified approach with regards to modelling is that it already allows for investigating 3-way G x S_1 x S_2 interaction effects (however limited to dichotomous S_1 and S_2). Prior to this work, there has no GWAS software tool or method been available that allows for fitting a 3-way interaction model on a genome-wide scale.

6.3 Stratum-specific effects in the genetics of obesity

Over the past years, several stratified GWAMAs have highlighted biologically relevant differences in genetic effects on obesity measures between men and women, between smokers and non-smokers or between age-groups.

In the following chapters a chronological review of identified sex-differences in genetic effects for body fat distribution is given and the observed sex-specific (chapter 6.3.1), smoking-status-specific (chapter 6.3.2) and age-specific effects (chapter 6.3.3) in the genetics of obesity are discussed.

6.3.1 Sex-differences in genetic effects for body fat distribution

Over the past years, several GIANT consortium GWAMA projects have increasingly identified sexually dimorphic genetic effects for WHR_{adjBMI} (as a measure of body fat distribution).

In 2009, the very first round of GIANT consortium GWAMAs for WHR_{adjBMI} – including up to 37,670 individuals in the discovery stage - was published (Lindgren et al., 2009). This work primarily aimed at the identification of associated loci. To accomplish this, overall (sex-combined) as well as sex-specific GWAMAs were performed. A single locus (near *LYPLAL1*) was identified to be associated with WHR_{adjBMI} in women only. This was the first locus in the genetics of WHR_{adjBMI} to display significant differences between sexes.

In 2010, a second round of GIANT consortium GWAMAs for WHR_{adjBMI} – now including up to 77,167 individuals in discovery stage - was conducted (Heid et al., 2010). Again, this work primarily aimed at the identification of overall (sex-combined) associated loci. In total, 14 loci displayed genome-wide significant overall effects on WHR_{adjBMI} (including the previously reported *LYPLAL1* locus and 13 novel associated loci). Interestingly, seven of the 14 loci displayed significant sex-differences, all of which showed effects in women and no or less pronounced effects in men.

The first systematic screen for sex-difference in the genetics of nine anthropometric traits (including WHR_{adjBMI}) was published in 2013 (Randall et al., 2013). Randall and colleagues conducted 2-stage sex-stratified GWAMAs based on sex-specific GWAS results that comprised up to 77,598 individuals in discovery stage. Variants were filtered for sex-specific association in the discovery stage and tested for sex-difference in an independent replication stage. Despite the increased multiple testing burdens (considering nine anthropometric traits in parallel), two novel sexually dimorphic (women-specific) loci for WHR_{adjBMI} were identified.

Importantly, the here presented power computations showed that Randall and colleagues did not implement the optimal approach. In the here presented work, application of the recommended 1-stage (or 2-stage) approaches to the Randall et al data sets showed

that four (or three) additional sexually dimorphic loci could have been identified. **Figure 24** illustrates that the additionally highlighted loci cover most of the sex-difference that was left undetected by Randall and colleagues. In particular, the recommended combination of 1-stage approaches (genome-wide screen for difference in combination with testing SNPs for difference that were filtered for overall association), exploit the underlying sex-difference much better than the analysis strategy conducted by Randall and colleagues.

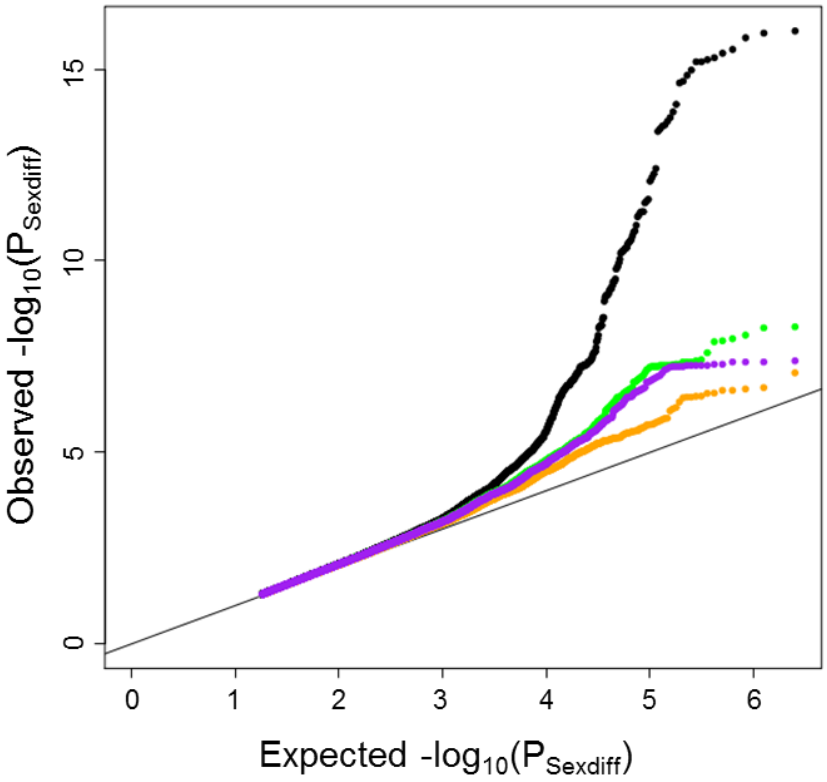


Figure 24. *QQ plot for WHR_{adjBMI} sex-difference P-Values from the combined (discovery + replication) stage data from Randall and colleagues: All genome-wide SNPs (black), after excluding the six loci reported by Randall and colleagues (green), after excluding the seven loci detected by the best 2-stage approaches (purple), and after excluding the 10 loci detected by the best 1-stage approaches (orange). The green curve reflects the sex-difference that was left undetected by Randall and colleagues. The 1-stage approaches detected the sex-difference most efficiently.*

In 2015, a third round of GIANT consortium GWAMAs for WHR_{adjBMI} – now including up to 224,459 individuals - was published (Shungin et al., 2015). Again, this work primarily aimed at the identification of associated loci and successfully identified 49 independent loci with genome-wide significant effects on WHR_{adjBMI}. Consistent with previous work, strong sexual dimorphism in genetic effects for WHR_{adjBMI} was observed among the overall associated variants: 19 of the 49 loci displayed stronger effects in women (and less

pronounced or no effects in men), and (for the first time) one locus displayed a men-specific effect (with no effect in women). The identification of such particular types of sexually dimorphic effects is consistent with the here presented power computations that showed that filtering for overall association prior to difference testing is optimal for SSE and CED type of effects.

Currently, a second systematic screen for sex-difference in the genetics of WHR_{adjBMI} is well underway (Winkler et al, in revision). In this work, age- and sex-stratified GWAMAs were conducted that include up to 216,654 individuals for WHR_{adjBMI}. Based on the here presented power computations, the best approaches were applied that aim at the identification of 2-way (G x AGE or G x SEX) or 3-way (G x AGE x SEX) interaction effects. For WHR_{adjBMI}, 44 loci were identified for significant sex-differences and the number of known sexually dimorphic effects was more than doubled (from 21 to 48, four were missed by the current study).

In summary, of the 48 identified sexually dimorphic WHR_{adjBMI} loci, 32 exhibited a stronger effect on WHR_{adjBMI} in women, five a stronger effect on WHR_{adjBMI} in men and 11 were classified as opposite effect loci.

A chronological overview on the identified sexually dimorphic loci for WHR_{adjBMI} from GIANT analyses is given in **Table 16.**

Table 16. *Overview of applied approaches and number of sexually dimorphic loci identified for WHR_{adjBMI} by the GIANT consortium.*

			Sample size		**Number of identified loci** (novel loci)			
Reference	**Description**	**Approach**	*Discovery*	*Replication*	*Sex-difference*	*Women-specific*	*Men-specific*	*Opposite effect*
(Lindgren et al., 2009)	Overall GWAMA to identify main effects for WHR_{adjBMI}	$\text{Overall}_{5e\text{-}8} \rightarrow \text{Diff}_{bonf}$	37,670	61,612	1 (1)	1 (1)	-	-
(Heid et al., 2010)	Overall GWAMA to identify main effects for WHR_{adjBMI}	$\text{Overall}_{5e\text{-}8} \rightarrow \text{Diff}_{bonf}$	77,167	113,636	7 (6)	7 (6)	-	-
(Randall et al., 2013)	Systematic screen for sex-difference for nine anthropometric traits (incl. WHR_{adjBMI})	$\text{Diff}_{FDR5\%}$ + $\text{Strat}_{FDR5\%} \rightarrow \text{Diff}_{FDR5\%}$	77,598	108,832	6 (2)	6 (2)	-	-
(Shungin et al., 2015)	Overall GWAMA to identify main effects for WHR_{adjBMI}	$\text{Overall}_{5e\text{-}8} \rightarrow \text{Diff}_{bonf}$	224,459	-	20 (12)	19 (11)	1 (1)	-
(Winkler et al, in revision)	Systematic screen for age-difference, sex-difference, and for 3-way G x AGE x SEX interaction effects for two obesity traits (incl. WHR_{adjBMI})	$\text{Diff}_{FDR5\%}$ + $\text{Overall}_{1e\text{-}5} \rightarrow \text{Diff}_{FDR5\%}$	216,654	-	44 (27)	28 (12)	5 (4)	11 (11)
		Total number of sexually dimorphic loci identified:			**48**	**32**	**5**	**11**

The enrichment of women-specific loci (32 among 48 loci with sex-difference, P_{Binom} = 0.015) is consistent with heritability analyses that showed a significantly higher heritability of WHR_{adjBMI} in women than in men (Shungin et al., 2015; Zillikens et al., 2008). Using the Framingham Heart study, family-based heritability of WHR_{adjBMI} in women was estimated to be 46% as compared to 19% for men (Shungin et al., 2015). Similar trends were observed in unrelated subjects and using all 2.5 million variants (Winkler et al, in revision). Moreover, variants found for overall association explained a significantly larger fraction of WHR_{adjBMI} variance in women than in men (2.4% explained variation in women, 0.8% in men, Winkler et al, in revision) (Shungin et al., 2015).

However, the identified sexually dimorphic effects explained only few of the missing heritability of WHR_{adjBMI}. A substantial fraction of the missing heritability may still be hidden in missed rare variant genetic effects, for which the current analyses were underpowered. For example, the most recent overall GIANT scan only identified main effect variants down to 5% allele frequency (Shungin et al., 2015) and the most recent systematic sex-difference scan only identified sex-specific effects (except for some opposite effects) down to 20% allele frequency (Winkler et al, in revision) (**Figure 25**). Thus, even larger sample sizes are needed to identify sex-difference for less common or rare variants or sex-difference that is based on even smaller – but realistic – genetic effect sizes.

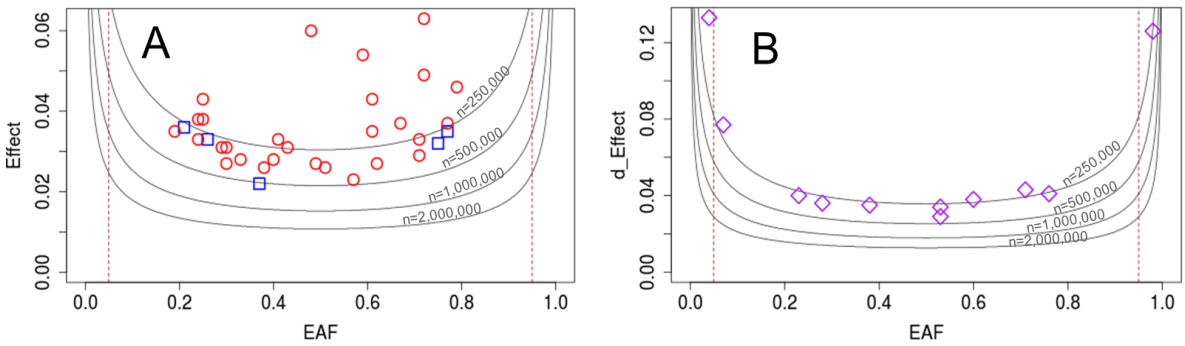


Figure 25. *Effect sizes versus allele frequencies: Effect size of the 44 sexually dimorphic WHR_{adjBMI} loci detected from the GIANT AGE x SEX meta-analyses, are depicted over allele frequency. A: Shown on the y-axis is the stronger sex-specific effect size of the 33 detected SSE and CED loci (red: 28 with stronger effect in women; blue: five with stronger effect in men). The curves reflect the power to find an SSE effect with the approach $[Overall_{1e\text{-}5} \rightarrow Diff_{\alpha_{Bonf}}]$ for varying total sample sizes. B: Shown on the y-axis is the effect difference for the 11 detected OED loci. The curves reflect the power to find an OED effect with the approach $[Diff_{5e\text{-}8}]$ for varying total sample sizes.*

Remarkably, many of the identified sexually dimorphic variants are located in biologically interesting regions.

For example, one locus found for its women-specific association with WHR_{adjBMI} is located in vicinity of the *GRB14* gene*,* of which the encoded protein has a reported role in insulin sensitivity and lipid metabolism (Goenaga et al., 2009; Ridker et al., 2009). The identified WHR_{adjBMI} variant showed a similar sex-specific and directionally consistent association pattern with HDL-cholesterol, triglycerides, fasting insulin and type 2 diabetes (Randall et al., 2013) as well as a significant (although not sex-specific) association with increased *GRB14* expression levels in human subcutaneous adipose tissue. In addition, a significant sexually dimorphic association of *GRB14* with lower expression in brown fat of female mice was observed that was lacking in male mice. Due to the reported role of brown fat in triglyceride metabolism, this is consistent with the reported women-specific association on HDL-cholesterol and triglycerides.

Another particularly interesting region is the women-specific association with WHR_{adjBMI} near *PPARG*, which is well-known for its relevant role in type 2 diabetes therapy. The encoded protein PPARγ regulates white adipocyte differentiation and influences adipogenesis, obesity as well as insulin sensitivity (Lehrke & Lazar, 2005; Nakagami, 2013). Follow-up analyses highlighted a significantly higher expression of *PPARG* in liver of female mice that was lacking in male mice (Randall et al., 2013). Interestingly, *PPARG* is predicted to bind at a transcription factor binding site that is in high LD with the lead SNP of another region (near *HSD17B4*) that was also found for its women-specific association on WHR_{adjBMI}. The enzyme *HSD17B4* itself is an interesting candidate because it converts the hormone estradiol into estrone and influences steroid metabolism as well as fatty acid oxidation (Leenders et al., 1996; Thompson, Dzubur, Wade, & Tomaszycki, 2011).

In summary, the identification of sexually dimorphic WHR_{adjBMI} regions supports the hypothesis that sex-differences in body fat distribution are regulated by a complex interplay of autosomal genetic factors.

6.3.2 Differences in genetic effects for obesity measures between smokers and non-smokers

Preliminary results from a smoking-status stratified GWAMA for obesity traits (Justice et al, in progress, own work), including BMI, waist circumference adjusted for BMI (WC_{adjBMI}) and waist-hip ratio adjusted for BMI (WHR_{adjBMI}), highlight a number of loci with significant differences in genetic effects between smokers and non-smokers.

For example, a variant near *PRNP* displayed a genome-wide significant difference in genetic effects on WC_{adjBMI} between smokers and non-smokers. The WC_{adjBMI} increasing allele in smokers showed a decreasing effect in non-smokers, which is consistent with the fact that smokers exhibit on average larger waist circumference than non-smokers. *PRNP* is highly expressed throughout the central nervous system (as are many obesity genes) and influences oxidative stress response (Kachiwala et al., 2005; Zomosa-Signoret, Arnaud, Fontes, Alvarez-Martinez, & Liautard, 2008). Taken together, this makes it a strong biological candidate gene for obesity and interaction with smoking.

Another interesting gene-smoking interaction effect has been identified for BMI and is located near *CHRNB4* (Cholinergic Nicotine Receptor B4), which confirms previous work (Freathy et al., 2011). The *CHRNB4* locus with a reported role in nicotine addiction (Picciotto & Kenny, 2013) is also an interesting finding from a methodological point of view. It was recently described that spurious gene-environment interaction can occur for a proxy marker if the true causal genetic variant is associated with the environmental exposure (Dudbridge & Fletcher, 2014). In the case of *CHRNB4,* the dependence of a potentially causal variant with the environmental factor smoking might be given through the reported association with nicotine addiction. Therefore, the observed gene-smoking interaction has to be interpreted with care and sensitivity analyses are required to establish the finding.

6.3.3 Age- and sex-specific genetic effects for obesity measures

Given that overall and central obesity measures change over time makes age another important factor that may influence genetic effects on obesity. Since changes in obesity measures differ between men and women, some genetic effects may possibly even be modified by both, age and sex simultaneously.

To systematically screen for 2-way G x SEX (sex-differences, independent of AGE), G x AGE (age-differences, independent of SEX) and 3-way G x AGE x SEX interaction effects, recently, a large-scale age- and sex-stratified GWAMAs were conducted that included up to 320,485 individuals for BMI and up to 216,654 individuals for WHR_{adjBMI} (Winkler et al, in revision). Herewith, age was dichotomized at a threshold of 50 years of age that reflects the mean age of menopause in women.

In summary, the study highlighted (i) 44 sex-dependent effects for WHR_{adjBMI} with predominantly larger effects in women (as noted and discussed before), and (ii) 15 age-dependent effects for BMI with predominantly larger effects in younger individuals. Although the study was well powered for any 2-way interaction effects, no age-difference was observed for WHR_{adjBMI} and no sex-difference was observed for BMI. Furthermore, no 3-way G x AGE x SEX interaction effects were discovered at all, which may however be attributable to low power of the study for plausible 1-stratum or 3-strata interaction effects.

The highlighted age-dependent BMI loci included some biologically interesting candidates, such as the most prominent BMI loci near *FTO* and near *MC4R*. The age-specifity of many of the identified loci is supported by longitudinal studies (den Hoed et al., 2010; Elks et al., 2012; Graff et al., 2013; Hardy et al., 2010; Hertel et al., 2011; Sovio et al., 2011).

The enrichment of loci showing stronger BMI effects in the younger subgroup (11 of the 15 loci) may mirror an accumulation of environmental and lifestyle factors over life that mask the genetic effect on obesity in the older group. Interestingly, the four loci exhibiting stronger effects in the older subgroup have a reported role on related diseases, such as type 2 diabetes or coronary heart disease, and may reflect disease relevant processes that are distinct from other BMI loci.

Admittedly, the age-dependency may also reflect a cohort effect that could have been introduced by the stratification of age into younger and older individuals - yielding different environmental or genetic make overs. For example, the increased exposure to high caloric food intake over the past 30 years may have influenced younger individuals differently than older individuals. Indeed, assuming no cohort effect, significant differences in genetic effects on BMI between two age groups imply change of BMI at some point during life. Yet, large-scale longitudinal studies are required to improve accuracy of age-dependent genetic effect estimates and to define the time in life with greater precision at which the genetic loci contribute to BMI changes.

6.4 Relevance of the developed Easy software packages

Due to the lack of software packages tailored for quality control and evaluation of large-scale stratified GWAMA data sets, two *R* packages called *EasyQC* and *EasyStrata* were developed. *EasyQC* provides extended features for quality control of multiple single-study GWAS results, as well as for stratified GWAMA results (Winkler, Day, et al., 2014). *EasyStrata* provides extended statistical and graphical features for the evaluation of stratified GWAMA results and can be utilized to apply the here discussed screening approaches and to extract significant results (Winkler, Kutalik, et al., 2014).

6. Discussion

The two software packages have specifically been developed for GIANT consortium GWAMA projects. They were applied for the quality control of study-specific GWAS or GWAMA results data sets as well as for evaluation of stratified GWAMA results for the most recent rounds of GIANT consortium meta-analyses (Locke et al., 2015; Shungin et al., 2015; Wood et al., 2014)(Winkler et al, in revision; Justice et al, in progress).

Both packages were integrated into the (specifically developed) *Easy* framework that contains efficient methods to handle big data and provides a scripting interface. The scripting interface makes the programs easy to use and simplifies handling of hundreds of large scale GWAS or GWAMA data sets. This diminishes computational burdens and allows less computationally experienced analysts to outline pre-defined quality control or statistical evaluation pipelines.

Notably, there are other tools available for GWAS data, which however only provide some parts of the quality control, graphical features or statistical functionality as captured by the two *Easy* packages (**Table 17**). Most of the other tools are either applicable for single study genotype files or for conducting meta-analyses of multiple GWAS.

Remarkably, *EasyStrata* is the only tool available to date that was specifically designed and developed for evaluation of stratified GWAMA results data, and is the only tool that provides the joint test (based on stratified outcomes) or the 3-way difference-of-difference test.

Table 17. *Comparison of GWAS software tools.*

	QC functionality			Plotting functionality				Statistical functionality					
Software	**File-level**[1]	**Meta-level**	**Meta-anal. level**	**QQ plot**	**Manhattan plot**	**Miami plot**	**Scatter plot**	**Meta-anal.**	**Joint test**	**Diff test**	**Diff-Diff Test**	**Hetero-geneity**	**Reference**
Haploview	no	no	no	no	yes	no	no	no	no	no	no	no	(Barrett, Fry, Maller, & Daly, 2005)
PLINK	no	no	no	no	no	no	no	yes	no	yes	no	yes	(Purcell et al., 2007)
GenABEL	no	no	no	yes	yes	no	yes	yes[4]	no	no	no	no	(Aulchenko et al., 2007)
metal	no	no	no	no	no	no	no	yes	no	no	no	yes	(Willer et al., 2010)
GWAMA	no	no	no	yes[2]	yes[2]	no	no	yes	yes[3]	no	no	yes[3]	(Magi & Morris, 2010)
GWASpi	no	no	no	yes	yes	no	no	no	no	no	no	no	(Muniz-Fernandez, Carreno-Torres, Morcillo-Suarez, & Navarro, 2011)
GWAtoolbox	yes	yes	no	yes	no	no	no	no	no	no	no	no	(Fuchsberger, Taliun, Pramstaller, Pattaro, & consortium, 2012)
GWASTools	yes	no	no	yes	yes	no	no	no	no	no	no	no	(Gogarten et al., 2012)
EasyStrata	no	no	no	yes	yes	yes	yes	yes	yes	yes	yes	yes	(Winkler, Kutalik, et al., 2014)
EasyQC	yes	yes	yes	yes	no	no	yes	yes	no	no	no	no	(Winkler, Day, et al., 2014)

[1] File-level QC denotes QC of GWAs summary results data, not QC of the raw genotype data.
[2] Plots are available through separate R scripts.
[3] The software GWAMA provides an option "--sex" that allows for running a gender-differentiated analysis, that is similar to the joint test of two groups (i.e. men and women), and that allows for testing for heterogeneity between the sexes. Both methods are limited to two subgroups (sexes) and are not applicable to analyses involving more than two strata.
[4] Meta-analysis available through the MetABEL software.

6.5 Strengths and Limitations

In the following chapter, important strength and limitations of this work are discussed.

The major strength of this work is that it fills an important research gap for a specific study design configuration: It is the first structured methodological evaluation of genome-wide screening approaches to identify *gene-strata (G x S) interaction effects* for *continuous outcomes*, given a *stratified model* and a *meta-analysis framework*. In contrast, other methodological work has - so far - focused on single-study gene-environment (G x E) interaction methods for binary outcomes and for interaction models. Moreover, this is the first methodological work to consider 3-way G x S_1 x S_2 interaction effects at all.

Although this work focused on continuous phenotypes, the methodological results can readily be extended to stratified GWAMAs involving dichotomous phenotypes, such as binary disease outcomes. By log-transformation, the pooled genetic odds ratios and confidence intervals – as obtained from stratified GWAMAs for dichotomous phenotypes – can easily be transferred into pooled effect estimates and standard errors, and then be applied to the discussed screening approaches.

A specific strength is that the methodological work is accompanied by easy-to-use software packages that are able to deal with large-scale GWAMA results data and that make the methods easily accessible for other scientists.

Finally, evaluated methods and the developed software were applied and exemplified using unpublished and unique real sex-stratified, smoking-status-stratified or age- and sex-stratified GWAMA data sets from the GIANT consortium.

Admittedly, there are also some limitations to this work, most of which are related to assumptions made for the power computations.

First, power formulae were derived based on the assumption of equal phenotype distributions between strata. Due to requesting uniformly transformed phenotypes for each stratum, this assumption is always fulfilled in the GIANT consortium analyses. Notably, other GWAMA consortia may employ unequal phenotype distributions between strata.

Second, power formulae were derived based on the assumption of equal genotype distributions between strata, an assumption that builds upon random mating between strata. Random mating between men and women is fulfilled and is not an issue for comparisons between younger and older individuals. However, the concept of random mating might be violated for example between smokers and non-smokers, because smokers might rather mate smokers and non-smokers might rather mate non-smokers. The impact of non-random mating is largely ignored in GWAMA methodology altogether.

Third, power computations assumed similar populations, i.e., homogenous genotype distributions and equal genetic effects across studies. Given these assumptions, it has been

shown that a fixed effect meta-analysis of GWAS summary statistics from multiple studies yields approximatively identical results as one large GWAS including all individuals (Evangelou & Ioannidis, 2013). Similarly, meta-analysis of multiple study-specific interaction estimates have been demonstrated to yield approximatively identical results as fitting an interaction model in one large study (Sung et al., 2014). This assumption might be violated for large-scale meta-analyses that include diverse populations, such as multiple ethnicities across studies.

In principle, violation to any of these three assumptions necessitates adapting the analytical power formulae. However, on average, one can assume that the overall message would still be valid.

To avoid overly complexity, the methodological evaluation of approaches for 3-way interaction effects was limited to balanced 1-stage designs. Nevertheless, the methodological results were particularly interesting and may be an important start for future methodological work on 3-way interaction effects that may involve unbalanced and 2-stage designs.

Similarly, this work was limited to additively modelled genotypes and it was not evaluated how or whether the results transfer to recessive or dominant genotype effects. Typically, GWAMA projects focus on additively modelled genotypes because those are considered to be the best compromise to identify different types of genotype-phenotype effects.

Finally, the discussed stratified GWAMA approaches were limited to identify interaction with dichotomous environmental variables. However, it was already noted that continuous environmental variables can always be dichotomized and that in some instances, this might even be a better model fit to the reality than utilizing the continuous variable.

6.6 Conclusion and Outlook

In conclusion, a structured methodological evaluation of stratified GWAMA approaches to identify difference in genetic effects between strata has been outlined, its value has been documented using GIANT consortium data, and software to implement the approaches has been developed.

First, a systematic scheme of stratified GWAMA approaches to identify stratum-difference was developed. Simulation-based evaluations of type 1 error rates as well as analytical comparisons of power were performed and approaches were recommended for specific types of stratum-difference as well as for 1- or 2-stage study designs. For example, testing variants for stratum-difference that were initially filtered for overall (strata-combined) association, yielded valid type 1 error rates and displayed the largest power to identify variants that exhibited an effect in one stratum but lacked effect or showed a less pronounced effect in the other stratum.

6. Discussion

The methodological evaluation was extended to stratified GWAMAs involving two dichotomous stratification variables: AGE (younger vs older than 50 years of age, which reflects menopause in women) and SEX (men vs women). A statistical test for difference-of-difference was introduced that can be utilized to test for 3-way G x AGE x SEX interaction effects given a stratified GWAMA model. Again, multiple approaches to identify 3-way G x AGE x SEX interaction effects were defined and compared with regards to type 1 error and power. Recommendations have been given for specific types of 3-way G x AGE x SEX interaction effects. For example, testing variants for difference-of-difference that were pre-filtered for marginal joint association yielded valid type 1 error rates and displayed the largest power to identify variants that exhibit an effect in one stratum (e.g., younger women) but lack effect in all other strata, or variants that exhibit an identical effect in three strata but lack effect in one stratum. Although only balanced 1-stage age- and sex-stratified GWAMA designs were considered, these results may be an important start for follow up methodological work on 3-way interaction effects.

Application of the stratified GWAMA approaches to obesity traits has proven to successfully pinpoint biologically relevant stratum-differences, such as interactions with sex, age or smoking status. For example, significant sex-differences were observed at 48 independent loci for WHR_{adjBMI} (as a measure of central obesity), and significant age-differences were observed at 15 independent loci for BMI (as a measure of overall obesity). Vice versa, no age-difference was observed for WHR_{adjBMI} and no sex-difference was observed for BMI. Adding to the reported separation of central versus overall obesity genetics into adipose or insulin-related versus central nervous system related biology (Locke et al., 2015; Shungin et al., 2015), this further separates obesity genetics into central obesity genetics being sex-specific but not affected by age versus overall obesity genetics being age-specific, but not affected by sex.

Notably, many of the highlighted stratum-differences in genetic effects of obesity measures have been identified at disease-relevant loci, a circumstance that may be an important start to guide and to improve personalized treatment options. Generally, identification of genes that influence obesity and whose effects are modified by environmental factors will help understanding the complex interplay between genetic susceptibility, environmental factors, obesity and obesity-related diseases. Yet, more functional follow-up analyses are required to further unravel likely causal candidates and to elucidate the genetic functions affecting stratum-differences in obesity measures.

Despite the recent success, power computations also demonstrated that even larger sample sizes are required to identify stratum-differences for rare variants, stratum-differences comprising even smaller effect sizes, or biologically plausible 3-way G x AGE x SEX

interaction effects. The upcoming 1000 Genomes initiative of the GIANT consortium including up to 1 million subjects is expected to – at least in part - fill in that gap.

Besides the stratification by relevant dichotomous environmental factors, future studies may also be interested in stratification by disease status. For example, one may be interested in whether the genetic effect on obesity differs significantly between healthy individuals and individuals with type 2 diabetes. Such efforts can similarly be implemented into the stratified GWAMA setting and the here discussed approaches can directly be employed to investigate potential gene-disease interaction effects.

Two software packages – called *EasyQC* and *EasyStrata* – were developed that facilitate quality control and statistical or graphical evaluation of stratified GWAMA data. Both packages are increasingly being used by various GWAMA consortia and an implementation of *EasyQC* into a cloud-computing framework is already in work.

7 Summary

While genome-wide association meta-analyses (GWAMA) have largely contributed to the understanding of the genetics of complex diseases, such as obesity, little has been known about whether or not the genetic effects differ between strata, such as between men and women, or between smokers and non-smokers. Thus, this work focused on stratified GWAMA screening approaches to identify variants with significant stratum-difference in genetic effects.

A structured methodological evaluation of approaches to identify differences in genetic effects between two strata with regard to type 1 error and power was performed. This evaluation differentiated between situations where one data set was available (1-stage approaches) or where two independent data sets were to be utilized (2-stage approaches). For 1-stage designs, (i) as expected, a genome-wide screen for difference is the best approach to detect variants with opposite effect directions among strata, and (ii) surprisingly, the naive approach of filtering for overall (strata-combined) association followed by a test for difference is the best approach for variants with no or less pronounced (but concordant) effect in one stratum. Remarkably, filtering for joint association violated the type 1 error of the difference test when both steps are conducted in the same data set. For 2-stage designs, (i) filtering for joint association and for nominally significant difference in the discovery stage followed by a test for difference in the replication stage is the best approach for variants with no effect in one stratum or variants with opposite effect directions among strata; and (ii) filtering for overall association in the discovery stage followed by a test for difference using the combined (discovery + replication) stage data is the best approach for variants with less pronounced (but concordant) effect in one stratum. Interestingly, unequal stratum sizes did not impact the type 1 error of the difference test, but power computations showed that it is generally more likely to identify variants with stronger effect in the larger stratum.

The methodological evaluation was extended to GWAMAs stratified by AGE (younger vs older than 50 years of age, which reflects menopause) and by SEX (men vs women) aiming at the identification of 3-way G x AGE x SEX interaction effects. A difference-of-difference test was introduced that can be employed to test for 3-way G x AGE x SEX interaction effects under a stratified GWAMA setting. As expected, a genome-wide screen for difference-of-difference is the best approach for extreme 3-way interaction effects that involve opposite effect directions across AGE and SEX. Surprisingly, filtering for marginal joint association followed by a test for difference-of-difference turned out to be the best approach for biologically more plausible 3-way interaction effects that involve an effect in one

stratum (e.g., younger women) and no effect in the other three, or that involve an identical effect in three strata and no effect in one.

Application of recommended approaches to stratified GWAMA data sets for waist-hip ratio (adjusted for BMI) from the GIANT consortium underscored the power computations results, and highlighted a number of biologically interesting differences in common genetic effects between sexes (predominantly larger effects in women) or between smokers and non-smokers. Most likely, due to the low power to detect plausible 3-way interaction effects, no 3-way G x AGE x SEX interaction effects were identified. Still, larger sample sizes are required to efficiently identify stratum-difference for rare variants, or to identify plausible 3-way interaction effects.

Two powerful and easy-to-use software packages called *EasyQC* and *EasyStrata* were developed that facilitate conduct, quality control and evaluation of stratified GWAMAs. The software packages are already utilized by GIANT consortium analysts as well as by other genetic research consortia.

In summary, the methodological results, supported by the application to GIANT consortium data and the provided software implementation may guide and help the conduct of future stratified GWAMAs aiming at the identification of gene-strata interaction effects. A better understanding of stratum-specific genetic effects for diseases and disease-related traits will ultimately lead to improved knowledge of etiology of the diseases and pinpoint novel treatment options.

8 Zusammenfassung

Genomweite Assoziations Meta-Analysen (GWAMAs) haben wesentlich zum besseren Verständnis der Genetik von komplexen Krankheiten, wie z.B. Adipositas, beigetragen. Allerdings ist wenig darüber bekannt, ob sich genetische Effekte zwischen Subgruppen, wie z.B. zwischen Männern und Frauen oder zwischen Rauchern und Nichtrauchern, unterscheiden. Die vorliegende Arbeit befasste sich daher mit stratifizierten GWAMA Ansätzen zum Identifizieren von genetischen Varianten, deren Effekte sich signifikant zwischen Subgruppen unterscheiden.

Zunächst wurde eine strukturierte, methodische Evaluierung von stratifizierten GWAMA Ansätzen zum Detektieren von Unterschieden zwischen zwei Gruppen, in Bezug auf Typ 1 Fehler und Power, durchgeführt. Hierbei wurde zwischen Situationen unterschieden, in denen entweder ein Datensatz (1-Phasen Design) oder zwei unabhängige Datensätze (Discovery + Replication, 2-Phasen Design) verwendet wurden.

Für 1-Phasen Designs zeigte sich die genomweite Suche nach Effekt-Unterschied als am besten geeignet für Varianten deren Effekte zwischen den zwei Subgruppen in unterschiedliche Richtungen zeigen. Überaschenderweise zeigte sich der naive Ansatz, zuerst für allgemeine (subgruppen-kombinierte) Effekte zu filtern und dann auf Effekt-Differenz zu testen, als am besten geeignet für Varianten die keinen oder einen kleineren (aber gleichgerichteten) Effekt in einer Subgruppe zeigen. Bemerkenswert war, dass das Filtern für Joint-Assoziation eine Erhöhung des Typ 1 Fehlers des Effekt-Differenz Tests zur Folge hatte, wenn dieser im gleichen Datensatz durchgeführt wird.

Für 2-Phasen Designs zeigte sich der Ansatz, im Discovery-Datensatz für Joint-Assoziation sowie für nominal signifikante Effekt-Differenz zu filtern und im Replication-Datensatz auf Effekt-Differenz zu testen, als am besten geeignet für Varianten deren Effekte zwischen den zwei Subgruppen in unterschiedliche Richtungen zeigen und für Varianten die keinen Effekt in einer Subgruppe zeigen. Des Weiteren zeigte sich der Ansatz, im Discovery-Datensatz für allgemeine (subgruppen-kombinierte) Assoziation zu filtern und im kombinierten (Discovery + Replication) Datensatz auf Differenz zu testen, als am besten geeignet für Varianten die einen kleineren (aber gleichgerichteten) Effekt in einer Subgruppe zeigen.

Interessanterweise hatten ungleiche Subgruppengrößen keinen Einfluss auf den Typ 1 Fehler des Effekt-Differenz Tests. Powerberechnungen zeigten jedoch, dass es im Allgemeinen wahrscheinlicher ist Varianten zu finden, deren stärkerer Effekt in der größeren Subgruppe vorkommt.

8. Zusammenfassung

Die methodische Evaluierung wurde auf alters- und geschlechts-stratifizierte GWAMAs erweitert, welche auf die Identifizierung von 3-fach G x AGE x SEX Interaktionen abzielen (mit SEX = Geschlecht, AGE = Alter dichotomisiert bei 50 Jahren, entspricht Menopause bei Frauen). Ein Differenz-Differenz Test wurde eingeführt, welcher zum Testen auf 3-fach G x AGE x SEX Interaktionen innerhalb einer stratifizierten Auswertung verwendet werden kann. Wie zu erwarten war, zeigte sich die genomweite Suche nach Differenz-Differenz als am besten geeignet für Varianten mit extremen 3-fach Interaktionen, die mit unterschiedlichen Richtungen hinsichtlich AGE und SEX einhergehen. Überraschenderweise zeigte sich der Ansatz zuerst für marginale Joint-Effekte zu filtern und dann auf Differenz-Differenz zu testen als am besten geeignet für biologisch plausible 3-fach Interaktionen, welche entweder einen Effekt in einer Subgruppe (z.B. in jungen Frauen) und keinen Effekt in allen anderen zeigen, oder einen identischen Effekt in drei Subgruppen und keinen Effekt in einer Subgruppe zeigen.

Die Anwendung der besten Ansätze auf reale stratifizierte GWAMA Datensätze für Taille-Hüft-Verhältnis (adjustiert für BMI) aus dem GIANT Konsortium unterstrich die Ergebnisse der Powerberechnungen und identifizierte eine Reihe von häufig-vorkommenden, biologisch interessanten, genetischen Effekt-Differenzen zwischen Geschlechtern (mit überwiegend stärkeren Effekten bei Frauen), als auch zwischen Rauchern und Nichtrauchern. Wahrscheinlich wurden wegen zu geringer Power keine biologisch plausiblen 3-fach Interaktionen gefunden. Um Effekt-Differenzen für seltene Varianten oder plausible 3-fach Interaktionen detektieren zu können, werden noch größere Stichprobenumfänge gebraucht.

Zwei leistungsfähige und einfach zu verwendende Software-Pakete namens *EasyQC* und *EasyStrata* wurden entwickelt, welche die Durchführung, Qualitätskontrolle und Auswertung von stratifizierten GWAMAs ermöglichen. Die Softwarepakete werden bereits von Analysten des GIANT-Konsortiums als auch von anderen genetischen Forschungskonsortien verwendet.

Zusammenfassend lässt sich sagen, dass die methodischen Ergebnisse, bestätigt durch die Anwendung auf Daten des GIANT Konsortiums, und die zur Verfügung gestellte Softwareimplementierung, die Durchführung zukünftiger stratifizierter GWAMA Auswertungen zum Detektieren von Gen-Subgruppen Interaktionen, unterstützen und anleiten können. Ein besseres Verständnis der subgruppen-spezifischen genetischen Effekten von Krankheiten und krankheits-relevanten Merkmalen wird letztendlich zu einem besseren Verständnis der Ursachen und Entstehung von Krankheiten beitragen.

9 Appendix

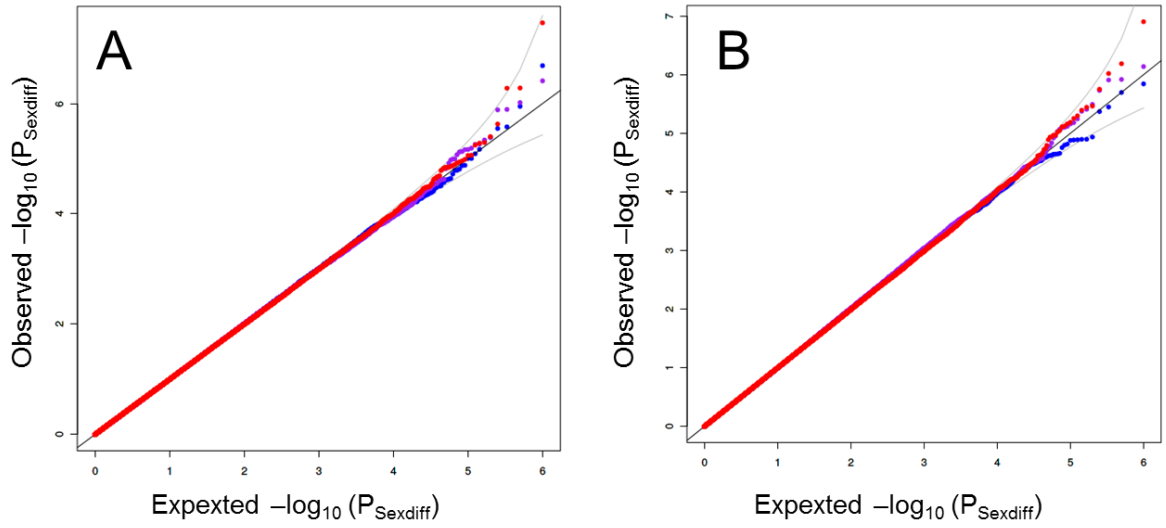
9.1 Validation of simulated data sets used for evaluation of type 1 error of approaches involving two strata

9.1.1 Validation of simulated data sets for 1-stage approaches

Prior to investigating type 1 error of the various 1-stage approaches, the simulated data sets were validated (**Appendix Table 1**). For each SNP, the sex-specific data sets – simulated under $H_o^{noeffect}$ – were shown to reflect non-associated sex-specific data sets (False-positive rate < 5.02%). Likewise, for each SNP, the sex-specific data sets – simulated under H_o^{effect} – were shown to reflect true associated sex-specific data sets that imply an effect that can approximately be detected with 80% power (True-positive rate > 71.6%). Further, the simulated data was shown to truly reflect the two implied sex-difference null hypotheses $H_o^{noeffect}: b_M = b_M = 0$ (no sex-difference, no sex-specific effect) and $H_o^{effect}: b_M = b_M \neq 0$ (no sex-difference, identical sex-specific effects). For each of the two implied null hypotheses and for each SNP, the type 1 error rate of the sex-difference test was estimated to be < 5.01% (**Appendix Table 1**). Further, no inflation of sex-difference P-values could be observed in the QQ plots (**Appendix Figure 1**). Contrariwise to the fact that identification of sex-difference from unassociated sex-specific data (given by $H_o^{noeffect}$) obviously is impossible, this suggests that significantly identified sex-difference automatically implies significant association in at least one of the sexes.

Appendix Table 1. *The table shows observed false-positive rates for the sex-specific data sets that were simulated under $H_o^{noeffect}$ (no sex-specific effects), observed true-positive rates for the sex-specific data sets that were simulated under H_o^{effect} (with sex-specific effects), and type 1 error rates (T1ER) for the resulting sex-difference test.*

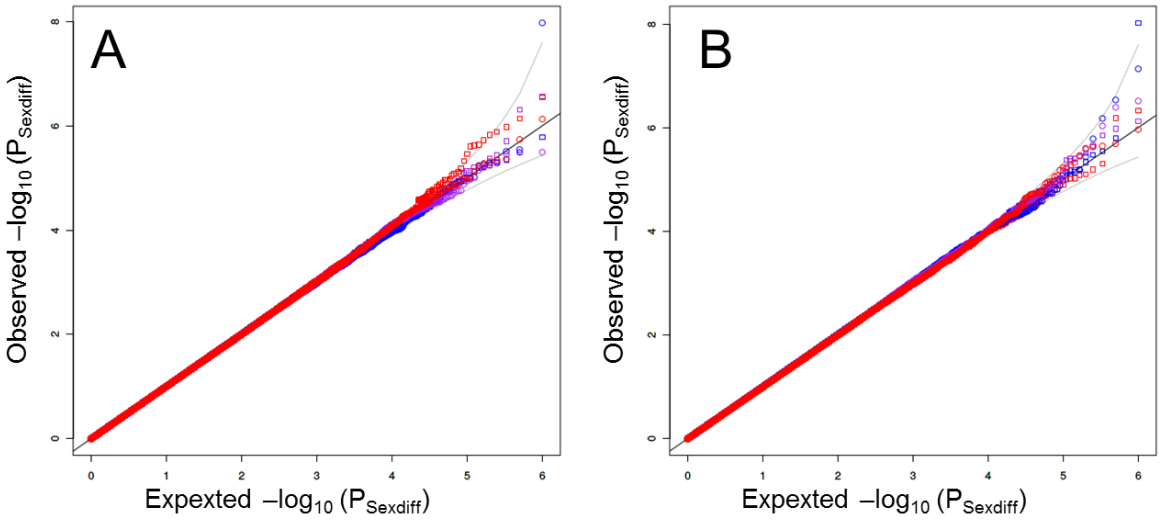
		$H_o^{noeffect}$			H_o^{effect}		
		False-Positive Rate		T1ER	True-Positive Rate		T1ER
SNP	**MAF**	*Men*	*Women*	*Sexdiff*	*Men*	*Women*	*Sexdiff*
rs6002481	0.02	5.00	5.02	5.01	71.6	82.0	5.00
rs8138968	0.30	4.99	5.01	5.01	78.5	80.0	4.99
rs6007738	0.50	5.01	5.02	5.01	79.3	79.3	4.98



Appendix Figure 1. *QQ plots showing the distribution of 1,000,000 sex-difference P-Values that are based on 1,500 men and 1,500 women from the KORA study. The distributions are depicted for each of the three SNPs, rs6002481 (blue), rs8138968 (purple) and rs6007738 (red) and for each of the two implied null hypotheses, A:* $H_o^{noeffect}$*, and B:* H_o^{effect}*.*

9.1.2 Validation of simulated data sets for 2-stage approaches

Again, prior to investigating type 1 error of the 2-stage approaches, the simulated data sets were validated. The sex-specific data sets simulated under $H_o^{noeffect}$ reflected true non-associated sex-specific data sets: False-positive rates for the sex-specific association tests ranged from 4.99 to 5.08% across SNPs and stages. The sex-specific data sets simulated under H_o^{effect} reflected true associated sex-specific data sets that imply an effect that can approximately be detected with 80% power: True-positive rates for the sex-specific association tests ranged from 62.0 to 85.5% across SNPs and stages. Further, the simulated data was shown to truly reflect the implied sex-difference null hypotheses $H_o^{noeffect}$: $b_M = b_M = 0$ (no sex-difference, no sex-specific effect) or H_o^{effect}: $b_M = b_M \neq 0$ (no sex-difference, identical sex-specific effects): Across stages, SNPs and null hypotheses, the type 1 error rate of the sex-difference test ranged from 4.97 to 5.09% and no inflation of sex-difference P-values could be observed in the QQ plots (**Appendix Figure 2**).



Appendix Figure 2. *QQ plots showing the distribution of 1,000,000 sex-difference P-Values, each based on 750 men and 750 women from the KORA study. The distributions are depicted for each SNP, rs6002481 (blue), rs8138968 (purple) and rs6007738 (red), for each stage, discovery (circles) and replication (squares) and for each of the two implied null hypotheses, A:* $H_o^{noeffect}$ *and B:* H_o^{effect}.

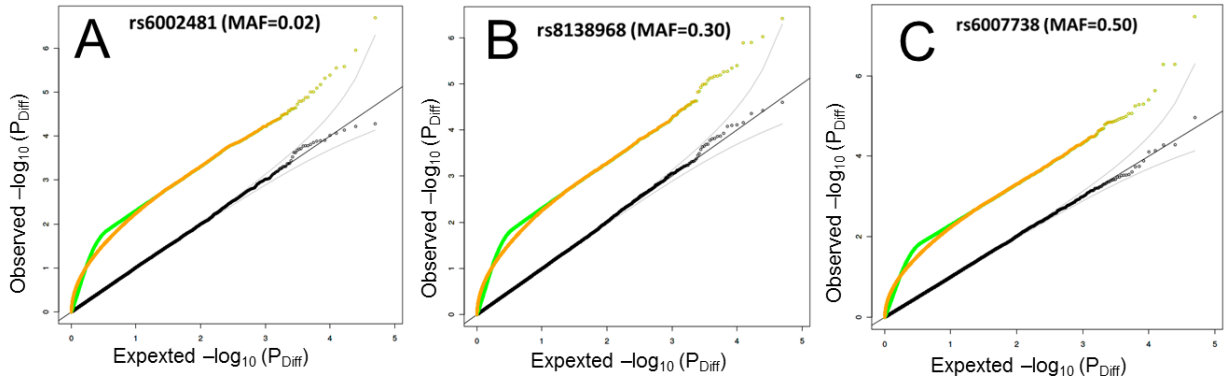
9.2 Simulation-based inference of statistical dependence between filtering and difference tests

Prior to computing power for the considered approaches, dependence of filtering steps and the difference test had to be inferred. This was important because formula (22) (to analytically calculate the power of an approach) can only just be applied if the implemented steps are independent.

Due to using identical or overlapping subjects, all multi-test 1- or combined 2-stage approaches (i.e., those that involve multiple statistical tests) require statistical independence of steps in order to calculate power analytically by formula (22). To evaluate this, statistical dependence of steps was inferred using the simulation-based data set from type 1 error evaluation. The simulated data sets (created under $H_o^{noeffect}$) were filtered for nominal significant overall, stratified or joint association (α_1 = 0.05), and the distribution of difference P-Values for the remaining subset of SNPs was depicted in QQ plots (**Appendix Figure 3**).

Clearly, filtering for overall association did not affect the distribution of difference P-Values. Thus, the overall association test can be considered as independent of the difference test and equation (22) can be employed to analytically calculate the power of approach $[\mathrm{Overall}_{\alpha_1} \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and of approach $[[\mathrm{Overall}_{\alpha_1}] \rightarrow \mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$.

In contrast, filtering for stratified or joint association yielded an inflated distribution of difference P-Values. Thus, the stratified or the joint test association tests have to be considered as dependent of the difference test and equation (22) cannot be employed to calculate the power for the discovery stage of approaches $[\mathrm{Strat}_{\alpha_1},\mathrm{Diff}_{\alpha_2}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$ and $[\mathrm{Joint}_{\alpha_1},\mathrm{Diff}_{\alpha_2}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{diff}}}]$. Power for the discovery stages of these approaches has to be estimated from simulated probability distributions (see **Appendix 9.3** for details).



Appendix Figure 3. *Inference of statistical dependence between filtering and difference tests. Shown is the distribution of difference P-Values after filtering for nominal significant overall association (black), stratified association (orange) and joint association (green). The plots depict simulation results for the 1-stage approaches based on* $H_o^{noeffect}$ *and for SNP A: rs6002481 (MAF=0.02), B: rs8138969 (MAF=0.3), and C: rs6007738 (MAF=0.5).*

9.3 Simulation-based estimation of power for approaches with dependent discovery steps

For the replication-based 2-stage approaches $[\mathrm{Strat}_{\alpha_1},\mathrm{Diff}_{\alpha_2}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ and $[\mathrm{Joint}_{\alpha_1},\mathrm{Diff}_{\alpha_2}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ that include discovery filtering steps on P_{Strat} and $P_{Joint,}$ respectively, accompanied by a second discovery filtering step on P_{Diff}, the power of the discovery stage needs to be calculated using simulated probability distributions. The reason for this is that the two discovery filtering steps are dependent.

9.3.1 Estimation of discovery power for approach $[\mathbf{Strat}_{\boldsymbol{\alpha_1}},\mathbf{Diff}_{\boldsymbol{\alpha_2}}]\rightarrow[\mathbf{Diff}_{\boldsymbol{\alpha_{bonf}}}]$

The discovery power of approach $[\mathrm{Strat}_{\alpha_1},\mathrm{Diff}_{\alpha_2}]\rightarrow[\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ $(Pwr_{Discovery}=Pwr_{Strat\cap Diff})$ that set sets the two discovery steps into an AND relationship has to be estimated based on

simulated probability distributions. Here, n_1 refers to the discovery sample size of stratum 1. The probability $Pwr_{Strat\cap Diff}$ can be written as

$$Pwr_{Strat\cap Diff}(\alpha_1,\alpha_2,R_1^2,R_2^2,n_1,f) = P\left(\left\{T_1 \le t_{n_1-2,\frac{\alpha_1/2}{2}} \cup T_1 \ge t_{n_1-2,1-\frac{\alpha_1/2}{2}} \cup T_2 \le t_{fn_1-2,\frac{\alpha_1/2}{2}} \cup T_2 \ge t_{fn_1-2,1-\frac{\alpha_1/2}{2}}\right\} \cap \left\{Z_{Diff} \le z_{\frac{\alpha_2}{2}} \cup Z_{Diff} \ge z_{1-\frac{\alpha_2}{2}}\right\}\right) =$$

$$= P\left(\left\{\left(T_1^* \le t_{n_1-2,\frac{\alpha_1/2}{2}} - k_t(R_1^2,n_1)\right) \cup \left(T_1^* \ge t_{n_1-2,1-\frac{\alpha_1/2}{2}} - k_t(R_1^2,n_1)\right) \cup \left(T_2^* \le t_{fn_1-2,\frac{\alpha_1/2}{2}} - k_t(R_2^2,fn_1)\right) \cup \left(T_2^* \ge t_{fn_1-2,1-\frac{\alpha_1/2}{2}} - k_t(R_2^2,fn_1)\right)\right\} \cap \left\{\left(Z_{diff}^* \le z_{\frac{\alpha_2}{2}} - k_{diff}(R_1^2,R_2^2,n_1,f)\right) \cup \left(Z_{diff}^* \ge z_{1-\frac{\alpha_2}{2}} - k_{diff}(R_1^2,R_2^2,n_1,f)\right)\right\}\right) \quad (43)$$

where – under the alternative hypothesis – the stratum-specific t statistics $T_i^* \sim t_{n_i-2}$ (i = 1,2, i.e., the two strata) follow a t distribution with n_i-2 df, and $Z_{diff}^* \sim N(0,1)$ follows a standard normal distribution. To account for the multiple testing of two strata, the stratum-specific α-levels are corrected for two tests. The constants k_t *and* k_{diff} can be calculated from the given stratum-specific explained variances R_i^2 as follows:

$$k_t(R_i^2,n_i) = \sqrt{n_i R_i^2/(1-R_i^2)} \quad (44)$$

$$k_{diff}(R_1^2,R_2^2,n_1,f) = \sqrt{n_1}\frac{|R_1|-|R_2|}{\sqrt{1-R_1^2+\frac{1}{f}(1-R_2^2)}} \quad (45).$$

The probability of the intersection is estimated from 100,000 simulated data points for T_1^*, T_2^* and Z_{diff}^* for which the distribution parameters and thresholds are known or inferred from the given values.

9.3.2 Estimation of discovery power for approach $[\mathbf{Joint}_{\boldsymbol{\alpha_1}}, \mathbf{Diff}_{\boldsymbol{\alpha_2}}] \rightarrow [\mathbf{Diff}_{\boldsymbol{\alpha_{bonf}}}]$

The discovery power of approach $[\mathrm{Joint}_{\alpha_1}, \mathrm{Diff}_{\alpha_2}] \rightarrow [\mathrm{Diff}_{\alpha_{\mathrm{Bonf}}}]$ ($Pwr_{Discovery} = Pwr_{Joint\cap Diff}$) that set sets the two discovery steps into an AND relationship has to be estimated based on simulated probability distributions. Here, n_1 refers to the stratum 1 sample size of the discovery stage. The probability $Pwr_{Joint\cap Diff}$ can be written as

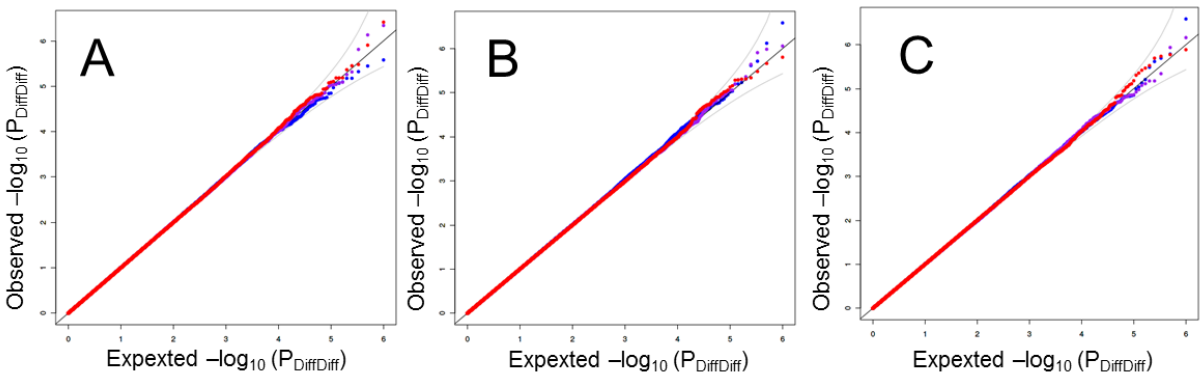
$$Pwr_{Joint\cap Diff}(\alpha_1,\alpha_2,R_1^2,R_2^2,n_1,f) =$$
$$= P\left(\left\{C_{Joint} \ge \chi_{2,1-\alpha_1}\right\} \cap \left\{Z_{Diff} \le z_{\frac{\alpha_2}{2}} \cup Z_{Diff} \ge z_{1-\frac{\alpha_2}{2}}\right\}\right) = \quad (46)$$

$$= P\left(\left\{C_{Joint} \geq \chi_{2,1-\alpha_1}\right\}\right.$$
$$\cap\left\{\left(Z^*_{Diff} \leq z_{\frac{\alpha_2}{2}} - k_{diff}(R_1^2, R_2^2, n_1, f)\right)\right.$$
$$\left.\left.\cup\left(Z^*_{Diff} \geq z_{1-\frac{\alpha_2}{2}} - k_{diff}(R_1^2, R_2^2, n_1, f)\right)\right\}\right)$$

where – under the alternative hypothesis - $C_{Joint} \sim \mathrm{X}^2_{2,\lambda(R_1^2,R_2^2,n_1,f)}$ follows a non-central chi-squared distribution with 2 df and non-centrality parameter λ that can be estimated using formula (20), and $Z^*_{Diff} \sim \mathrm{N}(0,1)$ follows a standard normal distribution. Again, the constant k_{diff} can be calculated using equation (45). The probability of the intersection is estimated from 100,000 simulated data points for C_{Joint} and Z^*_{Diff} for which the distribution parameters and thresholds are known or inferred from the given values.

9.4 Validation of simulated data sets used for evaluation of type 1 error of age- and sex-stratified GWAMA approaches

Prior to investigating type 1 error of the age- and sex-stratified GWAMA approaches, the simulated data sets were validated. Across SNPs and across implied null hypotheses, the type 1 error rate of the difference-of-difference test ranged from 4.99% to 5.05% and no inflation of difference-of-difference P-Values was observable in QQ plots (**Appendix Figure 4**). Thus, the simulated data sets were considered to reflect real 'difference-of-difference' null data sets.



Appendix Figure 4. *QQ plots showing the distribution of 1,000,000 age- and sex-difference P-Values that are based on 750 younger men, 750 older men, 750 younger women and 750 older women from the KORA study. The distributions are depicted for each SNP, rs6002481 (blue), rs8138968 (purple) and rs6007738 (red) and for each of the three implied null hypotheses, A:* $H_o^{noeffect}$*, B:* H_o^{effect} *and C:* $H_o^{sexdiff}$*.*

9.5 Simulation-based inference of statistical dependence between steps for the age- and sex-stratified GWAMA approaches

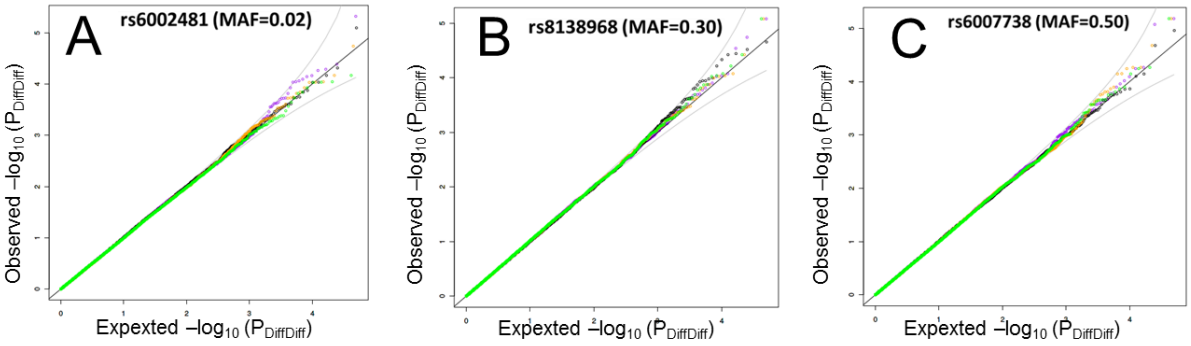
Prior to computing power for the validated approaches, dependence of filtering steps (marginal difference , marginal stratified, marginal joint and overall test) and the difference test had to be inferred. This was important because formula (22) (to analytically calculate the power of an approach) can only just be applied if the implemented steps are independent.

In addition, to calculate power of a marginal test, independence of the respective sex- and age-version is required. Each marginal test is a combination of an age- and a sex-version and the power formulae of the marginal tests include an intersection of the age- and sex-version, which again can only be multiplicatively calculated if the age- and sex-versions of the test are independent.

9.5.1 Dependence between filtering and difference-of-difference tests

Due to using identical subjects, all validated approaches ($[\mathrm{Overall}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$, $[\mathrm{MarDiff}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$, $[\mathrm{MarStrat}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ and $[\mathrm{MarJoint}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$) require statistical independence of steps in order to calculate power analytically by formula (22). To evaluate this, statistical dependence of steps was inferred using the simulation-based data set from type 1 error evaluation. The simulated data sets (created under $H_o^{noeffect}$) were filtered for nominal significant overall, marginal difference, marginal stratified, or marginal joint association (α_1 = 0.05), and the distribution of difference-of-difference P-Values for the remaining subset of SNPs was depicted in QQ plots (**Appendix Figure 5**).

Clearly, none of the four filtering tests did affect the distribution of difference-of-difference P-Values. Thus, filtering on overall association, marginal difference, marginal stratified or marginal joint association can be considered as independent of the difference-of-difference test and equation (22) can be employed to analytically calculate the power of the respective approach.



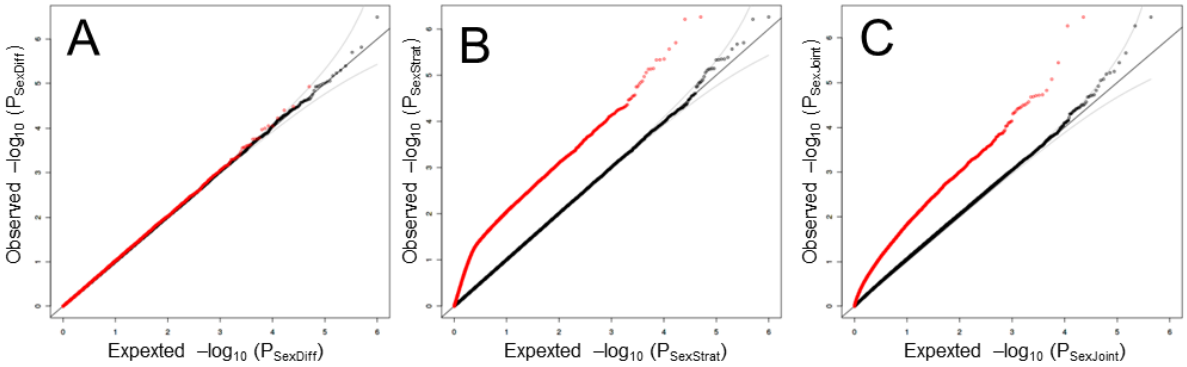
Appendix Figure 5. *Inference of statistical dependence between validated filtering and difference-of-difference tests. Shown is the distribution of difference-of-difference P-Values after filtering for nominal significant overall association (black), marginal difference (purple), marginal stratified association (orange) and marginal joint association (green). The plots depict simulation results based on $H_{0\text{-}noeffect}$ and for SNP A: rs6002481 (MAF=0.02), B: rs8138969 (MAF=0.3), and C: rs6007738 (MAF=0.5).*

9.5.2 Dependence between marginal age- and marginal sex-tests

Next, dependence of the marginal age- and sex-tests was evaluated for approaches $[\mathrm{MarStrat}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$, $[\mathrm{MarDiff}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ and $[\mathrm{MarJoint}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$. This is important because formula (22) can only just be applied - to analytically calculate power of step 1 - if the implemented step 1 tests are statistically independent. The simulated data from the type 1 error inference (for $H_o^{noeffect}$ and for rs8138968, MAF = 0.3) was used to evaluate dependence of tests. The distribution of marginal sex-tests was illustrated in dependence of marginal age-tests (**Appendix Figure 6**).

Clearly, the marginal age-difference and the marginal sex-difference test were found to be statistically independent, i.e. filtering on age-difference cannot bias the distribution of the sex-difference. Thus, power of step 1 (the marginal difference test) can completely be analytically calculated using formula (22) for approach $[\mathrm{MarDiff}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$.

In contrast, the marginal age-stratified and the marginal sex-stratified test, as well the marginal age-joint and the marginal sex-joint test, were found to be statistically dependent., Thus, the power computation has to be supplied by simulated probability distributions for approaches $[\mathrm{MarStrat}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ and $[\mathrm{MarJoint}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ (see **Appendix 9.6** for details).



Appendix Figure 6. *Inference of statistical dependence between marginal age- and marginal sex-tests. A: Distribution of marginal $P_{Sexdiff}$ with (red) and without (black) filtering for marginal $P_{Agediff} < 0.05$; B: Distribution of marginal $P_{Sexstrat}$ with (red) and without (black) filtering for marginal $P_{Agestrat} < 0.05$; and C: Distribution of marginal $P_{Sexjoint}$ with (red) and without (black) filtering for marginal $P_{Agejoint} < 0.05$.*

9.6 Simulation-based estimation of power for age- and sex-stratified GWAMA approaches that involve dependent statistical tests

For the age- and sex-stratified GWAMA approaches $[\mathrm{MarStrat}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ and $[\mathrm{MarJoint}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{DiffDiff}}}]$ that include initial filtering steps on $P_{MarStrat}$ and $P_{MarJoint,}$ respectively, the power of the filtering step needs to be calculated using simulated probability distributions. The reason for this is that for both approaches, the filtering steps comprise a combination of two statistically dependent tests.

9.6.1 Estimation of filtering power for approach $[\mathbf{MarStrat}_{\boldsymbol{\alpha_1}} \rightarrow \mathbf{DiffDiff}_{\boldsymbol{\alpha_{bonf}}}]$

The filtering power of approach $[\mathrm{MarStrat}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ ($Pwr_{MarStrat} = Pwr_{Agestrat \cup Sexstrat} = Pwr_{Agestrat} + Pwr_{Sexstrat} - Pwr_{Agestrat \cap Sexstrat}$) that sets the dependent marginal age-stratified and marginal sex-stratified tests into an OR relationship has to be supported by simulated probability distributions for the power of the intersection $Pwr_{Agestrat \cap Sexstrat}$. The probability $Pwr_{Agestrat \cap Sexstrat}$ can be written in dependence of the stratum-specific sample size *n*, the α-level and the stratum-specific explained variances as follows:

$$Pwr_{Agestrat\cap Sexstrat}(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) =$$

$$P\left(\left\{T_{\leq 50} \leq t_{2n-2,\frac{\alpha/4}{2}} \cup T_{\leq 50} \geq t_{2n-2,1-\frac{\alpha/4}{2}} \cup T_{>50} \leq t_{2n-2,\frac{\alpha/4}{2}} \cup T_{>50} \geq t_{2n-2,1-\frac{\alpha/4}{2}}\right\} \cap \left\{T_F \leq t_{2n-2,\frac{\alpha/4}{2}} \cup T_F \geq t_{2n-2,1-\frac{\alpha/4}{2}} \cup T_M \leq t_{2n-2,\frac{\alpha/4}{2}} \cup T_M \geq t_{2n-2,1-\frac{\alpha/4}{2}}\right\}\right) =$$

$$= P\left(\left\{\left(T^*_{\leq 50} \leq t_{2n-2,\frac{\alpha/4}{2}} - k_t(R^2_{M\leq 50}, R^2_{F\leq 50}, n)\right) \cup \left(T^*_{\leq 50} \geq t_{2n-2,1-\frac{\alpha/4}{2}} - k_t(R^2_{M\leq 50}, R^2_{F\leq 50}, n)\right) \cup \left(T^*_{>50} \leq t_{2n-2,\frac{\alpha/4}{2}} - k_t(R^2_{M>50}, R^2_{F>50}, n)\right) \cup \left(T^*_{>50} \geq t_{2n-2,1-\frac{\alpha/4}{2}} - k_t(R^2_{M>50}, R^2_{F>50}, n)\right)\right\} \cap \left\{\left(T^*_{M} \leq t_{2n-2,\frac{\alpha/4}{2}} - k_t(R^2_{M\leq 50}, R^2_{M>50}, n)\right) \cup \left(T^*_{M} \geq t_{2n-2,1-\frac{\alpha/4}{2}} - k_t(R^2_{M\leq 50}, R^2_{M>50}, n)\right) \cup \left(T^*_{F} \leq t_{2n-2,\frac{\alpha/4}{2}} - k_t(R^2_{F\leq 50}, R^2_{F>50}, n)\right) \cup \left(T^*_{F} \geq t_{2n-2,1-\frac{\alpha/4}{2}} - k_t(R^2_{F\leq 50}, R^2_{F>50}, n)\right)\right\}\right) \quad (47).$$

Assuming the alternative hypothesis, the marginal t statistics $T_i^* \sim t_{2n-2}$ (*i* = *≤50*, *>50*, *M* or *F,* i.e., the age- and sex-marginals) follow a t distribution with *2n-2* df. To account for the multiple testing of four marginal strata (younger, older, men, women), the α-levels are corrected for four tests. The constants k_t can be calculated from two stratum-specific explained variances R_i^2 each (e.g., for younger individual, from younger men and from younger women) and from the stratum-specific sample size n:

$$k_t(R_1^2, R_2^2, n) = \sqrt{n}\ \frac{\sum_{i=1}^{2} \frac{|R_i|}{(1-R_i^2)}}{\sqrt{\sum_{i=1}^{2} \frac{1}{(1-R_i^2)}}} \quad (48)$$

The probability of the intersection is estimated from 100,000 simulated data points for T_i^*, for which the distribution parameters and thresholds are known or inferred from the given values.

9.6.2 Estimation of filtering power for approach $[\mathbf{MarJoint_{\alpha_1}} \rightarrow \mathbf{DiffDiff_{\alpha_{bonf}}}]$

The filtering power of approach $[\mathrm{MarJoint}_{\alpha_1} \rightarrow \mathrm{DiffDiff}_{\alpha_{\mathrm{Bonf}}}]$ ($Pwr_{MarJoint} = Pwr_{Agejoint\cup Sexjoint} = Pwr_{Agejoint} + Pwr_{Sexjoint} - Pwr_{Agejoint\cap Sexjoint}$) that sets the dependent marginal age-joint and marginal sex-joint tests into an OR relationship has to be supported by simulated probability distributions for the power of the intersection $Pwr_{Agejoint\cap Sexjoint}$. The probability $Pwr_{Agejoint\cap Sexjoint}$ can be written in dependence of the stratum-specific sample size *n*, the α-level and the stratum-specific explained variances as follows:

$$Pwr_{Agejoint \cap Sexjoint}(\alpha, R^2_{M\leq 50}, R^2_{F\leq 50}, R^2_{M>50}, R^2_{F>50}, n) = \\ = P\left(C_{Agejoint} \geq \chi_{2,1-\frac{\alpha}{2}} \cap C_{Sexjoint} \geq \chi_{2,1-\frac{\alpha}{2}}\right) \quad (49)$$

, where – under the alternative hypothesis - $C_{Agejoint}$ and $C_{Sexjoint}$ follow a non-central chi-squared distribution with 2 df and non-centrality parameter λ that can be estimated using formula (20): $\lambda_{Agejoint} = \lambda\left((R^2_{M\leq 50}, R^2_{F\leq 50}), (R^2_{M>50}, R^2_{F>50}), n\right)$ and $\lambda_{Sexjoint} = \lambda\left((R^2_{M\leq 50}, R^2_{M>50}), (R^2_{F\leq 50}, R^2_{F>50}), n\right)$. To account for the multiple testing of two 2-way joint tests, the α-levels are corrected for two tests. The probability of the intersection is estimated from 100,000 simulated data points for $C_{Agejoint}$ and $C_{Sexjoint}$, for which the distribution parameters and thresholds are known or inferred from the given values.

9.7 Example EasyStrata pipeline to evaluate sex-difference

In the following, an example ecf-file pipeline (called *sexdiff.ecf*) to evaluate sex-stratified GWAMA results data is shown:

```
##############################################################################################
##############################################################################################
#######################         Example EasyStrata ecf-pipeline        ######################
##############################################################################################
## EasyStrata configuration parameters:

DEFINE  --pathOut /path2output/results
        --acolIn SNP;A1;A2;EAF;BETA;SE;P
        --acolInClasses character;character;character;numeric;numeric;numeric;numeric

EASYIN  --fileIn /path2input/GWAMA_result_Men.txt
        --fileInTag MEN

EASYIN  --fileIn /path2input/GWAMA_result_Women.txt
        --fileInTag WOMEN

##############################################################################################
## EasyStrata scripting interface:

START EASYSTRATA

## Merging the two input files:
MERGEEASYIN --colInMarker SNP --blnMergeAll 0

## Annotating results with Chr and Pos using a reference file
MERGE --colInMarker SNP
      --fileRef /path2reffile/reference_chr_pos.txt
        --acolIn SNP;CHR;POS
        --acolInClasses character;character;integer
      --colRefMarker SNP
      --blnInAll 1
      --blnRefAll 0

## Adjust allele directions in men to women
ADJUSTALLELES  --colInA1 A1.MEN
               --colInA2 A2.MEN
               --colInFreq EAF.MEN
               --colInBeta BETA.MEN
               --colRefA1 A1.WOMEN
               --colRefA2 A2.WOMEN

## Perform difference test (yields sex-difference P-Values)
CALCPDIFF --acolBETAs BETA.MEN;BETA.WOMEN
          --acolSEs SE.MEN;SE.WOMEN
          --colOutPdiff PSexdiff
```

```
## Create QQ plot showing men- and women-specific association P-Values
##   as well as sex-difference P-Values
QQPLOT --acolQQPlot P.MEN;P.WOMEN;PSexdiff
       --astrColour blue;red;magenta
       --numPvalOffset 0.05

## Create a Miami-plot that contrasts men- and women-specific association P-Values
MIAMIPLOT --colMIAMIPlotUp P.WOMEN
          --colMIAMIPlotDown P.MEN
          --colInChr CHR
          --colInPos POS
          --numPvalOffset 0.05

## Create a Manhattan-plot od sex-difference P-Values
MHPLOT  --colMHPlot PSexdiff
       --colInChr CHR
       --colInPos POS
       --numPvalOffset 0.05

## Extract independent lead SNPs with genome-wide significant PSexdiff
##   using a distance-based criterion d<500KB
INDEP --rcdCriterion PSexdiff<5e-8
      --colIndep PSexdiff
      --colInChr CHR
      --colInPos POS
      --numPosLim 500000

STOP EASYSTRATA
##############################################################################################
##############################################################################################
##############################################################################################
```

The pipeline assumes that each sex-stratified GWAMA result data set contains SNP-specific results in rows and contains columns for the effect allele (column *A1*), other allele (column *A2*), the effect allele frequency (column *EAF*), the genetic effect (column *BETA*), with standard error (column *SE*), as well as the association P-value (column *P*).

The pipeline consecutively 1) merges the men- and women-specific GWAMA results, 2) adds chromosomal and base positions to each SNP using a reference mapping file, 3) adjusts allele directions between the men- and women-specific results, 4) performs a sex-difference difference test for each SNP (yields sex-difference P-values for each SNP), 5) creates a QQ plot that simultaneously displays men- and women-specific association P-Values as well as sex-difference P-Values for all SNPs genome-wide, 6) creates a Miami plot that contrast men- and women-specific association P-Values, 7) creates a Manhattan plot of sex-difference P-Values, and finally 8) extracts independent lead-SNPs that display genome-wide significant sex-difference.

To start the pipeline the pipeline is called from the *R* command line by the *EasyStrata* R function: > *EasyStrata("/path2ecffile/sexdiff.ecf").*

9.8 Extended features and applicability of EasyStrata

9.8.1 Extended applicability

9.8.1.1 Dichotomous outcomes.

In contrast to linear regression based stratified GWAMAs, stratified GWAMAs of dichotomous traits report stratum-specific, pooled odds ratio (OR) with respective confidence intervals for each SNP. Since stratum-specific effect sizes and standard errors – required by the statistical approaches implemented in EasyStrata – can be obtained from the OR and the confidence intervals using a log-transformation, the package is fully applicable to dichotomous traits GWAMAs.

9.8.1.2 General GWAMA

Except the functionality for between strata comparison, all implemented features are directly applicable for general GWAMA results. For example, for a single GWAMA result data set, *EasyStrata* can be used to i) extract significant results, e.g. by applying a multiple testing correction to the association P-Values and clumping significant SNPs into independent loci, and ii) to obtain standard graphical presentation of results, e.g., by creating Manhattan- and QQ-plots.

9.8.1.3 Multiple phenotypes

The features of the package to compare GWAMA results between multiple strata can readily be translated to comparing GWAMA results between multiple phenotypes. For example, plotting GWAMAs results on BMI and WHR_{adjBMI} in a single QQ plot or contrasting the results in a Miami-plot may be useful to compare their association content generally or by locus.

9.8.1.4 Interaction analyses

Meta-analyses that are based on study-specific gene-environment interaction testing, i.e. GWAs including an interaction term in the regression model, typically include SNP-specific summary statistics for the SNP main effect, for the SNP x E interaction effect and for the joint (main + interaction) effect (Manning et al., 2011). Again, the general functionality of EasyStrata to extract significant results and to graphically present results can be applied to each of these tests. Some of the functionality to contrast stratum-specific results may be used to compare results between the different tests, e.g., to contrast main association P-Values with the SNP x E interaction P-Values in a Miami- or QQ-plot.

9.8.1.5 Data handling

In addition to the specific functionality described, *EasyStrata* makes available a toolbox of functions for general data handling, data manipulation and data extraction (see **Appendix Table 2** for an overview of general data handling functions).

Appendix Table 2. *Overview of EasyStrata's general data handling functions. More details are shown in the EasyStrata manual that is accessible from www.genepi-regensburg.de/software.*

	Function Name	**Functionality**
Column handling:	ADDCOL	Add column
	EDITCOL	Edit column values
	GETCOLS	Keep subset of columns
	REMOVECOL	Remove column
	RENAMECOL	Rename column
	STRSPLITCOL	Split string column
Row handling:	CLEAN	Remove SNPs
	CRITERION	Extract SNPs
	EXTRACTSNPS	Extract subset of SNPs defined in a separate file
	FILTER	Filter SNPs
Joins:	EASYMERGE	Join data from one separate external file per input
	MERGE	Join data from a single external file to each input
	MERGEEASYIN	Join data from all input files

9.8.2 Extended statistical methods

In addition to the implemented statistical tests to follow up stratified GWAMA results, EasyStrata provides functions to control for multiple testing of large numbers of SNPs and to control for potential population stratification problems:

9.8.2.1 Multiple-testing correction

The statistical tests from above are mostly applied to multiple SNPs, either to all SNPs genome-wide using a hypothesis free screening approach, e.g., to screen for SNPs with significant G x S effects genome-wide, or to a subset of follow-up SNPs, e.g., to follow-up a limited number of known trait-associated SNPs for potential G x S effects. In both cases the test-statistics need to be corrected for the multiple testing of numerous SNPs. EasyStrata's *BONFERRONI* function allows correcting the calculated test statistics for the number of independent test using a conservative Bonferroni-correction (Johnson et al., 2010). Alternatively the function *FDR* offers a less conservative multiple-testing control using a False-Discovery-Rate (FDR) approach (Benjamini & Hochberg, 1995).

9.8.2.2 Genomic control correction

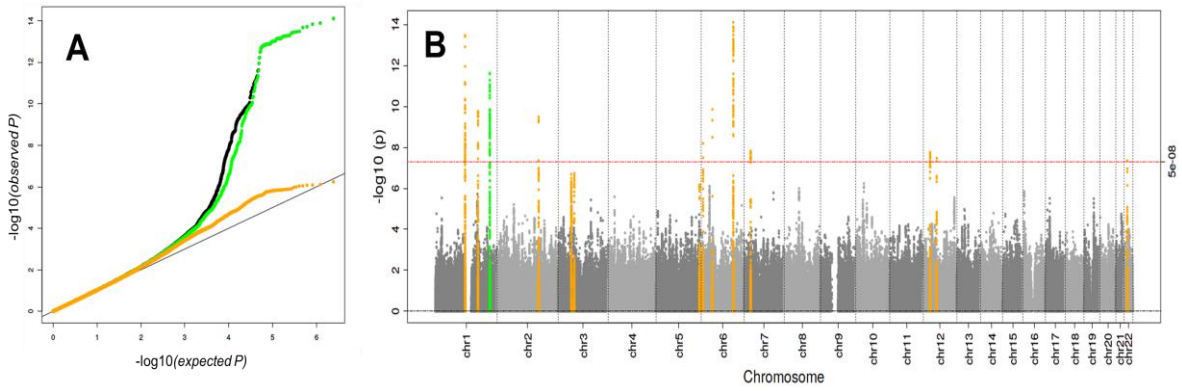
A common problem in GWAs and GWAMAs is the inflation of test statistics due to population stratification. While there are several methods to control for population stratification on the study level, e.g. adjusting for principal components or correcting test-statistics by a study-specific genomic-control inflation factor (single GC correction), the common approach in GWAMAs is to apply a second genomic-control correction to the meta-analyzed data set (double GC correction) (Devlin & Roeder, 1999). *EasyStrata*'s function *GC* facilitates genomic-control of the stratified GWAMA results.

9.8.3 Extended graphical features

In addition to the tailored functionality for between-strata comparison, further convenient graphical features are provided that are applicable not only to stratified but also to general GWAMA results.

9.8.3.1 Locus annotation

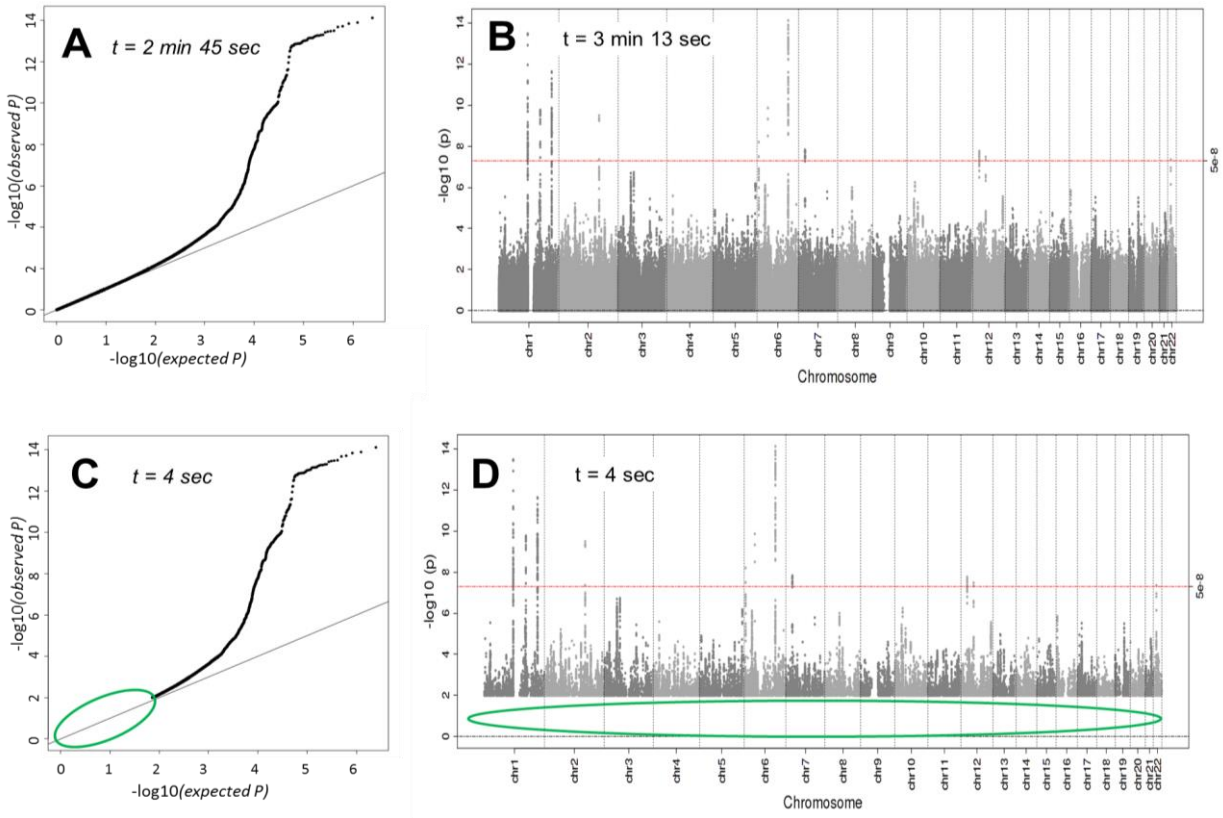
One of these additional graphical features is based on locus annotation. Particularly interesting loci, e.g. previously known or novel loci associated with the trait can be highlighted in Manhattan-or Miami-Plots or be omitted from QQ plots (**Appendix Figure 7**). For example, highlighting previously observed loci in a Manhattan plot allows the user to easily compare novel with previously identified associated loci and allows removing such loci from QQ plots, which is useful to visually examine the strength of the novel associations.



Appendix Figure 7. *QQ and Manhattan plot showing publically available GWAMA results for WHR_{adjBMI} (Heid et al., 2010). The QQ plot displays all SNPs (black), all SNPs after exclusion of the locus previously reported by Lindgren and colleagues (Lindgren et al., 2009)(green) and after exclusion of loci identified by Heid and colleagues. The coloring in the Manhattan plot indicates loci identified by Lindgren (green) and Heid (orange).*

9.8.3.2 Improved plotting speed

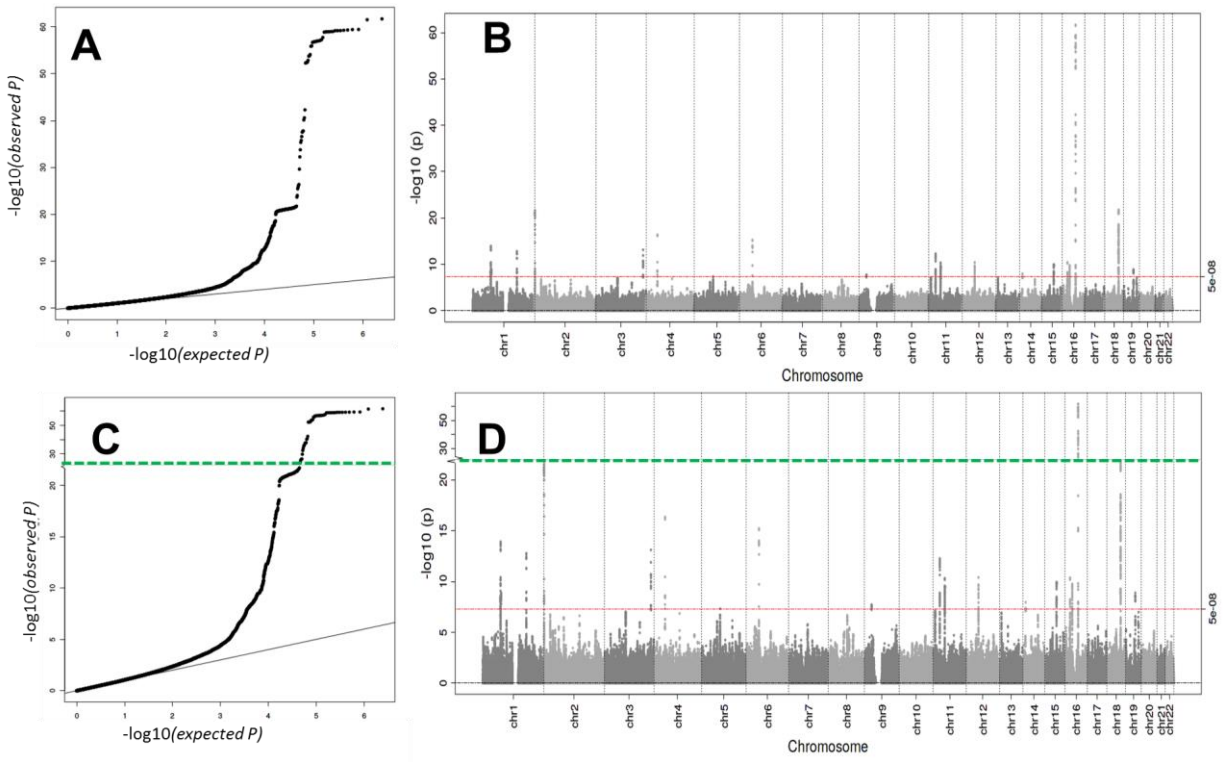
Less significant, i.e. less interesting, SNPs can be excluded from QQ-, Manhattan- and Miami-plots, which can substantially improve plotting speed and thus reduce computational time (**Appendix Figure 8**).



Appendix Figure 8. *Improving plotting speed. QQ and Manhattan plots for the publically available GIANT GWAMA results on WHR_{adjBMI} (Heid et al., 2010).* ***A./B.*** *QQ and Manhattan plots displaying all SNPs (computational times: 2min45sec and 3min13sec, respectively);* ***C./D.*** *QQ and Manhattan plot after exclusion of uninformative SNPs with P>0.05 (computational time: 4sec for both). The plot is based on Supplementary Figure 5 from the EasyStrata manuscript (Winkler, Kutalik, et al., 2014).*

9.8.3.3 Axes breaks

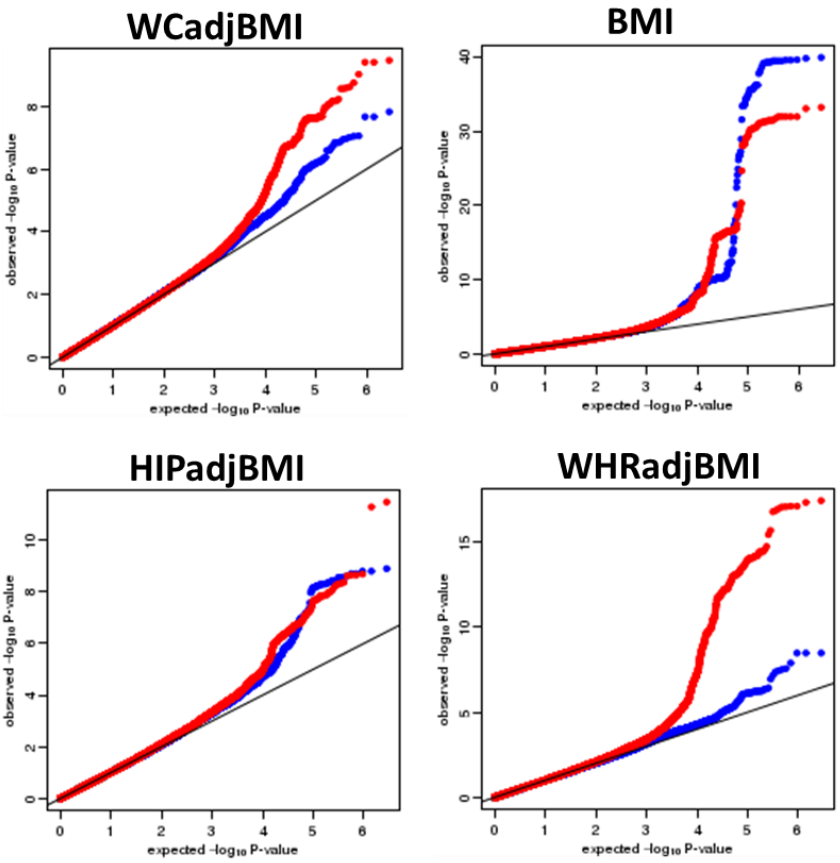
Another helpful feature is the ability to break up the scale of the y-axis in Manhattan-, Miami- and QQ-plots. This is useful when the data set contains loci with very low association P-Values, which would distort the scale of the y-axis and this limit the visibility of other associated loci (**Appendix Figure 9**).



Appendix Figure 9. *Plotting extreme P-values. QQ and Manhattan plots for the publically available GIANT GWAMA results on BMI (Speliotes et al., 2010).* ***A./B.*** *QQ and Manhattan plots using a constant scale;* ***C./D.*** *QQ and Manhattan plot with broken up y-axis scale at y = 22 yielding two different scales. The break in the y-axis scale improves the visibility of less significant loci in the Manhattan plot which are squeezed together otherwise. The plot is based on Supplementary Figure 6 from the EasyStrata manuscript (Winkler, Kutalik, et al., 2014).*

9.8.3.4 Panel of plots

To obtain a quick overview on numerous traits or analyses, QQ or scatter plots can automatically be displayed as panels of plots and stored into a single image file (**Appendix Figure 10**).



Appendix Figure 10. *Panel of QQ plots. This figure combines publically available sex-stratified GIANT GWAMA results (Randall et al., 2013), displaying QQ plots (red: women, blue: men) for BMI, waist and hip circumferences and waist-hip ratio adjusted for BMI. The plot is based on Supplementary Figure 7 from the EasyStrata manuscript (Winkler, Kutalik, et al., 2014).*

References

Abecasis, G. R., & Wigginton, J. E. (2005). Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet, 77*(5), 754-767

Adan, R. A., Tiesjema, B., Hillebrand, J. J., la Fleur, S. E., Kas, M. J., & de Krom, M. (2006). The MC4 receptor and control of appetite. *Br J Pharmacol, 149*(7), 815-827

Andreasen, C. H., Stender-Petersen, K. L., Mogensen, M. S., Torekov, S. S., Wegner, L., Andersen, G., . . . Hansen, T. (2008). Low physical activity accentuates the effect of the FTO rs9939609 polymorphism on body fat accumulation. *Diabetes, 57*(1), 95-101

Aschard, H., Hancock, D. B., London, S. J., & Kraft, P. (2010). Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Hum Hered, 70*(4), 292-300

Aulchenko, Y. S., Ripke, S., Isaacs, A., & van Duijn, C. M. (2007). GenABEL: an R library for genome-wide association analysis. *Bioinformatics, 23*(10), 1294-1296

Aulchenko, Y. S., Struchalin, M. V., & van Duijn, C. M. (2010). ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics, 11*, 134

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263-265

Behrens, G., Winkler, T. W., Gorski, M., Leitzmann, M. F., & Heid, I. M. (2011). To stratify or not to stratify: power considerations for population-based genome-wide association studies of quantitative traits. *Genet Epidemiol, 35*(8), 867-879

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

Caballero, B. (2005). A nutrition paradox--underweight and obesity in developing countries. *N Engl J Med, 352*(15), 1514-1516

Chiolero, A., Faeh, D., Paccaud, F., & Cornuz, J. (2008). Consequences of smoking for body weight, body fat distribution, and insulin resistance. *American Journal of Clinical Nutrition, 87*(4), 801-809

Cochran, W. G. (1954). The Combination of Estimates from Different Experiments. *Biometrics, 10*(1), 101-129

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. London, New York: Chapman and Hall; distributed in U.S. by Halsted Press.

Dai, J. Y., Logsdon, B. A., Huang, Y., Hsu, L., Reiner, A. P., Prentice, R. L., & Kooperberg, C. (2012). Simultaneously testing for marginal genetic association and gene-environment interaction. *Am J Epidemiol, 176*(2), 164-173

den Hoed, M., Ekelund, U., Brage, S., Grontved, A., Zhao, J. H., Sharp, S. J., . . . Loos, R. J. (2010). Genetic susceptibility to obesity and related traits in childhood and adolescence: influence of loci identified by genome-wide association studies. *Diabetes, 59*(11), 2980-2988

Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics, 55*(4), 997-1004

Distefano, J. K., & Taverna, D. M. (2011). Technological issues and experimental design of gene association studies. *Methods Mol Biol, 700*, 3-16

Dudbridge, F., & Fletcher, O. (2014). Gene-environment dependence creates spurious gene-environment interaction. *Am J Hum Genet, 95*(3), 301-307

Edwards, A. W. (2008). G. H. Hardy (1908) and Hardy-Weinberg equilibrium. *Genetics, 179*(3), 1143-1150

Elks, C. E., Loos, R. J. F., Hardy, R., Wills, A. K., Wong, A., Wareham, N. J., . . . Ong, K. K. (2012). Adult obesity susceptibility variants are associated with greater childhood weight gain and a faster tempo of growth: the 1946 British Birth Cohort Study. *American Journal of Clinical Nutrition, 95*(5), 1150-1156

Evangelou, E., & Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet, 14*(6), 379-389

Farooqi, I. S., Keogh, J. M., Yeo, G. S., Lank, E. J., Cheetham, T., & O'Rahilly, S. (2003). Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. *N Engl J Med, 348*(12), 1085-1095

Farooqi, I. S., & O'Rahilly, S. (2000). Recent advances in the genetics of severe childhood obesity. *Arch Dis Child, 83*(1), 31-34

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160

Fischer, J., Koch, L., Emmerling, C., Vierkotten, J., Peters, T., Bruning, J. C., & Ruther, U. (2009). Inactivation of the Fto gene protects from obesity. *Nature, 458*(7240), 894-898

Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., . . . McCarthy, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science, 316*(5826), 889-894

Freathy, R. M., Kazeem, G. R., Morris, R. W., Johnson, P. C. D., Paternoster, L., Ebrahim, S., . . . Munafo, M. (2011). Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index. *Int J Epidemiol, 40*(6), 1617-1628

Fuchsberger, C., Taliun, D., Pramstaller, P. P., Pattaro, C., & consortium, C. K. (2012). GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics, 28*(3), 444-445

Gauderman, W. J., Zhang, P., Morrison, J. L., & Lewinger, J. P. (2013). Finding novel genes by testing G x E interactions in a genome-wide association study. *Genet Epidemiol, 37*(6), 603-613

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073

Goenaga, D., Hampe, C., Carre, N., Cailliau, K., Browaeys-Poly, E., Perdereau, D., . . . Burnol, A. F. (2009). Molecular determinants of Grb14-mediated inhibition of insulin signaling. *Mol Endocrinol, 23*(7), 1043-1051

Gogarten, S. M., Bhangale, T., Conomos, M. P., Laurie, C. A., McHugh, C. P., Painter, I., . . . Laurie, C. C. (2012). GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics, 28*(24), 3329-3331

Graff, M., Gordon-Larsen, P., Lim, U., Fowke, J. H., Love, S. A., Fesinmeyer, M., . . . North, K. E. (2013). The influence of obesity-related single nucleotide polymorphisms on BMI across the life course: the PAGE study. *Diabetes, 62*(5), 1763-1767

Hamza, T. H., Chen, H., Hill-Burns, E. M., Rhodes, S. L., Montimurro, J., Kay, D. M., . . . Payami, H. (2011). Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet, 7*(8), e1002237

Hancock, D. B., Artigas, M. S., Gharib, S. A., Henry, A., Manichaikul, A., Ramasamy, A., . . . London, S. J. (2012). Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet, 8*(12), e1003098

Hardy, R., Wills, A. K., Wong, A., Elks, C. E., Wareham, N. J., Loos, R. J., . . . Ong, K. K. (2010). Life course variations in the associations between FTO and MC4R gene variants and body size. *Hum Mol Genet, 19*(3), 545-552

Haslam, D. W., & James, W. P. (2005). Obesity. *Lancet, 366*(9492), 1197-1209

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., . . . Lindgren, C. M. (2010). Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet, 42*(11), 949-960

Hertel, J. K., Johansson, S., Sonestedt, E., Jonsson, A., Lie, R. T., Platou, C. G., . . . Njolstad, P. R. (2011). FTO, type 2 diabetes, and weight gain throughout adult life: a meta-analysis of 41,504 subjects from the Scandinavian HUNT, MDC, and MPP studies. *Diabetes, 60*(5), 1637-1644

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., & Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A, 106*(23), 9362-9367

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet, 6*(2), 95-108

Hsu, L., Jiao, S., Dai, J. Y., Hutter, C., Peters, U., & Kooperberg, C. (2012). Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol, 36*(3), 183-194

Hutter, C. M., Mechanic, L. E., Chatterjee, N., Kraft, P., Gillanders, E. M., & Tank, N. C. I. G.-E. T. (2013). Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genet Epidemiol, 37*(7), 643-657

International HapMap, C. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320

International HapMap, C., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., . . . McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature, 467*(7311), 52-58

Johnson, R. C., Nelson, G. W., Troyer, J. L., Lautenberger, J. A., Kessing, B. D., Winkler, C. A., & O'Brien, S. J. (2010). Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics, 11*, 724

Kachiwala, S. J., Harris, S. E., Wright, A. F., Hayward, C., Starr, J. M., Whalley, L. J., & Deary, I. J. (2005). Genetic influences on oxidative stress and their association with normal cognitive ageing. *Neurosci Lett, 386*(2), 116-120

Keller, M. C. (2014). Gene x environment interaction studies have not properly controlled for potential confounders: the problem and the (simple) solution. *Biol Psychiatry, 75*(1), 18-24

Khoury, M. J., Beaty, T. H., & Cohen, B. H. (1993). *Fundamentals of genetic epidemiology*. New York: Oxford University Press.

Kirchengast, S. (2010). Gender Differences in Body Composition from Childhood to Old Age: An Evolutionary Point of View. *Journal of Life Sciences, 2*(1), 1-10

Kooperberg, C., & Leblanc, M. (2008). Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol, 32*(3), 255-263

Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J., & Gauderman, W. J. (2007). Exploiting gene-environment interaction to detect genetic associations. *Hum Hered, 63*(2), 111-119

Kuk, J. L., Saunders, T. J., Davidson, L. E., & Ross, R. (2009). Age-related changes in total and regional fat distribution. *Ageing Res Rev, 8*(4), 339-348

Kutalik, Z., Johnson, T., Bochud, M., Mooser, V., Vollenweider, P., Waeber, G., . . . Bergmann, S. (2011). Methods for testing association between uncertain genotypes and quantitative traits. *Biostatistics, 12*(1), 1-17

Lakka, T. A., & Bouchard, C. (2005). Physical activity, obesity and cardiovascular diseases. *Handb Exp Pharmacol*(170), 137-163

Leenders, F., Tesdorpf, J. G., Markus, M., Engel, T., Seedorf, U., & Adamski, J. (1996). Porcine 80-kDa protein reveals intrinsic 17 beta-hydroxysteroid dehydrogenase, fatty acyl-CoA-hydratase/dehydrogenase, and sterol transfer activities. *J Biol Chem, 271*(10), 5438-5442

Lehrke, M., & Lazar, M. A. (2005). The many faces of PPARgamma. *Cell, 123*(6), 993-999

Li, D., & Conti, D. V. (2009). Detecting gene-environment interactions using a combined case-only and case-control approach. *Am J Epidemiol, 169*(4), 497-504

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet, 10*, 387-406

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol, 34*(8), 816-834

Lindgren, C. M., Heid, I. M., Randall, J. C., Lamina, C., Steinthorsdottir, V., Qi, L., . . . Giant, C. (2009). Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet, 5*(6), e1000508

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., . . . Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature, 518*(7538), 197-206

Loomba-Albrecht, L. A., & Styne, D. M. (2009). Effect of puberty on body composition. *Curr Opin Endocrinol Diabetes Obes, 16*(1), 10-15

Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Zhao, J. H., Prokopenko, I., . . . Mohlke, K. L. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet, 40*(6), 768-775

Lovejoy, J. C., Sainsbury, A., & Stock Conference Working, G. (2009). Sex differences in obesity and the regulation of energy homeostasis. *Obes Rev, 10*(2), 154-167

Magi, R., Lindgren, C. M., & Morris, A. P. (2010). Meta-analysis of sex-specific genome-wide association studies. *Genet Epidemiol, 34*(8), 846-853

Magi, R., & Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics, 11*, 288

Manning, A. K., Hivert, M. F., Scott, R. A., Grimsby, J. L., Bouatia-Naji, N., Chen, H., . . . Langenberg, C. (2012). A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet, 44*(6), 659-669

Manning, A. K., LaValley, M., Liu, C. T., Rice, K., An, P., Liu, Y., . . . Dupuis, J. (2011). Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP x environment regression coefficients. *Genet Epidemiol, 35*(1), 11-18

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet, 39*(7), 906-913

Moriarty, J. P., Branda, M. E., Olsen, K. D., Shah, N. D., Borah, B. J., Wagie, A. E., . . . Naessens, J. M. (2012). The effects of incremental costs of smoking and obesity on health care costs among adults: a 7-year longitudinal study. *J Occup Environ Med, 54*(3), 286-291

Mukherjee, B., Ahn, J., Gruber, S. B., Rennert, G., Moreno, V., & Chatterjee, N. (2008). Tests for gene-environment interaction from case-control data: a novel study of type I error, power and designs. *Genet Epidemiol, 32*(7), 615-626

Mukherjee, B., & Chatterjee, N. (2008). Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics, 64*(3), 685-694

Muniz-Fernandez, F., Carreno-Torres, A., Morcillo-Suarez, C., & Navarro, A. (2011). Genome-wide association studies pipeline (GWASpi): a desktop application for genome-wide SNP analysis and management. *Bioinformatics, 27*(13), 1871-1872

Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C., & Gauderman, W. J. (2011). Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol, 35*(3), 201-210

Murcray, C. E., Lewinger, J. P., & Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am J Epidemiol, 169*(2), 219-226

Nakagami, H. (2013). The mechanism of white and brown adipocyte differentiation. *Diabetes Metab J, 37*(2), 85-90

Picciotto, M. R., & Kenny, P. J. (2013). Molecular mechanisms underlying behaviors related to nicotine addiction. *Cold Spring Harb Perspect Med, 3*(1), a012112

Piegorsch, W. W., Weinberg, C. R., & Taylor, J. A. (1994). Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat Med, 13*(2), 153-162

Pischon, T., Boeing, H., Hoffmann, K., Bergmann, M., Schulze, M. B., Overvad, K., . . . Riboli, E. (2008). General and abdominal adiposity and risk of death in Europe. *N Engl J Med, 359*(20), 2105-2120

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

Qi, L., Kraft, P., Hunter, D. J., & Hu, F. B. (2008). The common obesity variant near MC4R gene is associated with higher intakes of total energy and dietary fat, weight change and diabetes risk in women. *Hum Mol Genet, 17*(22), 3502-3508

Randall, J. C., Winkler, T. W., Kutalik, Z., Berndt, S. I., Jackson, A. U., Monda, K. L., . . . Heid, I. M. (2013). Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet, 9*(6), e1003500

Ridker, P. M., Pare, G., Parker, A. N., Zee, R. Y., Miletich, J. P., & Chasman, D. I. (2009). Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study. *Circ Cardiovasc Genet, 2*(1), 26-33

Rose, K. M., Newman, B., Mayer-Davis, E. J., & Selby, J. V. (1998). Genetic and behavioral determinants of waist-hip ratio and waist circumference in women twins. *Obes Res, 6*(6), 383-392

Rosner, B. (2006). *Fundamentals of biostatistics* (6th ed.). Belmont, CA: Thomson-Brooks/Cole.

Rubner-Institut, M. (2008). Nationale Verzehrs-Studie II.

Samanic, C., Chow, W. H., Gridley, G., Jarvholm, B., & Fraumeni, J. F., Jr. (2006). Relation of body mass index to cancer risk in 362,552 Swedish men. *Cancer Causes Control, 17*(7), 901-909

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res, 29*(1), 308-311

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Magi, R., . . . Mohlke, K. L. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature, 518*(7538), 187-196

Simino, J., Sung, Y. J., Kume, R., Schwander, K., & Rao, D. C. (2013). Gene-alcohol interactions identify several novel blood pressure loci including a promising locus near SLC16A9. *Front Genet, 4*, 277

Skol, A. D., Scott, L. J., Abecasis, G. R., & Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet, 38*(2), 209-213

Sovio, U., Mook-Kanamori, D. O., Warrington, N. M., Lawrence, R., Briollais, L., Palmer, C. N. A., . . . Consortium, E. G. G. (2011). Association between Common Variation at the FTO Locus and Changes in Body Mass Index from Infancy to Late Childhood: The Complex Nature of Genetic Association through Growth and Development. *PLoS Genet, 7*(2)

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., . . . Loos, R. J. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet, 42*(11), 937-948

Stouffer, S. A. (1949). *The American soldier*. Princeton,: Princeton University Press.

Sturm, R. (2002). The effects of obesity, smoking, and drinking on medical problems and costs. *Health Aff (Millwood), 21*(2), 245-253

Sung, Y. J., Schwander, K., Arnett, D. K., Kardia, S. L., Rankinen, T., Bouchard, C., . . . Rao, D. C. (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. *Genet Epidemiol, 38*(4), 369-378

Taylor, K. C., Carty, C. L., Dumitrescu, L., Buzkova, P., Cole, S. A., Hindorff, L., . . . Study, P. (2013). Investigation of gene-by-sex interactions for lipid traits in diverse populations from the population architecture using genomics and epidemiology study. *BMC Genet, 14*, 33

Thompson, J. B., Dzubur, E., Wade, J., & Tomaszycki, M. (2011). The effects of estradiol on 17beta-hydroxysteroid dehydrogenase type IV and androgen receptor expression in the developing zebra finch song system. *Brain Res, 1401*, 66-73

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *Am J Hum Genet, 90*(1), 7-24

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., . . . Boehnke, M. (2012). The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet, 8*(8), e1002793

Wells, J. C. (2007). Sexual dimorphism of body composition. *Best Pract Res Clin Endocrinol Metab, 21*(3), 415-430

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006

Wichmann, H. E., Gieger, C., Illig, T., & Group, M. K. S. (2005). KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen, 67 Suppl 1*, S26-30

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics, 26*(17), 2190-2191

Willer, C. J., Speliotes, E. K., Loos, R. J., Li, S., Lindgren, C. M., Heid, I. M., . . . Genetic Investigation of, A. T. C. (2009). Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat Genet, 41*(1), 25-34

Winham, S. J., Cuellar-Barboza, A. B., Oliveros, A., McElroy, S. L., Crow, S., Colby, C., . . . Biernacka, J. M. (2014). Genome-wide association study of bipolar disorder accounting for effect of body mass index identifies a new risk allele in TCF7L2. *Mol Psychiatry, 19*(9), 1010-1016

Winkler, T. W., Day, F. R., Croteau-Chonka, D. C., Wood, A. R., Locke, A. E., Magi, R., . . . Genetic Investigation of Anthropometric Traits, C. (2014). Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc, 9*(5), 1192-1212

Winkler, T. W., Kutalik, Z., Gorski, M., Lottaz, C., Kronenberg, F., & Heid, I. M. (2014). EasyStrata: Evaluation and Visualization of stratified genome-wide association meta-analysis data. *Bioinformatics*

Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., . . . Frayling, T. M. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet, 46*(11), 1173-1186

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet, 9*(5), e1003520

Zillikens, M. C., Yazdanpanah, M., Pardo, L. M., Rivadeneira, F., Aulchenko, Y. S., Oostra, B. A., . . . van Duijn, C. M. (2008). Sex-specific genetic effects influence variation in body composition. *Diabetologia, 51*(12), 2233-2241

Zomosa-Signoret, V., Arnaud, J. D., Fontes, P., Alvarez-Martinez, M. T., & Liautard, J. P. (2008). Physiological role of the cellular prion protein. *Vet Res, 39*(4), 9

List of abbreviations

α_{Bonf}	Bonferroni-corrected α-level
α_{gws}	Genome-wide significance level
BMI	Body mass index
CED	Concordant effect direction
G x E	Gene-environment
G x S	Gene-strata
GWAMA	Genome-wide association meta-analysis
GWAS	Genome-wide association study
LD	Linkage Disequilibrium
M	Number of independent difference tests performed (= number of independent lead-SNPs selected from filtering step)
MAF	Minor allele frequency
OED	Opposite-effect direction
QQ	Quantile-quantile
R^2	Explained variance
SNP	Single Nucleotide Polymorphism
SSE	Single stratum effect
T1ER	Type 1 error
T2D	Type 2 Diabetes
WHR	Waist-hip ratio
WHR_{adjBMI}	Waist-hip ratio adjusted for BMI
WC	Waist circumference
WC_{adjBMI}	Waist circumference adjusted for BMI

List of publications

Related publications and description of own contribution

The content of this work has already given rise to five publications in international peer-reviewed journals:

(1) Randall JC*, Winkler TW*, Kutalik Z*, Berndt SI* … Heid IM. *Sex-stratified Genome-wide Association Studies Including 270,000 Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits.* PLoS Genet. 2013;9(6): e1003500

→ Own contribution: Methods and Software development for quality control and for statistical analysis; Power calculations; Data preparation and quality control; Sex-specific genome-wide association meta-analysis; Interpretation of results; Paper writing

(2) Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, Ferreira T, Fall T, Graff M, Justice AE, Luan J, Gustafsson S, Randall JC, Vedantam S, Workalemahu T, Kilpeläinen TO, Scherag A, Esko T, Kutalik Z, Heid IM*, Loos RJ*. *Quality control and conduct of genome-wide association meta-analyses.* Nat Protoc. 2014 May;9(5):1192-212.

→ Own contribution: Led of all aspects of the paper including design of the paper; Collection and analysis of the example data; Development of the quality control methods and software; Paper writing

(3) Winkler TW, Kutalik Z, Gorski M, Lottaz C, Kronenberg F, Heid IM. *EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data.* Bioinformatics. 2015 Jan 15;31(2):259-61.

→ Own contribution: Led of all aspects of the paper including design of the paper and initiation of the research question; Collection and analysis of example data; Software development including documentation; Paper writing

(4) Shungin DS*, Winkler TW*, Croteau-Chonka DC*, Ferreira T*, Locke AE*, Mägi R* ... Mohlke KL. *New genetic loci link adipose and insulin biology to body fat distribution.* Nature. 2015 Feb 12;518(7538):187-96.

→ Own contribution: Data preparation and quality control; Genome-wide association meta-analyses; Lead analyst for sex-specific analyses; Power calculations; Genetic risk score analysis; Interpretation of results; Paper writing

(5) Winkler TW*, Justice AE*, Graff M*, Barata L* … Loos RJ. *The influence of age and sex on genetic associations with adult body size and shape: a large-scale genome-wide interaction study.* In revision at PLoS Genetics.

→ Own contribution: Design of the analysis; Statistical methods development and implementation in software; Data preparation and quality control; Genome-wide association meta-analyses; Power calculations; Interpretation of results; Paper writing

* These authors contributed equally

Other publications

Methods paper:

(1) Behrens G, Winkler TW, Gorski M, Leitzmann MF, Heid IM. *To stratify or not to stratify: power considerations for population-based genome-wide association studies of quantitative traits.* Genet Epidemiol. 2011 Dec;35(8):867-79.

(2) Gorski M, Winkler TW, Stark K, Müller-Nurasyid M, Ried JS, Grallert H, Weber BH, Heid IM. *Harmonization of study and reference data by PhaseLift: saving time when imputing study data.* Genet Epidemiol. 2014 Jul;38(5):381-8.

Original articles:

(1) Heid IM*, Jackson AU*, Randall JC*, Winkler TW* et al. *Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution.* Nat Genet. 2010 Nov;42(11):949-60. *contributed equally

(2) Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G … Winkler TW … Boehnke M, Stefansson K, North KE, McCarthy MI, Hirschhorn JN, Ingelsson E, Loos RJ. *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* Nat Genet. 2010 Nov;42(11):937-48.

(3) Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F … Winkler TW … Abecasis GR, Stefansson K, Frayling TM, Hirschhorn JN. *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature. 2010 Oct 14;467(7317):832-8.

(4) Dastani Z, Hivert MF, Timpson N, Perry JR, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikäinen LP, Fuchsberger C … Winkler TW … Kathiresan S. *Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals.* PLoS Genet. 2012;8(3):e1002607.

(5) Fritsche LG, Chen W, Schu M, Yaspan BL, Yu Y, … Winkler TW … Allikmets R, Baird PN, Gorin MB, Wang JJ, Klaver CC, Seddon JM, Pericak-Vance MA, Iyengar SK, Yates JR, Swaroop A, Weber BH, Kubo M, Deangelis MM, Léveillard T, Thorsteinsdottir U, Haines JL, Farrer LA, Heid IM, Abecasis GR; AMD Gene Consortium. *Seven new loci associated with age-related macular degeneration.* Nat Genet. 2013 Apr;45(4):433-9, 439e1-2.

(6) Berndt SI, Gustafsson S, Mägi R, Ganna A … Winkler TW … McCarthy MI, Speliotes EK, North KE, Loos RJ, Ingelsson E. *Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture.* Nat Genet. 2013 May;45(5):501-12.

(7) Monda KL, Chen GK, Taylor KC, Palmer C, Edwards TL … Winkler TW … North KE, Haiman CA. *A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry.* Nat Genet. 2013 Jun;45(6):690-6.

(8) Liu CT, Buchkovich ML, Winkler TW, Heid IM; African Ancestry Anthropometry Genetics Consortium; GIANT Consortium, Borecki IB, Fox CS, Mohlke KL, North KE, Adrienne Cupples L. *Multi-ethnic fine-mapping of 14 central adiposity loci.* Hum Mol Genet. 2014 Sep 1;23(17):4738-44.

(9) Hinney A, Albayrak O, Antel J, Volckmar AL, Sims R, Chapman J, Harold D, Gerrish A, Heid IM, Winkler TW, Scherag A, Wiltfang J, Williams J, Hebebrand J; GERAD Consortium; IGAP Consortium; GIANT Consortium. *Genetic variation at the CELF1 (CUGBP, elav-like family member 1 gene) locus is genome-wide associated with Alzheimer's disease and obesity.* Am J Med Genet B Neuropsychiatr Genet. 2014 Jun;165B(4):283-93.

(10) Hoggart CJ … Winkler TW … Kutalik Z. *Novel approach identifies SNPs in SLC2A10 and KCNK9 with evidence for parent-of-origin effect on body mass index.* PLoS Genet. 2014 Jul 31;10(7):e1004508.

(11) Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S … Winkler TW … Frayling TM. *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet. 2014 Nov;46(11):1173-86.

(12) Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH … Winkler TW … Speliotes EK. *Genetic studies of body mass index yield new insights for obesity biology*. Nature. 2015 Feb 12;518(7538):197-206.

Selbstständigkeitserklärung

Ich, Thomas Winkler, geboren am 05.10.1981 in Regensburg, erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Insbesondere habe ich nicht die entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten (Promotionsberater oder andere Personen) in Anspruch genommen.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Regensburg, 22.02.2015 Thomas Winkler

Acknowledgements

First of all, I want to thank Prof. Dr. Iris Heid, Head of the Department of Genetic Epidemiology at University of Regensburg, for motivating me to work on this thesis, for the constant support and most of all for introducing me into the great world of genetic epidemiology and the many possibilities it offers to a mathematician. Many thanks for supervising this work and for giving me the opportunity to participate at national and international conferences.

I would like to thank Univ.-Prof. Dr. Florian Kronenberg, Head of the Division of Genetic Epidemiology Innsbruck at Medical University Innsbruck, for mentoring the thesis and as a consequence thereof, for the reiterated travels to Regensburg. Thank you for many valuable advices and for giving me the opportunity to collaborate with other great scientists at the Division of Genetic Epidemiology Innsbruck.

Next, I would like to thank PD Dr. Jörg Marienhagen from the Department of Nuclear Medicine at University Hospital Regensburg, for mentoring the thesis and for the constant support.

Many thanks to two great collaborators: Prof. Dr. Ruth Loos from The Charles Bronfman Institute for Personalized Medicine at Mount Sinai, New York, for supporting me at all times and for hosting me three months at the Big Apple, and Ass.-Prof. Dr. Zoltan Kutalik from the University of Lausanne, for providing constant statistical supervision whenever I ended up in a statistical dead-end.

Also, I would like to thank my colleagues at the Department of Genetic Epidemiology in Regensburg – all of which being great! In particular, many thanks to Mathias Gorski and Dr. Matthias Olden for valuable computer science advices and for great discussions, Dr. Wilmar Igl for statistical advice and PD Dr. Klaus Stark for advice in biology (and in everything else).

Finally, I am grateful to my fiancée and future wife Katrin, who has never failed with encouragement, caring and advice.

Curriculum Vitae

Name:	Thomas Winkler
Date, Place of Birth:	October 5th, 1981 in Regensburg, Germany

Education:

06/2002	Abitur at the Goethe-Gymnasium, Regensburg
09/2003 - 06/2004	Civil Engineering, University of Applied Sciences Regensburg
09/2004 - 02/2009	Diploma in Mathematics, University of Applied Sciences Regensburg

Professional Experience:

08/2002 – 05/2003	Alternative community service at Hospital Barmherzige Brüder, Regensburg
09/2005 - 01/2006	Practical semester at Micron Electronic Devices AG, Neutraubling
03/2007 - 07/2007	Practical semester at BMW AG, Munich
03/2007 - 07/2007	Diploma project at Osram Opto Semiconductors GmbH, Regensburg
04/2009 – present	Scientific staff at the Department of Genetic Epidemiology, Institute of Epidemiology and Preventive Medicine, University of Regensburg
09/2012 – 11/2012	Internship at the Charles Bronfman Institute for Personalized Medicine at Mount Sinai School of Medicine, New York, USA