ENHANCING THE DETECTION OF COMPLEX DISEASE LOCI BY NEW APPROACHES TO IMPUTATION

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

der
Fakultät für Medizin
der Universität Regensburg

vorgelegt von
Mathias Gorski
aus
Neumarkt i. d. OPf

im Jahr
2015

Dekan: Prof. Dr. Dr. Torsten E. Reichert

Betreuer: *Prof. Dr. Iris Heid*

Tag der mündlichen Prüfung:

Contents

1 Introduction to genetic epidemiology

1.1 Genetic Epidemiology

The Human Genome Project decoded the complete human genome in 2003 [1] and showed that more than 99% of the human genetic code is identical across all humans. The remaining less than 1% are variants that account for our individuality. These variants do not only affect our weight, personality or the color of our eyes. They also point towards the development of complex diseases. But the interdependencies between the genetic code and the development of complex diseases are not yet understood in full detail. It is the overarching aim in Genetic Epidemiology to investigate the influence of variants on complex disease.

One example for a complex disease is the Age-related Macular Degeneration (AMD). It is a leading cause of vision loss with great impairment in daily life. AMD is the most common cause of blindness in the elderly in developed countries and it was shown, that AMD has a substantial heritable component [2]. Another example for a complex disease with a considerable heritable component is Chronic Kidney Disease (CKD) [3]. CKD is a progressive loss of kidney function: The kidneys can't clear waste materials from the body and maintain a normal balance of fluids in the body any more. Patients with a permanent decreased kidney filtration rate have a high risk to require dialysis and have an increased risk of death [4, 5].

In summary these examples of the genetic influence on eye and kidney disease underpin the importance to investigate the role of the genetic variations on the development of complex disease to improve prevention, diagnosis and therapy.

1.2 Genome wide association analyses and genome wide association meta-analyses

The aim of **genome wide association analyses** (**GWAS**) and **genome wide association meta-analyses** (**GWAMAs**) is to identify associations between genetic variants and a *phenotype*. A phenotype can be an outcome of interest (for example kidney filtration rate) or a disease (for example AMD). In this chapter I introduce how associations between a phenotype and variants in the human genome are identified.

1.2.1 The genetic code

The human genome is the complete set of genetic information of a subject encoded in the Deoxyribonucleic Acid (DNA) by a diploid, two-stranded set of 23 chromosomes, including 22 autosomes and one pair of gonosomes. Genetic information is encoded as nucleotides Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) – called alleles. The alleles are organized in 2 complementary

strands per chromosome, where Adenine binds with Thymine and Cytosine binds with Guanine. Each strand begins with a 5'-hydroxyl group (5' start) and ends with the 3'-hydroxyl group (3' end).

More than 99% of the human genome is identical in all humans. In the remaining less than 1% it is possible that single nucleotides are polymorph (Single Nucleotide Polymorphisms - SNPs). SNPs are the most common genetic variants and usually consist of two nucleotides (bi-allelic SNPs). Other types of variants are insertions (additional alleles in an individual's variants) and deletions (alleles are deleted from an individual's variants). In this work, I categorize variants with a minor allele frequency (MAF) higher than 20% as *very common*, variants with a MAF between 5% and 20% as *common*, variants with a MAF between 1% and 5% as *less common* and variants with MAF below 1% as *rare*. Half of the variants are inherited by the father, the other half from the mother. The sequence of adjacent alleles on either chromosome is called *haplotype*. **Figure 1** shows a short segment of the paternal and maternal genetic code in one subject, including the coding and complementary strand. The upper (in this case paternal) haplotype is *ACGTACGTA* and the lower (in this case maternal) haplotype is *ATGTGAACA*. The genotype of a variant in a subject is the unordered pair of alleles. The genotype of the exemplified SNP is C/T, the genotype of the insertion is T/TGA and the genotype of the deletion is CGT/C as shown in **Figure 1**.

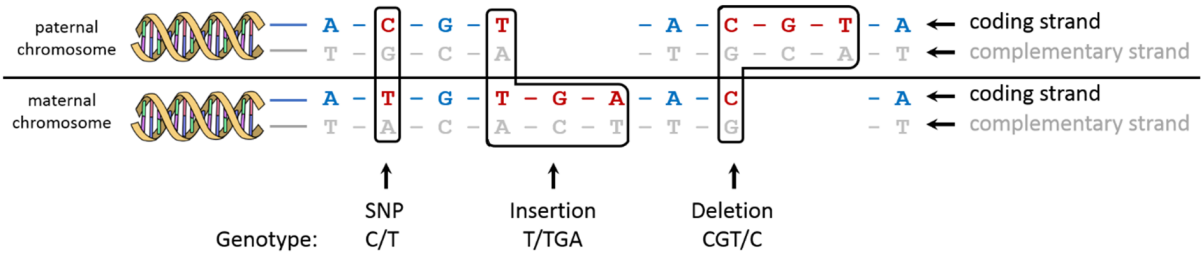

Figure 1. ***Schematic presentation of a heterozygote SNP, insertion and deletion****. Shown are the coding and complementary strand for the three variations on a paternal and maternal chromosome.*

In the beginning of the 1990s it was possible to determine up to several hundreds of genetic variants in the human genome with Polymerase Chain Reaction based assays [6]. Since then technical advances have been made [7] and finally in 2005 the first genome wide association study was conducted [8]. Today chip-based microarray technologies allow the assessment of several hundred thousand up to millions of variants in thousands of individuals with a single high-throughput genotyping chip [9]. Next Generation Sequencing techniques [10] allow the determination of the precise order of nucleotides in a person [11]. Whereas today only a small number of persons is fully sequenced, decreasing costs for sequencing technologies will allow the evaluation of the full genome in large populations in the near future [12].

1.2.2 Testing association between genotype and phenotype

As genotyping arrays could only detect a small number of variants per subjects, candidate gene approaches were commonly used in the 1990's. Prior knowledge about the influence of genes on the human biology were the basis in this hypothesis driven approach. Since technological advances lowered costs for genome wide scans of the human genome enormously, hypothesis free, genome wide scans became possible [8]. Numerous loci for complex diseases have been identified on a wide range of phenotypes on health and disease until today [13].

1.2.3 Genome wide association studies

It is the aim of **GWAS** to identify genetic loci, which are associated with a phenotype. The genotypes of cross sectional study subjects are tested with quantitative or binary phenotypes.

The association of the study genotypes with a **quantitative** phenotype (for example kidney function) is tested with a linear regression. This linear regression tests if the genetic effect on the quantitative phenotype is significantly different from zero in the model

$$Y = \alpha + \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2 + \cdots + \beta_n \hat{X}_n + \varepsilon, \varepsilon \sim N(0, \sigma^2),$$

where Y is the quantitative phenotype, α is the intercept of the regression line, β_1 is the effect of the genotypes $\hat{X}_1$ and β_2 , … , β_n are the effects of the covariates $\hat{X}_2$, … , $\hat{X}_n$ and ε is the σ^2 distributed error. The test of significance tests the null hypothesis H_0: $\beta_1 = 0$ from being different from the alternative hypothesis H_a: $\beta_1 \neq 0$ in applying the test statistic

$$T = \frac{\beta_1}{SE\ (\beta_1)}.$$

Assuming the null hypothesis, the test statistic *T* follows a t distribution with *n-2* degrees of freedom, i.e., $T \sim (n-2)|H0$. The t test yields a SNP association p-value.

The association of study genotypes with a **binary** phenotype (for example AMD) is tested with a logistic regression. Y = 1 means that a person is affected and Y = 0 means that a person is unaffected. Then the regression model reads as

$$logit(Y) = ln\frac{\mathrm{P(Y=1)}}{P(Y=0)} = \alpha + \beta_1 \hat{X}_1 + \beta_2 \hat{X}_2 + \cdots + \beta_n \hat{X}_n + \varepsilon, \varepsilon \sim N(0, \sigma^2),$$

where α is the study-specific intercept and β_1 is the genotype log odds ratio (OR). P(Y=1) and P(Y=0) are the probabilities for a person being affected or unaffected. The logit is the logarithm of an odds: $Logit = \ln\left(\frac{P}{1-P}\right), where\ P = P(Y=1)$. The parameter β_1 specifies the effect of the genotype ($\hat{X}_1$) and the model can be adjusted for a series of covariates ($\hat{X}_2$, … , $\hat{X}_n$) and ε is the σ^2 distributed error.

The test of significance tests the null hypothesis H_0: β = 0 from being different from the alternative hypothesis H_a: β ≠ 0. The Wald score test statistic

$$W = \frac{\beta_1}{SE\ (\beta_1)}$$

is evaluated relative to a standard normal distribution and yields a variant association p-value. The Firth bias-corrected likelihood ratio test is an extension to this Wald test, which solves the "phenomenon of separation" [14], which occurs if too many cells in the contingency (2x2) table have low counts, which is the case if a variant is rare. Commonly, variables like gender or age are covariates in the analysis to reduce the phenotypic variance or to account for potential confounders in the analysis.

In genome wide association analyses, several millions of variants are tested, but it is assumed that about 1 Mio independent tests are conducted to account for the linkage disequilibrium among genetic variants. Linkage disequilibrium (LD) is a measure if two genotypes at different positions occur more often than expected by chance. D prime (D') is a measure if two alleles occur on the same haplotype more often than expected by chance. To avoid huge numbers of false positive findings, genome wide results are corrected for multiple testing. In a genome wide setting the human genome wide significance level is set at $1x10^6/0.05=5x10^{-8}$ to correct to correct for 1 million independent tests [15]. Genomic inflation can systematically increase Type I and Type II errors [16]. It occurs if study samples have cryptic relatedness or population stratification (different ancestries causing systematically different allele frequencies between subpopulations). Genomic inflation can be visualized with a Quantile-Quantile-Plot (QQ-Plot), comparing the distribution of the test statistics with the expected one. It can be quantified by comparing the median of the chi-squared test statistics with the expected median. Thus the genomic inflation factor λ is defined as the ratio of the median of the empirically observed distribution of the test statistic divided by the expected median. The genome wide association results are then controlled for genomic inflation in dividing the observed test statistics with the observed genomic inflation factor λ.

1.2.4 Genome wide association meta-analysis

Although genome wide association studies have been successful in detecting genetic loci for complex diseases, single GWAS have only limited power to detect genome wide significant loci for complex diseases. Genome wide association meta-analysis (GWAMAs) increase the power to detect additional genetic loci associated with complex diseases in pooling the results of several GWAS into one big meta-analysis. Furthermore, the workflow of GWAMAs in consortia is designed to provide a practical way to avoid that study analysts must share their study genotypes and phenotypes with other study partners and to eliminate the complex integration of genotypic and phenotypic data from different studies:

First, study partners for the meta-analysis are identified. Second, an analysis plan is developed, where the participating studies get detailed instructions on how to perform the GWAS. Third, the study analysts perform the GWAS in the study centers and estimate the β_i's and SE's of all individual studies. Fourth, the study specific β_i's and SE'S are uploaded onto a central server, which is also accessible to the analyst from the consortium. Fifth, this consortium analyst pools the study specific β_i's in one meta-analysis. This workflow avoids that genotypes of study individuals need to be made available by the single studies, but allows to test the association of a phenotype with the genotype of all subjects from the participating studies in a meta-analysis [17-20].

To test for associations, all studies provide the effect estimates β_i and standard errors SE_i of all *i* variants in the study data. The pooled summary statistics are then computed with the inverse variance weighted method, assuming homogeneity of study effects across all input studies [21]. The z-statistic per variant yields the association p-value and is obtained from the association effect estimate β and standard error SE per variant across all studies:

$$\beta = \frac{\sum_i b_i w_i}{\sum_i w_i}$$

$$SE = \sqrt{\frac{1}{\sum_i w_i}}$$

$$w_i = \frac{1}{SE_i^2}$$

$$Z = \frac{\beta}{SE} \sim N(0,1) | H_0.$$

1.3 Genotype imputation

Genotyping chips directly depict a fraction of all variants in the human genome. Studies in a meta-analysis are typically genotyped on different genotyping chips. As a consequence not all genotypes are known in all subjects. Maximal power per variant is achieved when the genotypes are known in all subjects. It is thus necessary to have a consensus set of genotypes in all studies. Genotype imputation infers variants not detected by the genotyping chip to a consensus set of genotypes, to harmonize all study specific sets of variants. The 2-step-approach of imputation [22] separates the **phase estimation** from the **genotype imputation** step.

1.3.1 Phasing

The first step in the 2-step-approach of imputation is the study haplotype estimation (phasing). It is the statistical inference of haplotypes from study genotypes (see **Figure 2**). Phasing is independent from a reference panel. Haplotype reconstruction methods were developed to efficiently estimate haplotype probabilities or the most likely haplotypes per subject [23, 24].

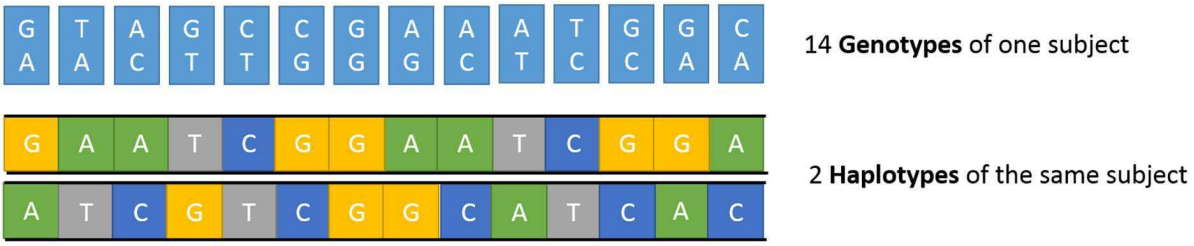

*Figure 2. **Inference of study haplotypes from study genotypes (Phase estimation)**. Shown is a short segment of 14 adjacent and polymorphic variants in one subject. Shown are these unordered pairs of alleles in the first row. Haplotype estimation reconstructs the set of variants on a single chromosome (haplotypes), as shown in the second row.*

All haplotype reconstruction methods have the common rational that subjects share stretches of their genetic code, inherited from common ancestors. These stretches are longer in a closely related pair of subjects (father-child) and shorter in distantly related or unrelated subjects. To infer haplotypes, phasing programs use the amount of LD between variants to infer shared stretches among subjects. This is can be accomplished by Clark's algorithm, an expectation-maximization (EM) algorithm, coalescence-based algorithm or by Hidden Markov Model approaches [25]. Haplotype reconstruction uses external data sets providing the genetic distance and recombination rates between two variants. The recombination rate is the probability of a cross-over between two variants.

Commonly used programs for phase estimation are PHASE [26], FastPHASE [27], Beagle [23], minimac [22], HapiUR [24] or ShapeIT [28]. Phase estimation is computationally challenging, as it needs to be done for full chromosomes, since this yields the best results. For example phasing of about 20,000 variants (the size of chromosome 19) in about 50,000 subjects with ShapeIT is completed on a 12 core server in about three days. Phasing the 22 autosomes would require ~ 2 month (without parallelization).

1.3.2 Genotype imputation

The second step in the 2-step-approach of imputation uses the estimated study haplotypes from the first step to estimate sites with unknown genotype in the study data with the help of external reference haplotypes (including a much denser set of typed variants). Based on the intuition that short stretches of haplotypes are inherited together, genotype imputation infers variants unknown in the study data but known in the reference panel by taking the LD between neighboring variations into account (see

Figure 3). The true haplotypes underlying the observed genotype data are assumed to be imperfect mosaics of the reference haplotypes [23, 27, 29-31].

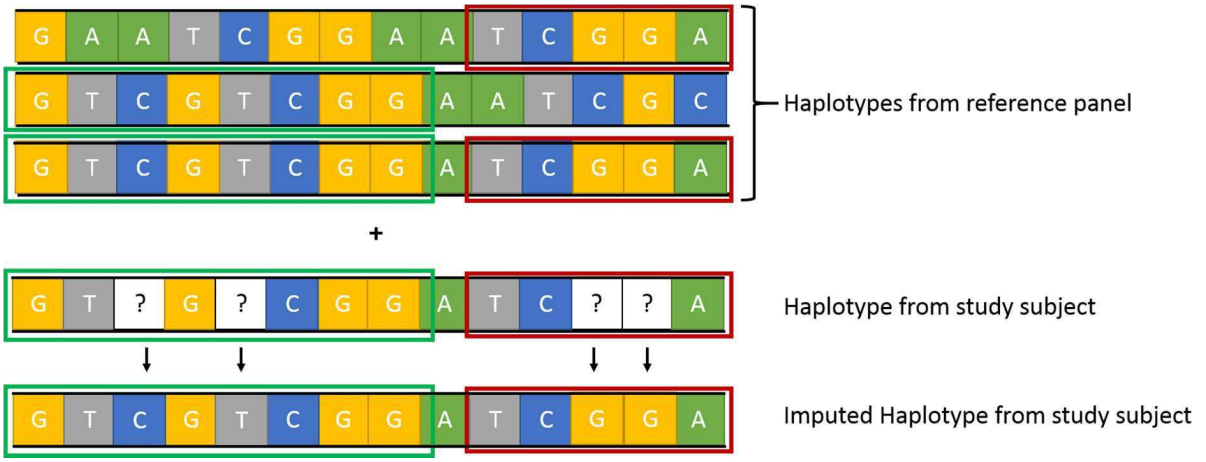

Figure 3. ***Estimation of untyped variants in the study with reference haplotypes (Genotype Imputation).*** *Shown is how 3 haplotypes from the reference panel are used to infer 4 untyped sites in a haplotype from a study subject. The first two untyped sites in the study subject are derived from identical segments in the second and third reference haplotype (green boxes). The third and fourth untyped sites in the study subject are derived from identical segments in the first and second reference haplotype (red boxes).*

Known genotypes are represented as the discrete values 0, 1 and 2, indicating the number of coded alleles in this variant. Imputed genotypes are represented as probabilities of the coded allele: e. g. (0.25/ 0.75) indicates that there is a 25%/ 75% probability of homozygosity of the coded allele or for heterozygosity in the variant, respectively (**Figure 4a**). A dosage of 1.25 indicates that there is a 62.5% probability (1.25*100/2) that there are 2 copies of the coded allele in the variant (dosage, **Figure 4b**). If it is necessary to work with discrete genotypes, it is possible to round the dosages to discrete best-guess genotypes, losing the uncertainty (**Figure 4c**).

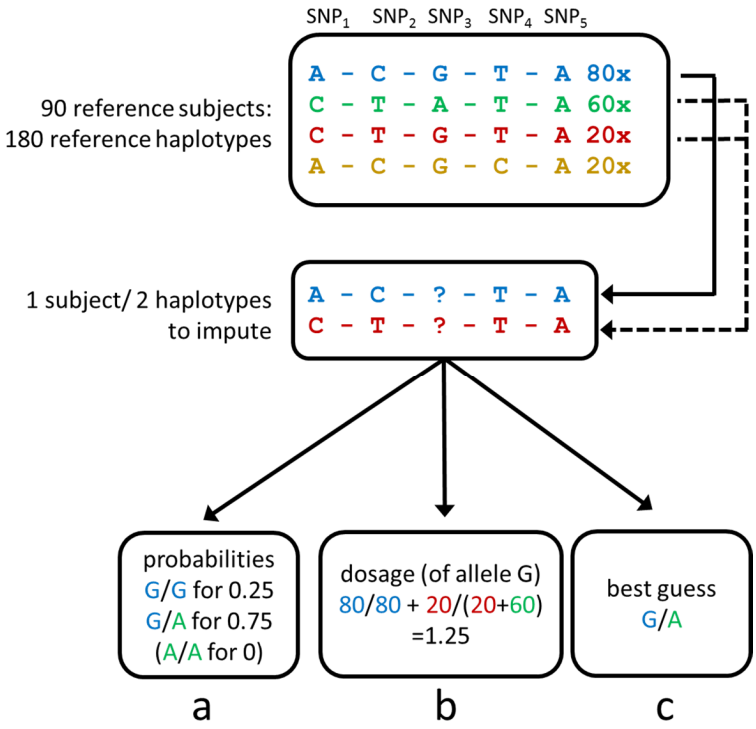

Figure 4. ***Schematic representation of genotype imputation of SNP₃ in the 2 haplotypes to impute****. The most common reference haplotype (blue) is an unambiguous match for the upper haplotype to impute. So all 80 reference haplotypes indicate the allele in the upper haplotype to impute is the coded allele with a probability of 1. The second and third reference haplotype match the lower haplotype to impute. So 20 of the 80 haplotypes indicate that the missing allele in the lower haplotype to impute is the coded allele with a probability of 0.25. It is shown how imputed genotypes can be represented as (a) probabilities, (b) dosages and as (c) a best guess discrete genotype. Figure is adapted from [33].*

1.3.3 Imputation quality

It is possible to evaluate the accuracy of an imputed genotype with the imputation quality metric. It is a measure which relates the imputed allele counts across all subjects in one imputed variant to the expected allele counts of a virtually perfectly imputed variant with identical allele frequency. It is defined as the variance in the imputed genotypes divided by what would be expected if the genotype would be observed without error following a binomial distribution, given the allele frequency

determined from the imputed variants. Let *d* be the imputed allele frequency in all study participants and let *p= mean(d)/2* be the mean allele frequency of the coded allele across all study participants, then the *estimated r^2 with true genotype* is defined as

$$E(r^2\ with\ true\ genotype) = \frac{Var(d)}{2*p*(1-p)}.$$

Typical measures are the *RSQ*-metric from *minimac* and the *Info*-score from *ImputeV2*. To analyze variants which are imputed at high accuracy, variants with an imputation quality smaller than 0.4 are typically excluded prior to meta-analysis [32].

1.3.4 Reference panels

Genotype imputation programs use external reference data sets to estimate the allele count in study participants at untyped sites. These external reference panels used for genotype imputation can be classified into low density reference data covering primarily common and very common genetic variants and high density reference data additionally including rare and less common variants in human beings.

The aim of the International HapMap Project [33] was to chart common and very common (with MAF ≥ 5%) human genetic variants across diverse populations. Genotypes, The LD between variants or reference panels for genotype imputation were released to other researchers for their studies on human health and disease from 2002 to 2010. With the releases of reference panels for genotype imputation the International HapMap Project laid the foundation for many successful GWAS and GWAMAs based on HapMap imputed genotypes [13]. HapMap reference panels typically provide 60 subjects of European descent, each genotyped and phased at about 3 Million common and very common variants [**see Web Resources]**.

Next generation sequencing techniques noticeably reduced the costs of sequencing human genomes. The 1000 Genomes Project [34] uses these new technologies to assemble a new generation of reference data sets since 2010 [35]. Thus there has been a constant update of reference panels with ever increasing number of subjects and variants over the last five years. These new reference data sets include common as well as rare variations (MAF < 1%) on autosomes, gonosomes and in the mitochondrion. For example the 1000 Genomes Project Phase I version 3 reference data set consists of about 39.7 Million variants in 379 European, 246 African, 286 Asian and 181 subjects of Mixed American populations [**see Web Resources**]. It is now a frequently used reference panel for genotype imputation.

1.3.5 Challenges of genotype imputation for genome wide association analyses

1.3.5.1 Challenges when imputing study data for genome wide association analyses

Study analysts must overcome several challenges when imputing study genotypes with **one reference panel**: Before imputation study genotypes are to be scrutinized and cleaned for variants or subjects with inferior call rate and for violation of Hardy-Weinberg equilibrium [36]. Prior to imputation, the build of the genotypes must match those in the reference panel. A build specifies the chromosome and position of a variant on the chromosome. But it is possible that variants are erroneously assigned to a wrong position. As advances in genotyping and sequencing technologies correct these dislocated variants, relocation of variants or the identification and merging of erroneously duplicated variants is possible. There are big leaps in builds [37] (e.g. from NCBI [38] build 36.3 to GRCh build 37.1), smaller changes (e.g. from build 37.1 to 37.2) as well as different chip manufacturer's annotation files. The manufacturer's annotation files often include chip-specific SNPs that are not yet contained in the official build information. If the study data is on an older build compared to the reference panel, the positions of the genotyped variants in the study data must be adjusted accordingly before imputation.

Phase estimation, genotype imputation and association analysis are computational demanding, requiring several weeks on a standard 8 core cluster. As imputed genotypes include millions of variants, up to 1 terabyte data storage is needed to store the imputed genotypes and association results. GWAS results are usually transferred and stored on file transfer server for meta-analysis. Thus also a fast internet connection is needed to transfer the data quickly to the server.

1.3.5.2 Challenges when imputing study data multiple times for genome wide association analyses

When study data needs to be **re-imputed with different reference panels**, the build of the study data can be harmonized to the build of only one reference panel. Theoretically the two-step approach of imputation allows for updating the imputation for a new reference panel without repeating the tedious phasing step. However, this advantage does no longer hold, when the build of the study data was adjusted to the build of the first reference panel and differs from the build of the second reference panel. In this case, the current approach is to harmonize the study data annotation with each reference panel (pre-phasing lift-over), requiring re-phasing and re-imputing. As the phasing is very time-consuming, study analysts are often overwhelmed with regular requests by GWAMA coordinators to lift, re-phase, and re-impute their study data, in order to optimize their study's contribution to the meta-analysis.

1.4 Application of imputed genotypes to complex disease

Imputed genotypes are analyzed in consortia in meta-analyses or mega-analyses to identify genetic loci associated with complex disease. While the genetic map of several phenotypes is subject to investigation in consortia [13], I focus in my work on meta-analyses on kidney function and on mega-analysis of AMD.

1.4.1 Meta-analysis of continuous trait on kidney function

Loss of kidney function can cause chronic kidney disease. Chronic kidney disease affects more than 10% of the adult population in America and Europe and can result in end stage renal disease, necessitating dialysis or a kidney transplant. It is associated with high morbidity and is also a major risk factor for myocardial infarction and stroke. Risk factors for CKD are obesity, cardiovascular disease, diabetes or hypertension. CKD is a complex disease with a strong genetic background [3, 39-41]. Stages of CKD are classified by the amount of glomerular filtration rate (eGFRcrea). This glomerular filtration rate is a measure for kidney function.

In the past years more than 50 studies have contributed genome wide association analysis results on kidney function for **meta-analyses** in the CKDGen consortium, one of the largest consortia focusing on kidney disease. These studies contributed HapMap imputed genotypes of a varying number of subjects, ranging from several hundreds to more than 20,000 subjects per study. Overall, more than 50 genetic loci associated with kidney function could be identified by several GWA consortia, including the CKDGen consortium [18, 20, 42-45].

1.4.2 Mega-analysis of binary trait on eye disease

The complex disease age-related macular degeneration (AMD) is the leading cause of blindness in the elderly. It causes irreversible loss of central vision, affecting more than 10% of subjects over age 80. AMD can affect one or both eyes. The heritability of AMD was shown to vary between 46% and 70% [2]. AMD is diagnosed by ophthalmologic inspection of the eye background. AMD is classified as wet AMD (choroidal neovascularization, CNV, when accompanied by angiogenesis) or as dry AMD (geographic atrophy, GA, when angiogenesis is absent) or both.

The genetic map of AMD is investigated for several years. First susceptibility loci for AMD were discovered in 2005 [8]. Today, analyses of common variation have uncovered 21 risk loci for AMD [17, 46-51] with a recent update to 34 risk loci [52].

1.5 Scientific gaps, aims and structure of this thesis

Hypothesis driven candidate gene approaches identified susceptibility loci for both AMD and kidney function. Hypothesis-free genome-wide association meta-analyses of studies imputed with low density reference panels charted genetic regions associated with complex diseases. Several common variants associated with genome-wide significance with kidney function and AMD were successfully identified by meta-analyses of HapMap imputed genotypes [45, 52]. Advances in sequencing technologies allow for the detection of more variants in the human genome. As a consequence a novel generation of reference panels for genotype imputation became available. These reference panels are used for genotype imputation in study data. But analyses based on these inferred genotypes in a large number of subjects are still missing. Also no comprehensive evaluation has been conducted to evaluate the gain of using the high density reference panels for genotype imputation compared to using low density reference panels. Genotype imputation and association analysis is either possible in meta-analyzing study data imputed per study or by mega-analyzing data pooled across several studies (mega-analysis of IPD). The gain in utilizing the mega-analysis approach compared to the meta-analysis approach has also not been investigated in a large number of subjects.

1.5.1 Scientific gaps

There are several gaps, which I address in my work:

First, meta-analysis of several studies imputed separately into low density reference panels (for example HapMap) allow the detection of common variants associated with complex disease. The latest released 1000 Genomes reference panels consist of more variants and more reference subjects compared to the HapMap reference panel. **But it is yet unclear how much we can gain in using high density reference panels for genotype imputation in contrast to using low density reference panels in a consortia setting.**

Second, constant updates of reference panels force study analysts to re-phase and re-impute the study data with each new release for such meta-analyses. Furthermore the phasing, imputation and analysis of genetic data is a complex and computational expensive task. **But there are no approaches to facilitate the time consuming imputation process and to help study analysts to quickly impute their study genotypes.**

Third, although meta-analyses have been very successful in identifying susceptibility loci for complex diseases, IPD-analyses promise to further alter our understanding of the genetics of complex diseases. **The question has not yet been addressed how much we gain with an IPD-analysis of a genome wide data set in contrast to a meta-analysis.**

1.5.2 Objectives and aims

To address the described research gaps, it **is the main objective of my work** to enhance the detection of disease loci by new approaches to imputation. This involves the following **specific aims**:

My first specific aim is to investigate how much we gain in using high density reference panels for imputation per study in contrast to using low density reference panels for imputation per study in a typical meta-analysis scenario: Specifically I am interested if I can identify additional genome wide significant associations with a complex trait, which might be due to the better fine mapping in the 1000 Genomes reference panels. It is also of interest if imputation with 1000 Genomes reference data produces generally better imputed variants and if it increases the power to detect common, less common and rare variants. I am using the data from the CKDGen consortium to achieve this aim, because I can compare meta-analysis results in a typical meta-analysis scenario on kidney function.

My second specific aim is to analyze how much gain if we would replace the commonly used meta-analysis approach in consortia with conducting a mega-analysis approach: to gather the data at participant level (generating IPD-data) and to mega-impute and mega-analyze it. I use the data of the IAMDGC on AMD as it is an optimal data set to examine the differences between meta-analysis and mega-analysis on variants with a wide spectrum of allele frequencies and effect sizes in a sample size realistic for consortia work.

My third specific aim is to provide computational methods and software to facilitate phase estimation, genotype imputation and association analysis. The constantly released reference panels result in the need to constantly re-phase and re-impute study data. I approach this problem by evaluating if the computational demanding task of re-phasing can be omitted. Phasing and imputation of IPD (mega-imputation) is therefore increasingly demanding with increasing number of subjects in the analysis. Thus I further aim at solving this challenge by highly parallelizing phasing and imputation and by exemplifying the needed time and computing facilities for a mega-imputation in a realistic scenario.

1.5.3 Outline of this work

In **chapter 2** of this thesis I present the investigation of the gain in GWAMAs, comparing imputation with high density reference data with low density reference data. After introducing the materials and methods needed in **chapter 2.1**, I contrast two meta-analysis results from the CKDGen consortium in **chapter 2.2**: the first one is a meta-analysis of HapMap imputed genotypes, the second one is the meta-analysis of 1000 Genomes imputed genotypes. This is exemplified on a search for genetic loci for kidney function. In **chapter 2.3** I describe the software *PhaseLift*, which facilitates genotype imputation: What the software does is harmonizing variant pairs on the haplotype level instead of on the genotype level. By this, a re-phasing can be omitted. I use data from the CKDGen consortium for all analyses in **chapter 2.**

In **chapter 3,** I focus on how much can be gained when applying the mega-analysis approach in contrast to the meta-analysis approach: After introducing the materials and methods needed in **chapter 3.1,** I evaluate the gain in imputing genotypes jointly compared to imputing them separately in **chapter 3.2**. Then I evaluate the influence of analyzing genotypes jointly compared to analyzing them separately by study. This is exemplified on a search for genetic loci for AMD. I **chapter 3.3** I discuss the need to parallelize genotype imputation, which is necessary to mega-impute an IPD set in reasonable time. I show how this parallelization can be achieved and how processing time can be optimized. I use the IAMDGC data set for all analyses in **chapter 3**.

Finally in **chapter 4**, I summarize the key findings of this thesis. I illustrate relevance, strength and limitation of my work and discuss my findings in the context of the current literature.

2 On the gain of imputing GWAS with high density reference data compared to low density reference data for meta-analyses

This chapter has two overarching aims: The first aim (**chapter 2.2**) is to evaluate the gain in using high density reference panels in contrast to using low density reference data for genotype imputation in large scale meta-analyses. For this investigation I utilize the data from the CKDGen consortium, which consists of up to 133,817 subjects from 56 studies and aims to identify genetic risk loci for kidney function by meta-analyses. The search for risk loci for kidney function in large scale meta-analyses is subject to research in consortia for several years [18, 20, 43-45] . The second aim (**chapter 2.3**) is to introduce a new software, which supports study analysts in imputing their study genotypes for GWAS in consortia in a fast and efficient way anytime a novel reference panel is released. For this investigation I utilize the data from 1,644 subjects in the population based KORA-F3 study. The methods and materials needed for these analyses are subject in **chapter 2.1**.

2.1 Methods and Material

In this section I introduce the phenotype for my comparison and the available low density and high density reference panels. Furthermore I contrast the studies from the CKDGen consortium used for GWAMAs of studies imputed with low density reference panels with those imputed with high density reference panels. I illustrate how differences between meta-analyses are quantified: I investigate the power to detect a genetic locus associated with a quantitative trait and I calculate the proportion of phenotypic variance explained.

2.1.1 Phenotype

Kidney function servers me as example for comparing results from meta-analysis of study data imputed with high density reference panels in contrast to meta-analysis results imputed with low density reference panels. Kidney function in humans is difficult to assess, because it strongly depends on the diet and calorie consumption of a person. Nevertheless it can be measured with serum creatinine, a waste molecule from the muscle metabolism. The kidneys filter most of the creatinine and secrete it by the urine. Serum creatinine is drawn from whole blood and is the basis of the estimated glomerular filtration rate eGFRcrea per subject, which is used in genetic association studies to quantify kidney function. As this quantitative parameter strongly dependents on age, sex and ethnicity of the subjects, it is standardized prior to analysis. Serum creatinine is usually standardized by the MDRD [53] or CKDEPI [54] formula to obtain the estimated glomerular filtration rate (eGFR). With the MDRD formula it is calculated as $186.3*(\text{serum creatinine [mg/dl]})^{-1.154} * \text{age}^{-0.203} * (0.742 \text{ if subject is female})$. Values below 15 ml/min/1.73 m^2 are set to 15 ml/min/1.73 m^2 and values above 200 ml/min/1.73 m^2 are set

to 200 ml/min/1.73 m^2 (winsorization). The estimated glomerular filtration rate based on creatinine measurement (eGFRcrea) is then a value between 15 and 200 ml/min/1.73 m^2.

2.1.2 Low density and high density reference panels

Let us define reference panels for genetic variants as low or high density reference panels. Low density reference panels have emerged from the International HapMap consortium [33] and consist of mostly 60 subjects and several million, mostly common SNPs (MAF > 5%). I exemplify the characteristics of low density reference panels on the example of the HapMap Phase II release 22 reference panel (see **Web Resources**) , which has been commonly used for recent GWA work [18, 20, 42-45] .

This reference panel consists of 60 subjects of European descent genotyped at 2,543,230 non-monomorphic bi-allelic SNPs on the 22 autosomes. The vast majority of SNPs is common, only 4.23% of all SNPs have a MAF below 1% in this reference panel (see **Table 1**). I will refer to the meta-analysis of studies imputed with low density reference panels as *HapMap meta-analysis* in the following.

*Table 1. **Summary of all autosomal SNPs in Hapmap Phase II release 22 reference panel.***

		Quantiles of MAF					Proportion of SNPs	Proportion of SNPs
Chr	#SNPs	min	25%	50%	75%	max	with MAF<5%	with MAF<1%
1	193,512	0.008	0.08	0.20	0.34	0.50	0.16	0.05
2	220,798	0.008	0.09	0.21	0.35	0.50	0.14	0.04
3	174,313	0.008	0.09	0.21	0.35	0.50	0.14	0.04
4	163,100	0.008	0.08	0.21	0.34	0.50	0.15	0.04
5	168,118	0.008	0.09	0.21	0.34	0.50	0.14	0.04
6	182,367	0.008	0.08	0.20	0.34	0.50	0.15	0.04
7	143,181	0.008	0.09	0.21	0.35	0.50	0.15	0.04
8	147,460	0.008	0.09	0.21	0.35	0.50	0.14	0.04
9	122,043	0.008	0.09	0.20	0.34	0.50	0.14	0.04
10	138,332	0.008	0.08	0.19	0.34	0.50	0.15	0.04
11	130,060	0.008	0.09	0.20	0.34	0.50	0.15	0.04
12	124,796	0.008	0.08	0.20	0.35	0.50	0.16	0.05
13	104,124	0.008	0.08	0.19	0.34	0.50	0.16	0.04
14	83,923	0.008	0.08	0.20	0.34	0.50	0.15	0.04
15	72,314	0.008	0.08	0.20	0.35	0.50	0.15	0.04
16	71,505	0.008	0.09	0.21	0.35	0.50	0.15	0.05
17	58,395	0.008	0.09	0.21	0.36	0.50	0.13	0.04
18	76,784	0.008	0.08	0.21	0.34	0.50	0.15	0.04
19	37,027	0.008	0.09	0.21	0.34	0.50	0.15	0.04
20	63,417	0.008	0.08	0.20	0.34	0.50	0.16	0.05
21	33,855	0.008	0.09	0.21	0.35	0.50	0.13	0.03
22	33,806	0.008	0.08	0.19	0.33	0.50	0.17	0.05
TOTAL	2,543,230						0.15	0.04

MAF= minor allele frequency.

High density reference panels have emerged from the 1000 Genomes project [35, 55]. They include more subjects and a higher number of variants compared to low density reference panels. Besides SNPs they also contain structural variants (insertions and deletions). I exemplify characteristics

of high density reference panels with the 1000 Genomes Phase I version 3 reference panel, which has been recently used for advanced GWA work. It consists of 1,092 subjects genotyped at 30,072,738 non-monomorphic variants. Of these, 4.77% are insertions or deletions. As shown in **Table 2,** the number of variants with a MAF below 1% is much higher than in the low density reference panel (57.32% vs. 4.23%).

Table 2. ***Summary of all autosomal variants in the 1000 Genomes Phase I V3 reference panel.***

			Quantiles of MAF					Proportion of variants with	Proportion of variants with
Chr	# Variants	Proportion of SNPs	min	25%	50%	75%	max	MAF<5%	MAF<1%
1	2,355,440	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.76	0.58
2	2,583,636	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.77	0.58
3	2,168,045	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.76	0.58
4	2,168,864	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.75	0.57
5	1,990,636	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.77	0.58
6	1,926,214	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.75	0.56
7	1,753,344	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.75	0.57
8	1,715,459	0.96	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.77	0.58
9	1,298,344	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.76	0.57
10	1,486,764	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.75	0.57
11	1,485,625	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.76	0.57
12	1,440,202	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.76	0.57
13	1,086,085	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.75	0.57
14	989,362	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.76	0.57
15	883,908	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.76	0.57
16	942,840	0.96	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.76	0.58
17	818,163	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.76	0.57
18	855,547	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.76	0.57
19	652,223	0.94	$9.00x10^{-4}$	$2.30x10^{-3}$	0.01	0.06	0.50	0.73	0.54
20	668,130	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.04	0.50	0.76	0.57
21	409,633	0.95	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.06	0.50	0.74	0.56
22	394,274	0.97	$9.00x10^{-4}$	$1.80x10^{-3}$	0.01	0.05	0.50	0.74	0.55
Total	30,072,738	0.95						0.76	0.57

MAF= minor allele frequency, Proportion of SNPs is the percentages of Single nucleotide polymorphisms among all variants.

In summary it can be seen, that there are differences between the high and low density reference panels concerning the number of subjects, the number of variants as well as the number of insertions and deletions. The majority of variants in the 1000 Genomes reference panel has a MAF below 1%, underpinning the increased possibility to detect rare variants associated with complex disease with 1000 Genomes imputed data compared to HapMap imputed data. I will refer to the meta-analysis of studies imputed with high density reference panels as *1000 Genomes meta-analysis* in the following.

2.1.3 The CKDGen data

The CKDGen consortium conducts meta-analyses of multiple GWAS on kidney function. Genotype imputation is performed per study, study specific association analyses are conducted and association analyses are pooled in 2 separate meta-analyses: one meta-analysis uses study data imputed with a low density reference panel, whereas the other meta-analysis uses study data imputed with a high density reference panel.

The HapMap meta-analysis is conducted by a 2 stage approach: In the discovery stage association with eGFRcrea is tested in a GWAMA including up to 133,831 European individuals from 50 studies. Analysis is done in all subjects and restricted to subjects free of diabetes. The overview of all studies is shown in **Table 3.** Genome-wide significant loci are identified from the discovery stage results and their lead variants (variants with smallest p-value in the region) are determined. In the second step (replication stage) these lead variants are *de novo* genotyped in up to 42,166 additional individuals from 15 studies. A lead variant is reported as genetic locus, if it is nominally significant or effect direction is consistent in the replication analysis and genome-wide significant in the combined meta-analysis of all discovery and replication stage studies.

The 1000 Genomes meta-analysis is conducted in up to 110,517 individuals from 33 studies. **Table 3** shows the number of subjects and variants analyzed per study. In the 1000 Genomes meta-analysis genome wide significant genetic loci are identified without a replication effort. Each genetic locus, which is identified additional to those from the HapMap meta-analysis, is at least 1 Megabase (MB) distant (down- and upstream) from a lead variant from the HapMap meta-analysis. So no genetic variant is in LD with a locus identified by the HapMap meta-analysis.

2.1.4 Study data to compare meta-analyses of GWAS imputed with HapMap and 1000 Genomes reference panels

Genotype imputation with 1000 Genomes reference panels result in a higher number of variants analyzable. This can be observed in **Table 3** as the number of analyzed variants in the CKDGen 1000 Genomes meta-analysis (median=25,469,713 variants) is about 10 times than the number of analyzed variants per studies in the CKDGen HapMap meta-analysis (median=2,543,887 variants). The maximum number of subjects in a variant can be 133,831 in the CKDGen HapMap meta-analysis and 110,517 in the CKDGen 1000 Genomes meta-analysis. These numbers can vary as not all variants were analyzed in all studies. There are 30 studies overlapping both meta-analyses. In the HapMap meta-analysis, 27 studies are included in the discovery stage and 3 in the replication stage.

Table 3. ***Study overview of the 50 studies from the discovery stage of the meta-analysis based on HapMap imputed data and the 33 studies from the meta-analysis based on 1000 Genomes imputed studies****. Overlapping studies between both meta-analyses are shown in one line. Given are the number of subjects and the number of analyzed variants per study for the 1000 Genomes and the HapMap meta-analysis, respectively.*

	Meta-analysis of studies imputed with			
	HapMap		1000 Genomes	
Study	#subjects	#variants	#subjects	#variants
ADVANCE	2,288	5,920,084	---	---
AGES	3,219	2,408,991	3,219	30,061,893
AMISH	1,211	2,543,013	---	---
ARIC	9,038	2,543,886	9,038	20,939,543
ASPS	850	2,543,887	829	9,332,962
AUSTWIN	9,592	2,373,240	---	---
BLSA	723	2,466,786	---	---
BMES	2,424	2,411,035	2,437	12,826,524
CHS	3,259	2,543,887	---	---
Colaus	---	---	5,409	29,739,752
Croatia-Split	478	2,543,887	---	---
DESIR	721	2,528,169	---	---
EGCUT-370k	863	2,552,470	---	---
EGCUT1	---	---	4,437	16,130,238
EGCUT-OMNI	261	2,584,841	1,018	13,895,165
ERF	2,561	2,543,887	---	---
FamHS	3,838	2,543,887	3,838	29,390,440
FHS	7,782	2,540,126	3,051	24,310,316
Gendian	---	---	450	28,587,788
GENOA	1,064	2,543,843	---	---
HABC	1,663	2,543,887	1,661	29,995,578
HCS	1,235	2,410,284	2,113	12,653,155
HPFS	818	2,570,544	818	29,269,208
$Hypergenes_{cases}$	1,591	2,579,321	---	---
$Hypergenes_{controls}$	1,662	2,579,321	---	---
INCIPE	940	3,071,172	---	---
Ingi-Carlantino	447	2,422,487	412	6,488,909
Ingi-Cilento	821	2,617,075	1,092	29,874,493
Ingi-Fvg	867	2,474,433	848	6,472,705
Ingi-Valborbera	1,636	2,542,680	1,754	27,362,400
IPM I	---	---	440	14,917,463
IPM II	---	---	1,307	25,469,713
JUPITER	8,780	4,513,673	---	---
KORA-F3	1,641	2,515,652	3,095	29,823,380
KORA-F4	1,814	2,543,887	2,936	29,926,398
Korcula	888	2,543,887	---	---
Lifelines	---	---	13,386	8,167,376
Mesa	2,520	2,594,696	2,520	30,065,554
Micros	1,201	2,543,887	1,185	30,061,860
Nesda	1,863	2,539,870	---	---
NHS	786	2,569,327	786	29,203,675
NSPHS	565	2,543,887	563	14,182,979
OGP-Talana	862	2,093,807	---	---
Orcades	704	2,543,887	---	---
Popgen	1,163	2,543,887	---	---
Prosper-Phase	5,237	2,543,887	---	---
Rotterdam I	4,390	2,543,887	4,595	27,647,251
Rotterdam II	1,863	2,543,835	---	---
Sapaldia	1,444	2,588,592	1,444	29,631,136
Ship	3,231	2,748,910	3,210	14,846,831
Ship-Trend	986	3,437,411	986	15,174,445
SORBS	878	2,457,689	---	---
ThreeCity	6,431	2,796,276	6,431	20,548,387
Croatia-Vis	768	2,543,887	---	---
WGHS	21,940	2,543,887	23,186	30,052,421
Young Finns*	2,023	2,543,887	2,023	15,758,340
SUM	**133,831**		**110,517**	

#subjects is the number of subjects analyzed in the study, #variants is the number of variants analyzed in the study; "---" means, that the study was not analyzed in this meta-analysis; *Young Finns contributed 13,681,401 variants, but only variants from the HapMap reference panel were analyzed in the CKDGen HapMap meta-analysis.

2.1.5 Susceptibility loci previously identified in CKDGen HapMap meta-analysis

Overall 53 genetic loci are identified with genome wide significance for eGFRcrea in the CKDGen HapMap meta-analysis (**Table 4**). All lead variants are common. The MAF of these lead variants range from 8% (WDR37) to 48% (NFKB1). All lead variants are changes in a single nucleotide (SNPs). The number of subjects analyzed per variant in the CKDGen HapMap meta-analysis varies between 127,290 and 175,670 subjects.

Table 4. ***The 53 genetic loci associated with eGFRcrea with genome wide significance from the meta-analysis based on low density reference panels in the CKDGen consortium****. *Results from the non-diabetics meta-analysis. Position is given on GRCh build 37. The gene closest to the variant is listed in italics if it was not tested for replication. Ref./ Non-Ref. (RAF) are the reference allele, the non-reference allele and the allele frequency of the reference allele. Beta is the effect of the reference allele. SE is the standard error of the SNP. P-values and standard errors are corrected for genomic inflation. I^2% is the percentage of the heterogeneity I-squared from the meta-analysis results. Results are sorted by chromosome and position.*

Variant ID	Chr	Position (bp)	Closest Gene	Ref./ Non-Ref.(RAF)	Effect (SE)	P-value	I^2%	#Subjects
rs1800615	1	15,832,281	*CASP9*	T/C(0.30)	-0.0058(0.0009)	$1.90x10^{-09}$	16.2	133,723
rs12136063	1	110,014,170	SYPL2	A/G(0.70)	0.0045(0.0008)	$4.71x10^{-08}$	0	175,426
rs267734	1	150,951,477	*ANXA9*	T/C(0.79)	-0.0079(0.0011)	$4.01x10^{-13}$	16.08	133,724
rs3850625*	1	201,016,296	CACNA1S	A/G(0.12)	0.0083(0.0013)	$6.82x10^{-11}$	0	153,107
rs2802729	1	243,501,763	SDCCAG8	A/C(0.44)	-0.0046(0.0008)	$2.20x10^{-08}$	9	174,808
rs807601	2	15,793,014	*DDX1*	T/G(0.34)	0.0064(0.0009)	$6.60x10^{-12}$	18.65	133,714
rs1260326	2	27,730,940	*GCKR*	T/C(0.42)	0.0068(0.0009)	$3.38x10^{-14}$	11.9	133,767
rs6546838	2	73,679,280	*ALMS1*	A/G(0.76)	-0.0093(0.0010)	$7.72x10^{-20}$	19.9	133,798
rs4667594	2	170,008,506	LRP2	A/T(0.53)	-0.0044(0.0008)	$3.52x10^{-08}$	4	175,337
rs7422339	2	211,540,507	*CPS1*	A/C(0.32)	-0.0106(0.0010)	$2.18x10^{-23}$	0	130,702
rs2712184*	2	217,682,779	IGFBP5	A/C(0.58)	-0.0053(0.0008)	$1.33x10^{-10}$	0	153,854
rs6795744	3	13,906,850	WNT7A	A/G(0.15)	0.006(0.0011)	$3.33x10^{-08}$	18	175,490
rs2861422	3	141,724,644	*TFDP2*	T/C(0.27)	0.0074(0.0010)	$9.12x10^{-14}$	0	132,783
rs9682041*	3	170,091,902	SKIL	T/C(0.87)	-0.0068(0.0012)	$2.58x10^{-08}$	2	150,911
rs10513801*	3	185,822,353	ETV5	T/G(0.87)	0.0072(0.0012)	$1.03x10^{-09}$	0	154,774
rs17319721	4	77,368,847	*SHROOM3*	A/G(0.43)	-0.0114(0.0009)	$1.32x10^{-37}$	0	133,700
rs228611	4	103,561,709	NFKB1	A/G(0.48)	-0.0056(0.0008)	$3.58x10^{-12}$	4	175,445
rs11959928	5	39,397,132	*DAB2*	A/T(0.44)	-0.0083(0.0009)	$1.66x10^{-20}$	31.85	133,703
rs6420094	5	176,817,636	*SLC34A1*	A/G(0.66)	0.0096(0.0010)	$4.92x10^{-22}$	12.76	131,066
rs7759001	6	27,341,409	ZNF204	A/G(0.76)	-0.0051(0.0009)	$1.75x10^{-08}$	0	175,483
rs9472135	6	43,809,802	*VEGFA*	T/C(0.71)	-0.0080(0.0010)	$3.34x10^{-15}$	11.39	133,689
rs316009	6	160,675,764	*SLC22A2*	T/C(0.10)	0.0131(0.0014)	$4.38x10^{-19}$	11.59	133,807
rs10277115	7	1,285,195	UNCX	A/T(0.24)	0.009(0.0012)	$8.72x10^{-14}$	0	156,521
rs3750082	7	32,919,927	KBTBD2	A/T(0.34)	0.0045(0.0008)	$3.22x10^{-08}$	2	168,494
rs848490	7	77,555,005	*TMEM60*	C/G(0.73)	0.0073(0.0010)	$7.80x10^{-13}$	10.7	127,291
rs7805747	7	151,407,801	*PRKAG2*	A/G(0.25)	-0.0130(0.0011)	$7.96x10^{-29}$	5.12	131,772
rs6459680	7	156,258,568	RNF32	T/G(0.74)	-0.0055(0.0009)	$1.07x10^{-09}$	0	175,347
rs3758086	8	23,714,992	*STC1*	A/G(0.42)	-0.0071(0.0009)	$1.71x10^{-15}$	11.22	133,679
rs4744712	9	71,434,707	*PIP5K1B*	A/C(0.40)	-0.0071(0.0009)	$4.29x10^{-15}$	32.77	133,720
rs1044261	10	1,065,710	*WDR37*	T/C(0.08)	-0.0113(0.0016)	$1.21x10^{-11}$	12.81	133,810
rs10994860*	10	52,645,424	A1CF	T/C(0.18)	0.0077(0.0011)	$1.07x10^{-12}$	2	154,644
rs163160*	11	2,789,955	KCNQ1	A/G(0.82)	0.0065(0.0011)	$2.26x10^{-09}$	14	154,684
rs963837	11	30,749,090	*MPPED2*	T/C(0.54)	-0.0078(0.0009)	$5.69x10^{-18}$	2.1	133,795
rs4014195	11	65,506,822	AP5B1	C/G(0.64)	0.0055(0.0008)	$1.10x10^{-11}$	0	175,400
rs10774021	12	349,298	*SLC6A13*	T/C(0.65)	-0.0063(0.0009)	$4.77x10^{-12}$	9.97	133,799
rs10491967	12	3,368,093	TSPAN9	A/G(0.11)	-0.0095(0.0013)	$5.18x10^{-14}$	0	175,670
rs7956634	12	15,321,194	PTPRO	T/C(0.81)	-0.0068(0.0010)	$7.17x10^{-12}$	0	175,448
rs1106766	12	57,809,456	INHBC	T/C(0.22)	0.0061(0.0010)	$2.41x10^{-09}$	11	154,665
rs716877	13	72,347,448	*DACH1*	C/G(0.40)	0.0049(0.0009)	$6.22x10^{-08}$	24.33	133,723
rs476633	15	41,392,134	*INO80*	C/G(0.57)	0.0051(0.0009)	$8.90x10^{-09}$	0	133,713
rs2467853	15	45,698,793	*GATM*	T/G(0.62)	0.0126(0.0009)	$1.05x10^{-42}$	4.89	132,748
rs491567	15	53,946,593	*WDR72*	A/C(0.78)	-0.0084(0.0010)	$2.86x10^{-15}$	10.67	133,723
rs1394125	15	76,158,983	*UBE2Q2*	A/G(0.35)	-0.0073(0.0010)	$5.47x10^{-14}$	13.57	133,740
rs13329952	16	20,366,507	*UMOD*	T/C(0.81)	-0.0158(0.0011)	$9.47x10^{-43}$	65.53	132,769
rs164748*	16	89,708,292	DPEP1	C/G(0.53)	0.0046(0.0008)	$1.95x10^{-08}$	17	154,497
rs2453580	17	19,438,321	*SLC47A1*	T/C(0.59)	0.0064(0.0009)	$2.93x10^{-11}$	0	132,640
rs9916302	17	37,499,949	*CDK12/ FBXL20*	T/C(0.74)	-0.008(0.0010)	$4.78x10^{-15}$	11.36	127,290
rs11657044	17	59,450,105	*BCAS3*	T/C(0.19)	-0.0115(0.0012)	$7.89x10^{-22}$	0.91	132,577
rs8091180*	18	77,164,243	NFATC1	A/G(0.56)	-0.006(0.0010)	$1.28x10^{-09}$	0	153,715
rs12460876	19	33,356,891	*SLC7A9*	T/C(0.60)	-0.0066(0.0009)	$1.86x10^{-13}$	0	133,702
rs11666497	19	38,464,262	SIPA1L3	T/C(0.18)	-0.0058(0.0011)	$4.25x10^{-08}$	24	168,911
rs6088580	20	33,285,053	TP53INP2	C/G(0.47)	-0.0049(0.0008)	$1.79x10^{-09}$	0	167,365
rs17216707	20	52,732,362	BCAS1	T/C(0.79)	-0.0077(0.0010)	$8.83x10^{-15}$	1	173,627

2.1.6 Methods to quantify the gain between meta-analyses

In the following, I show methods to compare meta-analyses: The differences of imputation qualities of 1000 Genomes compared to HapMap imputed variants. The proportion of the phenotypic variance explained is evaluated, which is a measure of how much of the genetic influence on the phenotype is revealed by the meta-analysis. Independent signals in previously reported loci are identified. First I focus on power, which is the probability to detect a genetic locus associated with an outcome. It depends on the number of subjects analyzed, allele frequency and phenotypic variance and explains why loci could be identified in a meta-analysis.

2.1.6.1 Power for continuous traits

A t test was used to infer whether the genetic variants were associated with a continuous phenotype, i.e., whether the observed genetic effect was significantly different from zero: $H_0: b = 0\ vs.\ H_A: b \neq 0$. Assuming the null hypothesis, the test statistic T = b/se follows a t-distribution with n-2 degrees-of-freedom (df): $T = b/se \sim t(n-2)\ |H_0$. Assuming the alternative hypothesis, the power of the t test is given by

$$\mathrm{P}(\mathrm{T} \leq \mathrm{t}_{\frac{\alpha}{2}}|\mathrm{H}_A) + \mathrm{P}(\mathrm{T} \geq -\mathrm{t}_{\frac{\alpha}{2}}|\ \mathrm{H}_A) =$$

$$\mathrm{P}(\mathrm{T} \leq \mathrm{t}_{\frac{\alpha}{2}}|\mathrm{H}_A) + \mathrm{P}(\mathrm{T} \geq \mathrm{t}_{1-\frac{\alpha}{2}}|\ \mathrm{H}_A)$$

where α is the α-level of the t test (e.g., genome-wide significance level α =5 x 10-8) and t_q is the q-th quantile of a cumulative t distribution with n-2 degrees of freedom. Assuming a true effect β with true standard error SE(β), the power can be written as

$$P\left(T^* \leq t_{\frac{\alpha}{2}} - \frac{\beta}{SE(\beta)}\right) + P\left(T^* \leq t_{\frac{\alpha}{2}} + \frac{\beta}{SE(\beta)}\right).$$

The formula can also be used to calculate the post-hoc power of an observed (estimated) genetic effect with observed standard error:

$$P\left(T^* \leq t_{\frac{\alpha}{2}} - \frac{\beta}{SE(\beta)}\right) + P\left(T^* \leq t_{\frac{\alpha}{2}} + \frac{\beta}{SE(\beta)}\right).$$

Here, T*~t(n-2) is known and follows a t-distribution with n-2 degrees of freedom (df). I also consider the observed allele frequency *AF* and the variance of the phenotype *var(y)*. The explained variance R^2, the squared standard error SE^2 and the variance of the observed genotype *var(x)* are known as

$$R^2 = \frac{\beta^2 * var(X)}{var(Y)}\ ;\ SE^2 = \sqrt{\frac{(1-R^2) * var(X)}{n * var(X)}}\ ;\ var(x) = 2 * AF * (1-AF)$$

Thus the test statistic can be transformed to

$$\frac{\beta_O}{SE_O} = \frac{\beta_O}{\sqrt{\frac{1 - R^2 * var(y)}{n * var(x)}}} = \sqrt{\frac{{\beta_O}^2 * n * var(x)}{1 - R^2 * var(y)}} = \sqrt{\frac{R^2 * n}{1 - R^2}}$$

The observed effect (β_o) and the observed standard error (SE_o) were used to obtain the final power as distribution function with *n*-2 degrees of freedom, where *n* is the observed number of individuals. The post-hoc power was calculated with the reported number of subjects per variant. To account for the uncertainty generated in imputed genotypes, the effective power was calculated with the effective number of subjects (imputation quality * number of subjects per variant).

2.1.6.2 Comparing imputation qualities between 1000 Genomes and HapMap imputed genotypes

The quality of genotype imputation of imputed variants was quantified with the imputation qualities (see **chapter 1.3.3**). For all studies that contributed to the HapMap [33] and 1000 Genomes meta-analysis, the median imputation qualities per variant were compared by MAF and imputation quality categories.

2.1.6.3 Evaluating meta-analyses dissecting studies from the HapMap and the 1000 Genomes meta-analysis

Genetic risk loci for kidney function were evaluated in the HapMap and the 1000 Genomes meta-analysis. I was interest to known why a locus was identified in either one of the meta-analyses exclusively or why a locus was identified by both meta-analyses. The meta-analysis of all studies in the HapMap (n = 133,831) and 1000 Genomes (n = 110,517) meta-analysis were compared to the meta-analyses of studies, which were analyzed in both HapMap (n = 84,461) and 1000 Genomes meta-analysis (n = 85,088). Also the meta-analyses of studies, which were analyzed exclusively in the HapMap (n = 49,370) and the 1000 Genomes (n = 25,429) meta-analysis were evaluated. Additionally, the study specific estimated effects, standard errors and the number of subjects per study were illustrated with forestplots.

2.1.6.4 Evaluating potential bias between 1000 Genomes and HapMap meta-analysis

To evaluate if there is a potential bias (i. e. if the results are systematically distorted) the effects per variant were compared between the CKDGen 1000 Genomes and the CKDGen HapMap meta-analysis. The effects were compared between the complete CKDGen HapMap and the complete CKDGen 1000 Genomes meta-analyses and between the meta-analyses of all studies, which contributed to both HapMap and 1000 Genomes meta-analyses.

2.1.6.5 Calculating the proportion of Phenotypic Variance Explained

Another measure to compare meta-analyses is the proportion of phenotypic variance explained. The question is, if the association analysis of the 1000 Genomes meta-analysis can explain a higher

proportion of the phenotypic variance as the HapMap meta-analysis. The proportion of phenotypic variance explained by a number an independent loci is estimated as $\frac{\beta^2 * var(SNP)}{var(\text{phenotype}))}$, where the variance of the variant $var(variant)$ is $2 * MAF * (1 - MAF)$ and β is the effect of the variant [56]. To estimate the proportion of phenotypic variance explained by the lead variants the variance of the residuals of log (eGFRcrea) is taken from the ARIC study (one of the largest study in both meta-analyses, n=9,038), because this big population based study serves as a good predictor of the real phenotypic variance across all studies. All lead variants are assumed to have independent effects on the phenotype.

2.2 Results on comparing 1000 Genomes with HapMap meta-analysis

To understand how much can be gained by 1000 Genomes meta-analysis compared the HapMap meta-analysis I analyzed the CKDGen data twice. First, using GWAS on HapMap imputed genotypes and second, using 1000 Genomes imputed genotypes. I compared the imputation quality and the detectability of association signals.

My evaluation focused on comparing the imputation quality of the analyzed genotypes, identifying risk loci with kidney function (eGFRcrea), identifying genetic risk loci for kidney function, which can be identified additional to those from the HapMap meta-analysis and evaluating why these were not be identified before. I analyzed why loci were identified with genome-wide significance in the HapMap, but not in the 1000 Genomes meta-analysis and compared lead variants detected with genome-wide significance by both meta-analyses. Finally I identified if the overall genetic contribution to kidney function is increased with meta-analyzing 1000 Genomes imputed GWAS.

2.2.1 Comparing imputation qualities

First I focused on the question whether 1000 Genomes imputed variants from the CKDGen consortium exhibit a higher imputation quality compared to HapMap imputed genotypes in the CKDGen consortium. To have the most informative set of variants, all variants present in at least half of all subjects in the 1000 Genomes and HapMap meta-analyses (number of subjects ≥ 66,910 and ≥ 55,260) were analyzed.

First, the median imputation qualities per variant were compared between 10,971,307 and 2,433,307 variants from the 1000 Genomes and the HapMap imputed genotypes, respectively. Among those, there is a lower proportion of well imputed (RSQ >0.8) variants and a higher proportion of medium well (0.4 < RSQ ≤ 0.8) imputed variants (73.86% vs. 92.41% and 25.85% vs. 6.33%) for 1000 Genomes and HapMap imputed variants, respectively. These proportions can also be observed in the subsets of very common and common variants (92.53% vs. 94.60% and 7.47% vs. 4.76%) and for less frequent variants (62.54% vs. 74.02% and 37.33% vs. 19.59%, **Table 5**).

Nevertheless, **the absolute number** of well and medium well imputed variants was much higher in the 1000 Genomes imputed genotypes compared to the HapMap imputed genotypes due to the higher number of variants in the 1000 Genomes reference panel (8,103,139 vs. 2,249,027; 2,836,412 vs. 154,161 and 31,784 vs. 30,570 for well, medium well and badly imputed variants, respectively).

The CKDGen 1000 Genomes imputed genotypes yielded 2,075,815 **rare variants**. 30.47% of them are well, 68.17% were medium well and only 1.36% of them were poorly imputed. In contrast, the HapMap imputed genotypes yielded no rare variants.

The **overlap** of variants present in both meta-analyses consisted of 2,408,573 variants. Thus 98.97% of variants in the HapMap meta-analysis results were also present in the 1000 Genomes meta-analysis. I was interested to know if the imputation quality can be increased by using the 1000 Genomes reference data compared to using the HapMap reference data for genotype imputation. In the overlap there were 3.63% more well imputed variants in the 1000 Genomes imputed genotypes compared to the HapMap imputed genotypes (96.94% vs. 93.31%); underpinning the superior imputation quality of the 1000 Genomes reference data (**Table 5**). These differences are remarkable as the number of subjects in the CKDGen 1000 Genomes imputed genotypes is **lower**, compared to the CKDGen HapMap imputed genotypes. These imputation qualities contribute to the effective number of subjects (imputation quality * number of subjects), which also influences the power (see **chapter 2.2.3** and **Appendix 7.1**). In summary, imputation qualities in the CKDGen 1000 Genomes imputed genotypes are generally higher compared to the CKDGen HapMap imputed genotypes and the 1000 Genomes imputed genotypes additionally yield imputed rare variants.

Table 5. ***Comparing the distribution of imputation quality between 1000 Genomes imputed with HapMap imputed genotype****. Shown are absolute numbers and the relative frequencies in poorly (RSQ≤ 0.4), medium (0.4 < RSQ≤ 0.8) and well imputed variants (0.8 < RSQ) in total and by categories of MAF (rare: MAF<=0.01, less frequent: 0.01 < MAF < 0.05 and common: MAF >= 0.05) of the median imputation qualities per variant across all studies in the 1000 Genomes and HapMap meta-analysis results alone and in the overlap of variants in the meta-analyses based on HapMap and 1000 Genomes imputed genotypes. All variants were meta-analyzed in at least half of all subjects in both 1000 Genomes and HapMap meta-analyses (number of subjects ≥ 66,910 and ≥ 55,260).*

MAF	RSQ	All variants in 1000 Genomes meta-analysis	All variants in HapMap meta-analysis	Overlapping variants in 1000 Genomes meta-analysis	Overlapping variants in HapMap meta-analysis
All	RSQ>0.8	8,103,124 (73.86%)	2,249,027 (92.41%)	2,334,834 (96.94%)	2,247,511 (93.31%)
	0.4<RSQ≤0.8	2,836,399 (25.85%)	154,161 (6.33%)	73,657 (3.06%)	147,152 (6.11%)
	RSQ≤0.4	31,784 (0.29%)	30,570 (1.26%)	82 (<0.01%)	13,910 (0.58%)
MAF ≥ 0.05	RSQ>0.8	5,885,422 (92.53%)	2,057,447 (94.6%)	2,118,463 (97.92%)	2,056,299 (95.04%)
	0.4<RSQ≤0.8	475,160 (7.47%)	103,467 (4.76%)	45,070 (2.08%)	100,472 (4.64%)
	RSQ≤0.4	63 (<0.01%)	14,018 (0.64%)	7 (<0.01%)	6,769 (0.31%)
0.01 < MAF < 0.05	RSQ>0.8	1,585,176 (62.54%)	191,580 (74.02%)	216,371 (88.3%)	191,212 (78.04%)
	0.4<RSQ≤0.8	946,240 (37.33%)	50,694 (19.59%)	28,587 (11.67%)	46,680 (19.05%)
	RSQ≤0.4	3,431 (0.13%)	16,552 (6.4%)	75 (0.03%)	7,141 (2.91%)
MAF ≤ 0.01	RSQ>0.8	632,526 (30.47%)	0 (0%)	0 (0%)	0 (0%)
	0.4<RSQ≤0.8	1,414,999 (68.17%)	0 (0%)	0 (0%)	0 (0%)
	RSQ≤0.4	28,290 (1.36%)	0 (0%)	0 (0%)	0 (0%)
Number of variants		10,971,307	2,433,307	2,408,573	2,408,573

2.2.2 Confirming known and identifying additional susceptibility loci for kidney function by the CKDGen 1000 Genomes meta-analysis

There were 53 genetic loci associated with eGFRcrea, which were identified by HapMap meta-analyses or candidate gene analyses [18, 20, 42-45]. **Table 4** shows these 53 genetic loci from the CKDGen HapMap meta-analysis of either all or the non-diabetic subjects, depending on which analysis yielded the smaller p-value (see **chapter 2.1.5**). In order to understand how the CKDGen 1000 Genomes meta-analysis compares to the CKDGen HapMap meta-analysis, I evaluated, if the susceptibility loci from the HapMap meta-analysis can be confirmed and if additional loci can be identified by the 1000 Genomes meta-analysis.

2.2.2.1 Genomic inflation

To ensure, that the association statistics are not distorted by genomic inflation in the CKDGen 1000 Genomes meta-analysis, the genomic inflation factor was computed on all variants in the CKDGen 1000 Genomes meta-analysis present in at least half of all subjects (n ≥ 55,260). The resulting λ = 1.12 did not indicate substantial inflation of the test statistic. P-values and standard errors were corrected for genomic control for this λ throughout all further analyses. The variants in the CKDGen 1000 Genomes meta-analysis excluding are all lead variants from the 53 genome wide significant loci ± 1MB down- and upstream show a similar inflation (λ=1.11) as the overall set of variants. Overall no substantial inflation was observed in the data and the λ of 1.12 was used to correct the p-values and standard errors for all further analyses.

2.2.2.2 Genome wide significant loci from the 1000 Genomes meta-analysis

The question is, if the CKDGen 1000 Genomes meta-analysis can extend our knowledge of the genetic architecture of kidney function (on the outcome eGFRcrea) compared to the previous CKDGen HapMap meta-analysis. Specifically the questions arise, if loci from the HapMap meta-analysis (see **Table 4**) can be confirmed and if additional loci might be identified.

Overall ten loci (where the variant with the lowest p-value is genome-wide significant, **Table 6**) were identified, which are not found in the HapMap meta-analysis. The Manhattan Plot of the CKDGen 1000 Genomes meta-analysis highlighting the 10 additional loci compared to the CKDGen HapMap meta-analysis is in **Appendix 7.2**.

The lead variants in 39 genetic loci among the 53 loci from the CKDGen HapMap meta-analysis were identified with a genomic controlled p-value ≤ $5x10^{-8}$. The lead variants (with the smallest p-value in the locus) were different in all but one loci. The lead variant rs807601 in the *DDX1* locus was the lead

variant in both HapMap and the 1000 Genomes meta-analyses. Furthermore, it was genome-wide significant in both meta-analyses. The p-values in the 14 not genome-wide significant lead variants ranged from $6.69x10^{-8}$ to $6.23x10^{-4}$ (see **Table 7**). In the CKDGen 1000 Genomes meta-analysis, the lead variants in 49 of the 53 top variants from the CKDGen HapMap meta-analysis are SNPs. Among those, 38 loci were genome-wide significant. One of the genome-wide significant lead variants is an insertion (rs139926232 in *WDR72* on chromosome 15). There were two not genome-wide significant lead variants, which were insertions (rs72144865 and rs4000129 in *WNT7A* and *KBTB2* on chromosome 3 and 7, respectively). And one not genome-wide significant lead variant is a deletion (rs33945021 in NFKB1 on chromosome 4). All loci showed no or low heterogeneity. Only the *UMOD* locus, known to show biological relevance with kidney filtration rate [57], showed moderate heterogeneity of 63.3%. Although the CKDGen 1000 Genomes meta-analyses yielded more than 2 Million rare variants, no genome wide significant lead variant was rare.

In summary, there were 10 loci identified as novel loci in the 1000 Genomes meta-analysis, which were not identified with genome-wide significance in the HapMap meta-analysis, 39 loci could be identified by both HapMap and 1000 Genomes meta-analysis and 14 loci could be identified by the HapMap meta-analysis but not in the 1000 Genomes meta-analysis.

*Table 6.**Variants with smallest p-values in novel loci associated with estimated glomerular filtration rate (eGFRcrea) in individuals of European ancestry**. P-values are corrected for inflation using genomic control. The effective power is calculated using the effective number of subjects (number of subjects * median imputation quality) per variant. The median imputation qualities and the number of analyzed studies are shown in **Appendix 7.1**.*

Variant ID	Chr:Position (bp)	Closest Gene	Ref./ Non-Ref. (RAF)	Effect(SE)	P-value	I^2%	#subjects	Effective power
rs10874312	1:82,944,571	*LPHN2*	A/G(0.67)	-0.0057(0.0011)	$2.20x10^{-08}$	19	107,335	0.7939
rs12144044	1:113,248,791	*RHOC*	A/C(0.28)	-0.0061(0.0011)	$2.87x10^{-08}$	0	110,517	0.8155
rs187355703	2:176,993,583	*HOXD8*	C/G(0.97)	0.0182(0.0030)	$5.15x10^{-10}$	2	109,257	0.9535
rs111366116	5:53,295,546	*ARL15*	T/C(0.11)	0.0094(0.0015)	$6.27x10^{-10}$	22	110,517	0.9395
rs113246091	5:67,739,274	*PIK3R1*	A/G(0.10)	-0.0095(0.0016)	$1.98x10^{-09}$	43	110,105	0.9105
rs7764488	6:133,812,872	*EYA4*	A/G(0.32)	0.0061(0.0011)	$4.08x10^{-09}$	1	110,516	0.8895
rs13298297	9:119,264,108	*ASTN2*	A/G(0.20)	-0.0075(0.0014)	$1.53x10^{-08}$	0	110,514	0.8202
rs1111571	16:68,363,181	*SLC7A6*	A/G(0.71)	0.0061(0.0011)	$6.20x10^{-09}$	0	109,275	0.8591
rs9962915	18:5,593,171	*EPB41L3*	T/C(0.48)	-0.0055(0.0010)	$7.19x10^{-09}$	0	110,516	0.8411
rs12458009	18:59,350,507	*RNF152*	T/G(0.78)	-0.0064(0.0012)	$2.90x10^{-08}$	22	107,325	0.7790

Chr:Position is given as chromosome: position on GRCh build 37. The closest gene closest to the SNP is listed. SE = standard error. I^2% = Heterogeneity I-Squared. #subjects is the number of subjects in the analysis. Effective power is estimated with imputation quality and the number of subjects in the analysis (see chapter 2.1.6).

Table 7. ***Lead variants from the 1000 Genomes meta-analysis in the 53 known loci identified by previous analyses.***

Variant ID	Chr	Position (bp)	Closest Gene	Ref. / Non-Ref.(RAF)	Effect(SE)	P-value	I^2%	#Subjects
rs7546668	1	15,855,123	*CASP9*	C/G(0.3111)	-0.0063(0.0011)	$1.14x10^{-09}$	0	110,517
rs10127790	1	109,891,133	*SYPL2*	T/C(0.7018)	0.0061(0.0011)	$7.58x10^{-09}$	0	110,517
rs267738	1	150,940,625	*ANXA9*	T/G(0.7903)	-0.0091(0.0012)	$1.48x10^{-14}$	20.4	107,336
rs3850625	1	201,016,296	*CACNA1S*	A/G(0.1197)	0.0088(0.0016)	$2.24x10^{-08}$	15.4	107,335
rs2783971	1	243,474,536	*SDCCAG8*	A/C(0.4879)	-0.0037(0.0010)	$1.20x10^{-04}$	12	110,517
rs807601	2	15,793,014	*DDX1*	T/G(0.3392)	0.0067(0.0011)	$3.84x10^{-11}$	0	110,517
rs780093	2	27,742,603	*GCKR*	T/C(0.4002)	0.0081(0.0010)	$1.57x10^{-16}$	13.6	110,517
rs4500972	2	73,767,897	*ALMS1*	A/G(0.2107)	0.0108(0.0013)	$3.20x10^{-18}$	14.7	110,517
rs35472707	2	169,995,581	*LRP2*	T/C(0.0475)	-0.0108(0.0023)	$3.93x10^{-06}$	38	109,257
rs1047891	2	211,540,507	*CPS1*	A/C(0.3197)	-0.0089(0.0011)	$1.90x10^{-16}$	0	110,517
rs2541381	2	217,683,836	*IGFBP5*	T/G(0.5286)	-0.0047(0.0010)	$1.77x10^{-06}$	0	110,517
rs72144865	3	13,918,234	*WNT7A*	C/CG(0.1804)	0.0064(0.0015)	$1.50x10^{-05}$	4.1	92,536
rs7640665	3	141,813,172	*TFDP2*	A/G(0.7153)	-0.0072(0.0011)	$4.66x10^{-11}$	17.3	110,517
rs6770214	3	171,006,768	*SKIL*	A/G(0.8114)	-0.0044(0.0013)	$6.23x10^{-04}$	0	110,517
rs6809651	3	185,814,642	*ETV5*	A/G(0.1277)	-0.0081(0.0015)	$2.34x10^{-08}$	0	109,275
rs13146355	4	77,412,140	*SHROOM3*	A/G(0.4421)	-0.0121(0.0010)	$3.18x10^{-37}$	26.2	110,517
rs33945021	4	103,573,122	*NFKB1*	TTTAA/T(0.477)	-0.0049(0.0011)	$9.54x10^{-06}$	6.5	88,495
rs700236	5	39,367,739	*DAB2*	A/G(0.5751)	0.0084(0.0010)	$1.74x10^{-18}$	39.1	110,517
rs3812036	5	176,813,404	*SLC34A1*	T/C(0.2594)	-0.0102(0.0012)	$8.90x10^{-19}$	0	108,404
rs9348765	6	27,314,650	*ZNF204*	A/T(0.6946)	-0.0048(0.0011)	$1.06x10^{-05}$	0	110,516
rs1317983	6	43,806,335	*VEGFA*	T/C(0.3053)	0.008(0.0011)	$1.10x10^{-13}$	11.8	110,516
rs2279463	6	160,668,389	*SLC22A2*	A/G(0.8764)	0.0118(0.0015)	$1.07x10^{-15}$	15.9	110,516
rs62435145	7	1,286,567	*UNCX*	T/G(0.6606)	-0.0077(0.0014)	$2.71x10^{-08}$	22.6	95,370
rs4000129	7	33,113,699	*KBTBD2*	T/TAA(0.7752)	0.0048(0.0013)	$2.63x10^{-04}$	1	92,536
rs112029703	7	77,238,678	*TMEM60*	A/T(0.2834)	-0.0065(0.0011)	$1.38x10^{-09}$	0	110,517
rs10254101	7	151,415,536	*PRKAG2*	T/C(0.2868)	-0.0104(0.0012)	$6.09x10^{-20}$	16.2	110,517
rs6971211	7	155,664,686	*RNF32*	T/C(0.4033)	-0.0055(0.0011)	$6.69x10^{-08}$	31.5	107,336
rs36071802	8	23,715,871	*STC1*	T/C(0.5776)	0.0079(0.0010)	$1.16x10^{-15}$	8.2	110,517
rs10746942	9	71,434,465	*PIP5K1B*	A/G(0.6246)	0.0086(0.0010)	$3.56x10^{-18}$	7.4	110,514
rs80282103	10	899,071	*WDR37*	A/T(0.9115)	0.0123(0.0018)	$1.12x10^{-11}$	0	110,517
rs10994856	10	52,645,248	*A1CF*	A/G(0.1807)	0.0075(0.0013)	$4.77x10^{-09}$	0	110,517
rs84178	11	2,774,374	*KCNQ1*	A/G(0.1616)	-0.0078(0.0013)	$4.29x10^{-09}$	31.7	110,516
rs3925584	11	30,760,335	*MPPED2*	T/C(0.5466)	-0.0079(0.0010)	$2.09x10^{-16}$	23.9	110,516
rs11604462	11	65,551,648	*AP5B1*	A/G(0.3584)	-0.006(0.0010)	$1.90x10^{-09}$	2.6	110,516
rs11062167	12	364,739	*SLC6A13*	A/G(0.5205)	-0.0055(0.0010)	$1.12x10^{-08}$	20.2	110,515
rs67551338	12	3,393,100	*TSPAN9*	T/C(0.0633)	-0.0124(0.0021)	$2.17x10^{-09}$	0	110,515
rs12826808	12	15,323,380	*PTPRO*	A/T(0.8071)	-0.0064(0.0012)	$1.36x10^{-07}$	14.9	110,515
rs3741414	12	57,844,049	*INHBC*	T/C(0.2289)	0.0064(0.0013)	$1.59x10^{-07}$	0	109,685
rs9529913	13	72,345,089	*DACH1*	T/C(0.5973)	-0.0066(0.0010)	$2.51x10^{-11}$	13.4	110,516
rs6492982	15	41,399,951	*INO80*	T/C(0.5628)	-0.0048(0.0011)	$2.39x10^{-06}$	0	110,517
rs2453533	15	45,641,225	*GATM*	A/C(0.3868)	-0.0135(0.0010)	$2.65x10^{-43}$	0	110,517
rs139926232	15	53,922,280	*WDR72*	A/AATAGCT(0.2557)	0.0083(0.0013)	$7.20x10^{-11}$	0	87,986
rs10851885	15	76,304,503	*UBE2Q2*	A/G(0.7604)	$8.10x10^{-3}$(0.0012)	$2.92x10^{-12}$	16.3	107,336
rs77924615	16	20,392,332	*UMOD*	A/G(0.2027)	0.0176(0.0014)	$4.57x10^{-40}$	63.2	110,517
rs428232	16	89,713,969	*DPEP1*	T/C(0.5134)	0.0052(0.0011)	$2.87x10^{-07}$	20.8	108,404
rs894680	17	19,440,538	*SLC47A1*	A/G(0.3902)	-0.0074(0.0011)	$5.46x10^{-12}$	0.3	102,993
rs12451586	17	37,633,835	*CDK12/ FBXL20*	A/T(0.6889)	-0.0092(0.0012)	$2.78x10^{-15}$	9.8	110,515
rs9895661	17	59,456,589	*BCAS3*	T/C(0.8091)	0.0125(0.0013)	$4.37x10^{-21}$	0	110,515
rs71359461	18	77,156,103	*NFATC1*	C/G(0.4751)	-0.0086(0.0014)	$3.67x10^{-10}$	15.6	78,608
rs7247977	19	33,358,355	*SLC7A9*	T/C(0.6167)	-0.007(0.0010)	$2.35x10^{-12}$	3.2	110,517
rs151087334	19	38,205,244	*SIPA1L3*	A/G(0.4488)	0.0061(0.0013)	$2.26x10^{-06}$	34.6	110,517
rs6058093	20	33,213,196	*TP53INP2*	A/C(0.5354)	-0.0074(0.0011)	$2.26x10^{-13}$	0	110,515
rs6127099	20	52,731,402	*BCAS1*	A/T(0.7094)	-0.0095(0.0012)	$2.91x10^{-17}$	6.9	110,515

Position is on GRCh build 37. The gene closest to the variant is listed. Ref./ Non-Ref. (RAF) = reference/ non-reference allele and frequency of the reference allele. Beta is the effect of the reference allele. SE is the standard error of the variant.I2% is the percentage of the heterogeneity I-squared from the meta-analysis results. #Subjects are the number of analyzed subjects

2.2.2.3 *Evaluating the imputation quality in the kidney function disease loci*

I was interested, if the imputation qualities in the top variants identified by the HapMap or the 1000 Genomes meta-analysis were comparable. Overall, 53 top variants from the CKDGen HapMap meta-analysis and 23 top variants from the CKDGen 1000 Genomes meta-analysis, which were also analyzed in the Hapmap meta-meta-analysis, were compared (**Figure 5**). The median of the imputation qualities per variant across all studies in the meta-analysis were compared. From the overall 76 variants, 59 showed a higher median imputation quality in the 1000 Genomes imputed studies. The majority of variants was well imputed (RSQ >0.8): There were seven variants in the HapMap imputed genotypes with a median imputation quality smaller than 0.8 (min: 0.5, rs10277115 in *UNCX* locus) and one variant in the 1000 Genomes imputed data with a median imputation quality smaller than 0.8 (0.64, rs10277115 in *UNCX* locus). In summary, the lead variants from the CKDGen HapMap meta-analysis and from the CKDGen 1000 Genomes meta-analysis showed high median imputation qualities, both at a comparable level.

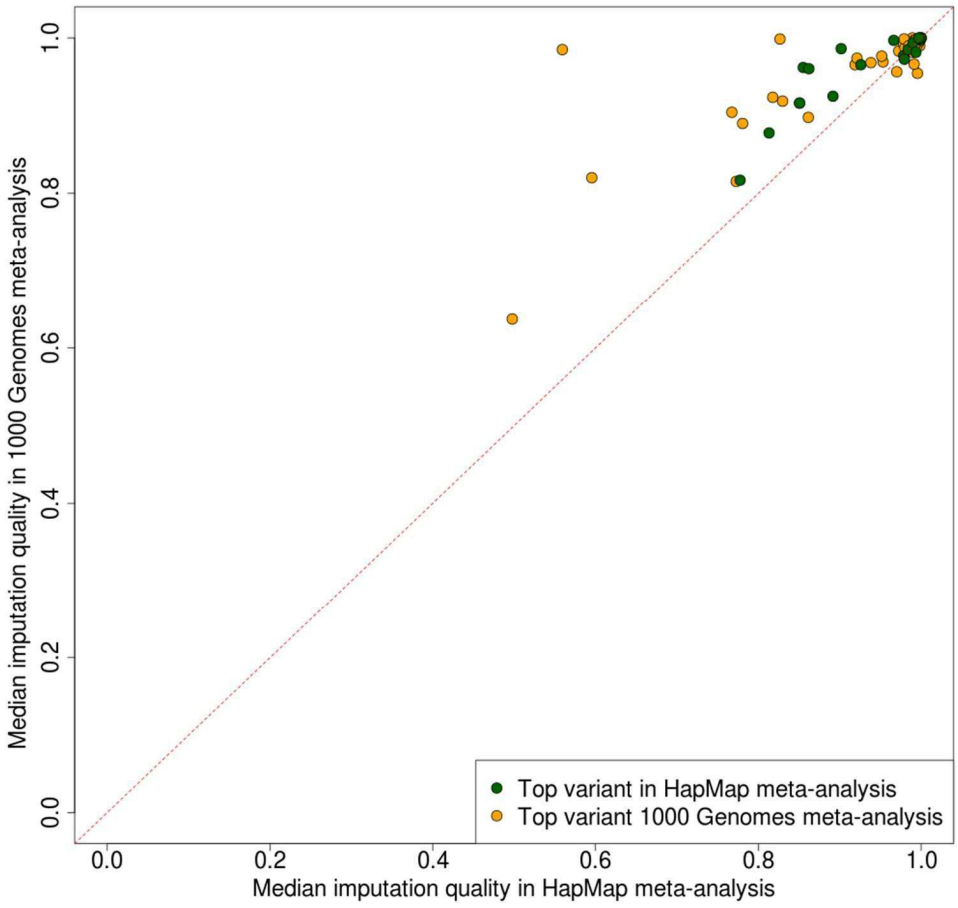

Figure 5. ***Comparison of imputation quality in the 53 lead variants from the CKDGen HapMap meta-analysis and in the 23 lead variants in the CKDGen 1000 Genomes meta-analysis.***

2.2.2.4 Characterizing the 10 additional hits from the CKDGen 1000 Genomes meta-analysis

Next, I focused on the 10 lead variants, which were genome-wide significant in the CKDGen 1000 Genomes meta-analysis and not genome-wide significant in the CKDGen HapMap meta-analysis. I was interested in the reasons why a genome-wide significant locus can be identified in the CKDGen 1000 Genomes meta-analysis, but not in the CKDGen HapMap meta-analysis: First, six of the 10 signals identified by the CKDGen 1000 Genomes meta-analysis were present in the 1000 Genomes reference panel, but not in the HapMap reference panel (see **Appendix 7.1)**. Furthermore, no proxies (i. e. other variants) with pairwise LD ≥ 0.4 for these variants could be identified. LD was estimated in subjects with European ancestry in the HapMap reference data sets (see **Web Resources**). So there was no substitute for those additional variants in the HapMap reference panel.

In the following, meta-analysis results from the HapMap and 1000 Genomes meta-analyses will be compared by three meta-analyses: in the meta-analysis of all studies (HapMap and 1000 Genomes meta-analysis); in the meta-analysis of studies analyzed in the HapMap meta-analysis but not in the 1000 Genomes meta-analysis and *vice versa* (in the following: *HapMap exclusive meta-analysis* and *1000 Genomes exclusive meta-analysis*); in studies, which were analyzed in both meta-analyses (meta-analysis of the overlapping studies).

The comparison of the HapMap with the 1000 Genomes meta-analysis yielded four variants, which were genome-wide significant in the 1000 Genomes meta-analysis, but not genome-wide significant in the HapMap meta-analysis and analyzed in both meta-analyses. **Table 8** shows the meta-analysis results from the lead variants in the four loci, which were identified by the 1000 Genomes meta-analysis, but not in the HapMap meta-analysis. The lead variants rs10874312, rs1214404, rs1111571 and rs1248009 in these four 4 loci were also analyzed in the CKDGen HapMap meta-analysis (in the *SLC7A6*, *RHOC*, *RNF152* and *LPHN2* loci*)*. All meta-analysis results are given in **Table 8.** For all four SNPs, the meta-analysis of studies analyzed in both meta-analyses yielded comparable effect estimates and p-values (see **Appendix 7.2.1**). The p-values were smaller in the 1000 Genomes exclusive meta-analysis compared to the HapMap exclusive meta-analysis (rs10874312: $6.90x10^{-03}$ vs. $3.57x10^{-01}$; rs1214404: $6.57x10^{-04}$ vs. $7.13x10^{-02}$; rs1111571: $1.92x10^{-02}$ vs. $3.00x10^{-01}$; rs1248009: $1.18x10^{-05}$ vs. $7.00x10^{-03}$). The effects were stronger in the 1000 Genomes exclusive meta-analysis compared to the HapMap exclusive meta-analysis (rs10874312: -0.0048 vs. -0.0014; rs1214404: 0.0064 vs. 0.003; rs1111571: -0.0045 vs. -0.0016; rs1248009: 0.0089 vs. 0.0046). As a consequence, first, these four SNPs were genome-wide significant in the overall CKDGen 1000 Genomes meta-analysis but not genome-wide significant in the overall CKD HapMap meta-analysis. Second, the estimated effect in the overall 1000 Genomes meta-analysis was higher compared to the estimated effect in the CKDGen HapMap meta-analysis.

So I was interested, why the stronger effects and smaller p-values were estimated in the 1000 Genomes exclusive meta-analysis compared to the HapMap exclusive meta-analysis. **Table 9** shows the studies, which contributed to the 1000 Genomes exclusive meta-analysis and to the HapMap exclusive meta-analysis, respectively. The estimated effects in the HapMap exclusive meta-analysis were small in some studies with high number of subjects. For example was the estimated effect of the AUSTWIN study 0.0010 in rs12144044 or the estimated effect of JUPITER was 0.0019 in rs12458009). Additionally, the estimated effects were not effect direction consistent across most studies for example in rs10874312 and rs1111571. In contrast the studies in the 1000 Genomes exclusive meta-analysis, which contributed most to the inverse variance weighted meta-analysis results (studies with a high number of subjects, or equivalent with small standard errors), were mostly effect direction consistent and comprised stronger effects. For example the three largest studies in the 1000 Genomes exclusive meta-analysis (LIFELINES, COLAUS and EGCUT1) estimated strong effects (-0.0038, -0.0058 and -0.0121) and contributed more than 90% (23,232 of 25,429) of subjects to the analysis. This can also observed comparable in the other three SNPs. All meta-analyses of these four SNPs were illustrated as forestplots in **Appendix 7.2.1**.

In summary, the four SNPs could be identified with genome-wide significance and stronger estimated effects in the CKDGen 1000 Genomes meta-analysis and could not be identified with genome-wide significance in the CKDGen HapMap meta-analysis. This was due to a randomly better composition of studies analyzed in the 1000 Genomes meta-analysis, not analyzed in the HapMap meta-analysis compared to the study composition analyzed in the HapMap meta-analysis, not analyzed in the 1000 Genomes meta-analysis.

Table 8. ***Meta-analysis results of a variant genome wide significant in the 1000 Genomes meta-analysis and not genome-wide significant in the HapMap meta-analysis****. Shown are the meta-analysis results on all subjects, subjects in the overlap between the meta-analyses and of those exclusively in the HapMap or 1000 Genomes meta-analysis, respectively.*

Variant ID	Chr	Position (bp)	Ref./ Non-Ref (RAF)	Effect	Standard error	P-value	#studies	I^2%	#subjects
Results from the HapMap meta-analysis of all subjects									
rs10874312	1	82717159	A/G(0.67)	-0.0041	0.0009	$5.60x10^{-06}$	50	0	133,803
rs12144044	1	113050314	C/A(0.73)	0.0051	0.0019	$6.62x10^{-07}$	50	0	133,583
rs1111571	16	66920682	G/A(0.29)	-0.0049	0.0009	$1.36x10^{-07}$	50	0	133,806
rs12458009	18	57501487	G/T(0.22)	0.0050	0.0019	$1.56x10^{-06}$	50	7.5	133,765
Results from the HapMap meta-analysis of subjects only in the HapMap meta-analysis									
rs10874312	1	82717159	A/G(0.66)	-0.0014	0.0015	$3.57x10^{-01}$	23	0	49,342
rs12144044	1	113050314	C/A(0.73)	0.0030	0.0017	$7.13x10^{-02}$	23	17.5	49,202
rs1111571	16	66920682	G/A(0.29)	-0.0016	0.0015	$3.00x10^{-01}$	23	0	49,343
rs12458009	18	57501487	G/T(0.22)	0.0046	0.0017	$7.00x10^{-03}$	23	0	49,341
Results from the HapMap meta-analysis of subjects in the overlapping studies between the HapMap and the 1000 Genomes meta-analyses									
rs10874312	1	82717159	A/G(0.67)	-0.0059	0.0011	$2.37x10^{-07}$	27	6.9	84,461
rs12144044	1	113050314	C/A(0.73)	0.0059	0.0013	$6.36x10^{-06}$	27	0	84,381
rs1111571	16	66920682	G/A(0.29)	-0.0068	0.0012	**$8.04x10^{-09}$**	27	0	84,461
rs12458009	18	57501487	G/T(0.22)	0.0048	0.0013	$2.40x10^{-04}$	27	28.5	84,424
Results from the 1000 Genomes meta-analysis of all subjects									
rs10874312	1	82944571	A/G(0.67)	-0.0057	0.0011	**$2.20x10^{-08}$**	29	19	107,335
rs12144044	1	113248791	C/A(0.72)	0.0061	0.0011	**$2.87x10^{-08}$**	33	0	110,517
rs1111571	16	68363181	G/A(0.29)	-0.0061	0.0011	**$6.20x10^{-09}$**	31	0	109,275
rs12458009	18	59350507	G/T(0.22)	0.0064	0.0012	**$2.90x10^{-08}$**	29	21.5	107,325
Results from the 1000 Genomes meta-analysis of subjects only in the 1000 Genomes meta-analysis									
rs10874312	1	82944571	A/G(0.67)	-0.0048	0.0018	$6.90x10^{-03}$	6	60.2	25,429
rs12144044	1	113248791	C/A(0.70)	0.0064	0.0019	$6.57x10^{-04}$	6	0	25,429
rs1111571	16	68363181	G/A(0.28)	-0.0045	0.0019	$1.92x10^{-02}$	6	0	25,429
rs12458009	18	59350507	G/T(0.22)	0.0089	0.0020	$1.18x10^{-05}$	6	0	25,429
Results from 1000 Genomes meta-analysis of subjects in the overlapping studies between the HapMap and the 1000 Genomes meta-analyses									
rs10874312	1	82944571	A/G(0.67)	-0.0061	0.0011	$1.08x10^{-07}$	23	0	81,906
rs12144044	1	113248791	C/A(0.73)	0.0060	0.0012	$1.58x10^{-06}$	27	14.7	85,088
rs1111571	16	68363181	G/A(0.29)	-0.0067	0.0012	**$7.22x10^{-09}$**	25	0	83,846
rs12458009	18	59350507	G/T(0.22)	0.0054	0.0013	$2.85x10^{-05}$	23	29	81,896

Chr = chromosome, Position is reported on GRCh 37 in the 1000 Genomes meta-analysis results and on NCBI build 36 in the HapMap meta-analysis, Ref./ Non-Ref (RAF) are the reference allele and non-reference allele and the frequency of the reference allele, #studies is the number of studies in the meta-analysis, I^2% is the heterogeneity as reported by the meta-analysis software, #subjects are the number of subjects in the analysis. Genome-wide significant p-values are bold.

Table 9. ***Overview over the study specific effects, standard errors and number of subjects in the meta-analysis of studies exclusively meta-analyzed in the HapMap and 1000 Genomes meta-analysis****. Effect (allele) = estimated beta for the reference allele. Data is sorted decreasing by number of subjects per study.*

	rs10874312, LPHN2			rs12144044, RHOC			rs1111571, SLC7A6			rs12458009, RNF152		
Study	**Effect (A)**	**SE**	**#subjects**	**Effect (C)**	**SE**	**#subjects**	**Effect (G)**	**SE**	**#subjects**	**Effect (G)**	**SE**	**#subjects**
HapMap exclusive studies: Studies in the CKDGen HapMap meta-analysis and not in the CKDGen 1000 Genomes meta-analysis												
AUSTWIN	-0.0070	0.0040	9,592	0.0010	0.0050	9,453	-0.0050	0.0040	9,592	0.0100	0.0050	9,592
JUPITER	-0.0028	0.0030	8,780	0.0035	0.0033	8,780	-0.0035	0.0031	8,780	0.0019	0.0034	8,780
PROSPER-PHASE	0.0002	0.0050	5,237	0.0036	0.0059	5,237	-0.0033	0.0054	5,237	0.0168	0.0059	5,237
CHS	0.0080	0.0073	3,259	0.0168	0.0089	3,259	0.0027	0.0073	3,259	0.0041	0.0079	3,259
ERF	0.0048	0.0063	2,561	0.0013	0.0071	2,561	-0.0011	0.0064	2,561	0.0115	0.0066	2,561
ADVANCE	-0.0061	0.0082	2,287	-0.0166	0.0084	2,287	0.0003	0.0087	2,287	-0.0087	0.0094	2,285
RS-II	-0.0063	0.0065	1,863	0.0030	0.0076	1,863	0.0100	0.0071	1,863	-0.0029	0.0073	1,863
NESDA	0.0039	0.0064	1,855	0.0132	0.0069	1,854	-0.0127	0.0066	1,856	0.0053	0.0072	1,856
HYPERGENES-CONTROLS	0.0108	0.0071	1,662	0.0079	0.0075	1,662	-0.0073	0.0069	1,662	0.0037	0.0082	1,662
HYPERGENES-CASES	-0.0037	0.0085	1,591	-0.0033	0.0091	1,591	-0.0051	0.0084	1,591	0.0045	0.0097	1,591
AMISH	0.0001	0.0093	1,211	-0.0048	0.0104	1,211	0.0097	0.0095	1,211	0.0044	0.0119	1,211
POPGEN	-0.0001	0.0077	1,163	-0.0050	0.0081	1,163	0.0073	0.0080	1,163	0.0002	0.0088	1,163
GENOA	-0.0064	0.0129	1,064	0.0119	0.0137	1,064	0.0122	0.0131	1,064	0.0070	0.0151	1,064
INCIPE	0.0016	0.0101	940	-0.0109	0.0111	940	0.0069	0.0098	940	-0.0098	0.0116	940
KORCULA	-0.0076	0.0102	888	0.0133	0.0113	888	0.0073	0.0109	888	0.0188	0.0111	888
SORBS	-0.0041	0.0086	878	0.0085	0.0094	878	-0.0084	0.0090	878	-0.0046	0.0094	878
OGP-TALANA	0.0048	0.0128	862	-0.0242	0.0148	862	-0.0042	0.0137	862	-0.0010	0.0155	862
VIS	0.0014	0.0127	768	0.0075	0.0144	768	0.0030	0.0126	768	0.0140	0.0140	768
BLSA	-0.0070	0.0160	723	0.0210	0.0190	723	0.0120	0.0160	723	-0.0130	0.0180	723
DESIR	-0.0086	0.0092	715	0.0120	0.0113	715	-0.0003	0.0096	715	-0.0025	0.0110	715
ORCADES	0.0116	0.0122	704	0.0274	0.0135	704	-0.0008	0.0132	704	0.0001	0.0143	704
CROATIA-SPLIT	0.0014	0.0126	478	-0.0224	0.0141	478	0.0040	0.0123	478	0.0267	0.0137	478
EGCUT-OMNI	-0.0416	0.0291	261	0.0008	0.0323	261	-0.0525	0.0327	261	-0.0270	0.0322	261
1000 Genomes exclusive studies: Studies in the CKDGen 1000 Genomes meta-analysis and not in the CKDGen HapMap meta-analysis												
LIFELINES	-0.0038	0.0021	13,386	0.0066	0.0022	13,386	-0.0019	0.0023	13,386	0.0092	0.0024	13,386
COLAUS	-0.0058	0.0040	5,409	0.0077	0.0042	5,409	-0.0095	0.0040	5,409	0.0081	0.0045	5,409
EGCUT1	-0.0121	0.0060	4,437	0.0017	0.0062	4,437	-0.0067	0.0066	4,437	0.0093	0.0068	4,437
IPM II	-0.0235	0.0151	1,307	0.0092	0.0149	1,307	-0.0232	0.0152	1,307	0.0247	0.0167	1,307
GENDIAN	0.0074	0.0206	450	-0.0031	0.0219	450	-0.0106	0.0202	450	-0.0185	0.0224	450
IPM I	0.0919	0.0323	440	0.0162	0.0336	440	-0.0050	0.0336	440	-0.0051	0.0373	440

2.2.2.5 *Characterizing the 39 SNPs genome-wide significant in both Hapmap and 1000 Genomes meta-analysis*

There were 39 variants, which were genome-wide significant in both 1000 Genomes and HapMap meta-analysis. Here, these variants are exemplified on the example of rs807601 (in the *DDX1* locus), which is the lead variant in both the CKDGen HapMap and the 1000 Genomes meta-analysis. **Table 10** gives all meta-analysis result of this variant. In line with the previous results, the meta-analysis of studies analyzed in both meta-analyses yielded comparable effect estimates and p-values. A comparison of the estimated effects, standard errors, imputation qualities and p-values in the overlapping studies between the CKDGen HapMap and the CKDGen 1000 Genomes meta-analysis of this SNP were illustrated in **Appendix 7.2.2.** Also the 1000 Genomes exclusive meta-analysis and the HapMap exclusive meta-analysis yielded comparable effect estimated (HapMap: -0.004, 1000 Genomes: -0.005) and p-values (Hapmap: 7.80×10^{-3}, 1000 Genomes: 7.37×10^{-3}). The effect estimates and standard errors in the HapMap exclusive studies and the 1000 Genomes exclusive studies were equally balanced (**Table 11**) and thus yielded comparable meta-analysis results, which led to a genome wide significant p-value and comparable effects in the overall HapMap and 1000 Genomes meta-analysis, respectively.

A detailed comparison of the studies analyzed by both the CKDGen HapMap and 1000 Genomes meta-analysis and forestplots of all studies in both meta-analyses were given in **Appendix 7.2.2.** One variant was exemplified here, but the results can be generalized to the other 38 variants, as well.

Table 10. Meta-analysis results of a variant genome wide significant in both HapMap and 1000 Genomes meta-analysis. Shown are the meta-analysis results on all subjects in the HapMap meta-analysis, those which were exclusively in the HapMap meta-analysis, those which were in the overlap between HapMap and 1000 genomes meta-analysis, all subjects, which were in the 1000 Genomes meta-analysis, those which were exclusively in the 1000 Genomes meta-analysis and those, which were in the overlap between the HapMap and the 1000 Genomes meta-analysis.

Variant ID	Chr	Position (bp)	Ref./ Non-Ref (RAF)	Effect	Standard error	P-value	#studies	I^2%	#subjects
Results from the HapMap meta-analysis of all subjects									
rs807601	2	15710465	G/T(0.66)	-0.0064	0.0009	**6.60×10^{-12}**	50	15.3	133,714
Results from the HapMap meta-analysis of subjects only in the HapMap meta-analysis									
rs807601	2	15710465	G/T(0.66)	-0.0040	0.0015	7.80×10^{-3}	23	35.5	49,342
Results from the HapMap meta-analysis of subjects in the overlapping studies between the HapMap and 1000 Genomes meta									
rs807601	2	15710465	G/T(0.66)	-0.0076	0.0011	**3.62×10^{-11}**	27	0	84,372
Results from the 1000 Genomes meta-analysis of all subjects									
rs807601	2	15793014	G/T(0.66)	-0.0067	0.0011	**3.84×10^{-11}**	33	0	110,517
Results from the 1000 Genomes meta-analysis of subjects only in the 1000 Genomes meta-analysis									
rs807601	2	15793014	G/T(0.67)	-0.0050	0.0018	7.37×10^{-03}	6	0	25,429
Results from 1000 Genomes meta-analysis of subjects in the overlapping studies between the HapMap and 1000 Genomes meta									
rs807601	2	15793014	G/T(0.66)	-0.0074	0.0011	**4.98×10^{-11}**	27	5.3	85,088

Chr = chromosome, Position is reported on GRCh 37 in the 1000 Genomes meta-analysis results and on NCBI build 36 in the HapMap meta-analysis, Ref./ Non-Ref (RAF) are the reference allele and non-reference allele and the frequency of the reference allele, #studies is the number of studies in the meta-analysis, I^2% is the heterogeneity as reported by the meta-analysis software, #subjects are the number of subjects in the analysis. Genome-wide significant p-values are bold.

Table 11. ***Overview over the study specific effects, standard errors and number of subjects in the meta-analysis of studies exclusively meta-analyzed in the HapMap and 1000 Genomes meta-analysis of rs807601.***

	rs807601		
Study	**Effect (G)**	**SE**	**#subjects**
HapMap exclusive studies			
AUSTWIN	-0.0130	0.0040	9,592
JUPITER	-0.0013	0.0029	8,780
PROSPER-PHASE	0.0002	0.0052	5,237
CHS	0.0053	0.0071	3,259
ERF	-0.0055	0.0063	2,561
ADVANCE	0.0000	0.0082	2,287
RS-II	-0.0024	0.0068	1,863
NESDA	-0.0005	0.0065	1,856
HYPERGENES-CONTROLS	0.0054	0.0071	1,662
HYPERGENES-CASES	-0.0162	0.0085	1,591
AMISH	0.0012	0.0112	1,211
POPGEN	-0.0130	0.0077	1,163
GENOA	-0.0193	0.0124	1,064
INCIPE	0.0089	0.0106	940
KORCULA	0.0034	0.0101	888
SORBS	-0.0164	0.0085	878
OGP-TALANA	0.0103	0.0135	862
VIS	-0.0020	0.0129	768
BLSA	0.0100	0.0160	723
DESIR	-0.0037	0.0095	715
ORCADES	-0.0253	0.0120	704
CROATIA-SPLIT	-0.0287	0.0122	478
EGCUT-OMNI	0.0738	0.0321	261
1000 Genomes exclusive studies			
LIFELINES	-0.0306	0.0022	13,386
COLAUS	-0.0018	0.0039	5,409
EGCUT1	-0.0033	0.0064	4,437
IPM II	-0.0301	0.0146	1,307
GENDIAN	-0.0049	0.0197	450
IPM I	-0.0068	0.0313	440

Effect (allele) = estimated beta for the reference allele. SE = standard error, #subjects = number of subjects in the study. Data is sorted decreasing by number of subjects per study for HapMap exclusive and 1000 Genomes exclusive studies separately.

2.2.2.6 *Characterizing the 14 SNPs genome wide significant in the HapMap meta-analysis and not genome-wide significant in the 1000 Genomes meta-analysis*

Overall 14 variants were not genome-wide significant in the 1000 Genomes meta-analysis and genome-wide significant in the HapMap meta-analysis. I exemplify these variants with the lead variant in the *LRP2* locus (rs4667594). All meta-analysis results of this variant are shown in **Table 12.** In line with the previous results, the meta-analysis of studies analyzed in both meta-analyses yielded comparable effect estimates and p-values. The HapMap exclusive meta-analysis yielded a smaller p-value compared to the1000 Genomes exclusive meta-analysis ($2.10x10^{-05}$ vs. $1.02x10^{-02}$) with a stronger effect (0.0061 vs 0.0044). In **Table 13** is shown, that the studies in the HapMap meta-analysis yielded mainly effect direction consistent effect estimates. Additionally, the overall number of subjects in the HapMap meta-analysis analysis was higher compared to the number of subjects analyzed in the 1000 Genomes meta-analysis (49,343 subjects vs. 25,429 subjects). Altogether, this finally led to a genome wide significant p-value in the CKDGen HapMap meta-analysis with a stronger effect compared to the CKDGen 1000 Genomes meta-analysis. A detailed comparison of the studies analyzed by both CKDGen HapMap and 1000 Genomes meta-analysis and forestplots of all studies in both meta-analyses are given in **Appendix 7.2.3.** One variant was exemplified, but the results can be generalized to the other 13 variants, as well.

*Table 12. **Meta-analysis results of a variant genome wide significant in the HapMap meta-analysis and not genome-wide significant in the 1000 Genomes meta-analysis**. Shown are the meta-analysis results on all subjects in the HapMap meta-analysis, those which were exclusively in the HapMap meta-analysis, those which were in the overlap between HapMap and 1000 genomes meta-analysis, all subjects, which were in the 1000 Genomes meta-analysis, those which were exclusively in the 1000 Genomes meta-analysis and those, which were in the overlap between the HapMap and the 1000 Genomes meta-analysis.*

Variant ID	Chr	Position (bp)	Ref./ Non-Ref (RAF)	Effect	Standard error	P-value	#studies	I^2%	#subjects
Results from the HapMap meta-analysis of all subjects									
rs4667594	2	169716752	T/A(0.47)	0.0044	0.0009	**$3.52x10^{-08}$**	50	4.3	133,715
Results from the HapMap meta-analysis of subjects only in the HapMap meta-analysis									
rs4667594	2	169716752	T/A(0.47)	0.0061	0.0014	$2.10x10^{-05}$	23	48.2	49,343
Results from the HapMap meta-analysis of subjects in the overlapping studies between the HapMap and 1000 Genomes meta									
rs4667594	2	169716752	T/A(0.47)	0.0032	0.0011	$3.00x10^{-03}$	27	0	84,372
Results from the 1000 Genomes meta-analysis of all subjects									
rs4667594	2	170008506	T/A(0.47)	0.0033	0.0009	$2.50x10^{-04}$	33	0	110,517
Results from the 1000 Genomes meta-analysis of subjects only in the 1000 Genomes meta-analysis									
rs4667594	2	170008506	T/A(0.47)	0.0044	0.0017	$1.02x10^{-02}$	6	37.2	25,429
Results from 1000 Genomes meta-analysis of subjects in the overlapping studies between the HapMap and 1000 Genomes meta									
rs4667594	2	170008506	T/A(0.47)	0.0029	0.0011	$6.63x10^{-03}$	27	0	85,088

Chr = chromosome, Position is reported on GRCh 37 in the 1000 Genomes meta-analysis results and on NCBI build 36 in the HapMap meta-analysis, Ref./ Non-Ref (RAF) are the reference allele and non-reference allele and the frequency of the reference allele, #studies is the number of studies in the meta-analysis, I^2% is the heterogeneity as reported by the meta-analysis software, #subjects are the number of subjects in the analysis. Genome-wide significant p-values are bold.

Table 13. ***Overview over the study specific effects, standard errors and number of subjects in the meta-analysis of studies exclusively meta-analyzed in the HapMap and 1000 Genomes meta-analysis of rs4667594***

	rs4667594		
Study	**Effect (T)**	**SE**	**#subjects**
HapMap exclusive studies			
AUSTWIN	0.0100	0.0040	9,592
JUPITER	0.0029	0.0028	8,780
PROSPER-PHASE	0.0040	0.0049	5,237
CHS	0.0248	0.0067	3,259
ERF	0.0034	0.0059	2,561
ADVANCE	-0.0212	0.0079	2,287
RS-II	0.0076	0.0064	1,863
NESDA	0.0055	0.0061	1,856
HYPERGENES-CONTROLS	0.0058	0.0065	1,662
HYPERGENES-CASES	-0.0028	0.0077	1,591
AMISH	0.0073	0.0098	1,211
POPGEN	0.0034	0.0071	1,163
GENOA	0.0147	0.0125	1,064
INCIPE	0.0013	0.0096	940
KORCULA	0.0209	0.0096	888
SORBS	0.0253	0.0081	878
OGP-TALANA	0.0284	0.0120	862
VIS	-0.0041	0.0117	768
BLSA	0.0230	0.0150	723
DESIR	-0.0066	0.0089	715
ORCADES	0.0252	0.0117	704
CROATIA-SPLIT	-0.0046	0.0115	478
EGCUT-OMNI	0.0181	0.0278	261
1000 Genomes exclusive studies			
LIFELINES	0.0041	0.0020	13,386
COLAUS	0.0104	0.0037	5,409
EGCUT1	-0.0085	0.0058	4,437
IPM II	0.0106	0.0134	1,307
GENDIAN	-0.0003	0.0187	450
IPM I	-0.0049	0.0315	440

Effect (allele) = estimated beta for the reference allele. SE = standard error, #subjects = number of subjects in the study. Data is sorted decreasing by number of subjects per study for HapMap exclusive and 1000 Genomes exclusive studies separately.

2.2.3 Comparing the power to detect a genome wide significant locus

The comparison of the HapMap with the 1000 Genomes meta-analysis yielded 4 variants, which were genome-wide significant in the 1000 Genomes meta-analysis, but not genome-wide significant in the HapMap meta-analysis. I was interested in the post-hoc power including the estimated effect to detect these variants with genome-wide significance. The post-hoc power analysis focused specifically on the lead variants in these 4 additional loci (*SLC7A6*, *RHOC*, *RNF152* and *LPHN2).* These 4 lead variants had a power of 85.9%, 81.5%, 77.9% and 79.4% in the 1000 Genomes meta-analysis (**Figure 6A, Figure 7A**). In the CKDGen HapMap meta-analysis these lead variants had power below 80% (64.4%, 67.1%, 47.8% and 34.4%, see **Figure 6B** and **Figure 7B**), where the locus was not associated with eGFRcrea with genome wide significance. Generally the CKDGen 1000 Genomes meta-analysis yielded 8 of the 10 novel and 39 of the 53 known loci with a lead variant with more than 80% power for genome wide significance (p-value ≤ $5x10^{-8}$) in.

To account for the uncertainty of imputed genotypes, the effective power of a variant was calculated with the effective number of subjects (number of subjects in the analysis of the variant * imputation quality of the variant). In the 10 additional loci, only the lead variants in *LPHN2* and *RNF152* had an effective power below 80%. The effective power in the 8 other loci ranged from 81.6% to 95.4% with a median of 87.4% (**Figure 6A and Figure 7A**). The lead variant in 3 known loci, which had theoretically more than 80% power are *not* genome-wide significant (p-values = $1.59x10^{-7}$, $6.69x10^{-8}$ and $2.26x10^{-6}$, respectively) in the 1000 Genomes meta-analysis (grey loci *R3HDM2*, *AC005534.6 and SIPA1L3* in **Figure 6A***)*. Comparing the effective power in those 3 loci, the lead variants in *R3HDM2* and *AC005534.6* showed an effective power of 78.0% and 80.3%, respectively. On the other hand, the lead variant in *SIPA1L3* in the 1000 Genomes meta-analysis had a power of only 52.8% due to its low imputation quality (RSQ=0.59, **Appendix 7.1 and Figure 7A**). The lead variants in the HapMap meta-analysis in those 3 loci were well imputed (with RSQ of 0.83, 0.98 and 0.95, respectively, also compare **Figure 7B**) and could reach genome-wide significance. The 6 novel loci, which were present in the 1000 Genomes reference panel but not in the HapMap reference panel were well powered (effective power > 79% in *HOXD8*, *ARL15*, *PIK3R1*, *EYA4*, *ASTN2* and *EPB41L3*, **Table 6**) in the 1000 Genomes meta-analysis.

In summary, power analysis shows, that the CKDGen 1000 Genomes meta-analysis yielded 39 loci among the 53 loci identified by the CKDGen HapMap meta-analysis, with a power more than 80% to detect the lead variant with genome-wide significance. In 4 loci, which were identified by the CKDGen 1000 Genomes meta-analysis, but not in the CKDGen HapMap meta-analysis, the lead variants yielded more than 77.9% power, whereas the CKDGen HapMap meta-analysis yielded power, which was smaller than 67.1% in those 4 loci. This increase in power was mainly due to larger effect sizes (in line with previous results) in the 1000 Genomes analysis compared to the HapMap analysis (0.0061 *vs*.

0.0049, 0.0061 *vs*. 0.0051, 0.0064 *vs*. 0.0050 and 0.0057 *vs*. 0.0041 in the lead variants in *SLC7A6*, *RHOC*, *RNF152* and *LPHN2,* respectively) and was driven by the different study composition.

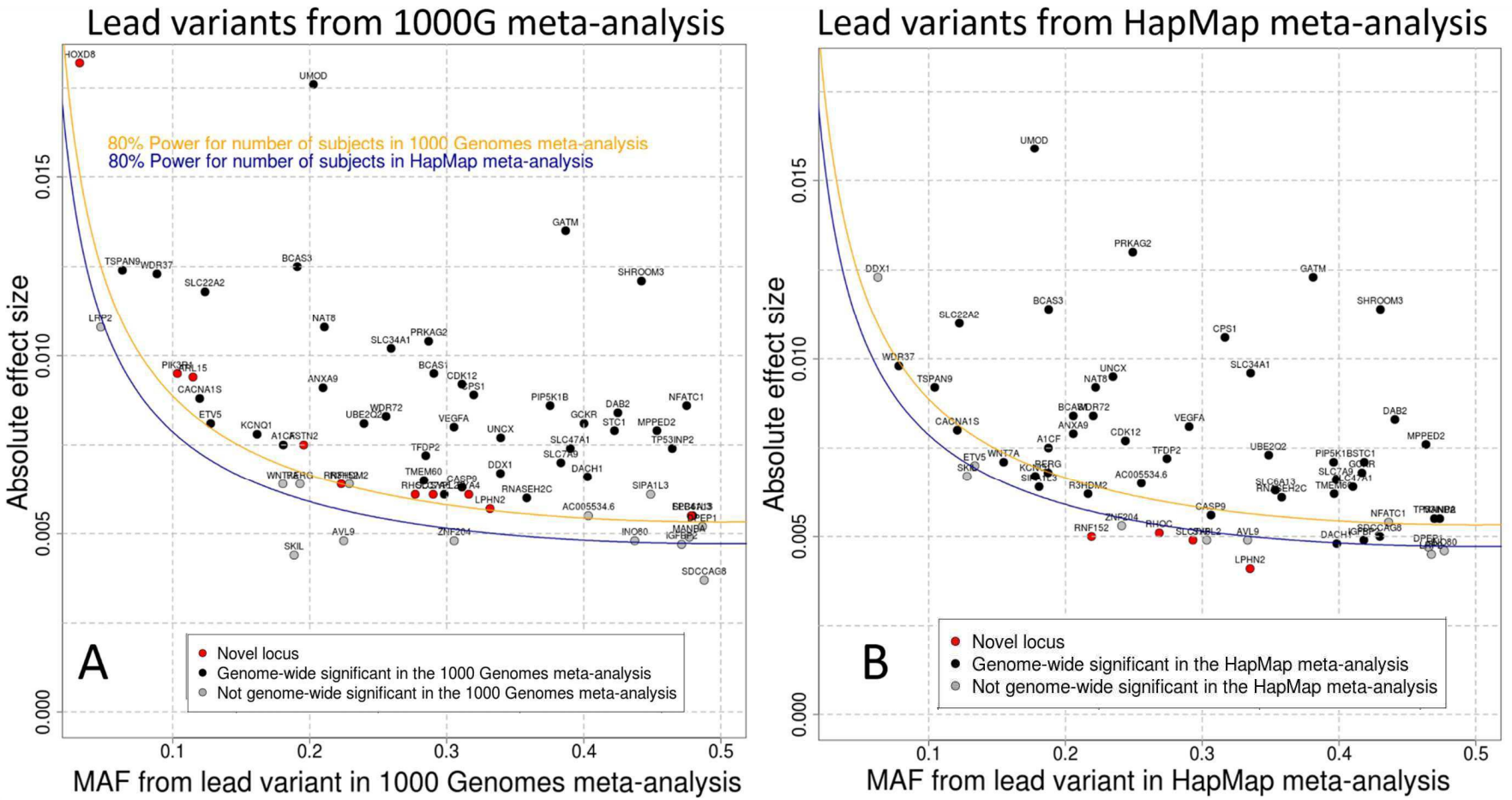

Figure 6. ***Post-hoc Power to detect lead variants in 1000 Genomes and HapMap meta-analysis****. Shown are scatterplots of MAF (x-axis) vs. the absolute effect (y-axis) of the lead variants from the 1000 Genomes meta-analysis (Panel A) and of the lead variants from the HapMap meta-analysis on eGFRcrea in all subjects (Panel B). Colored in red are the 10 lead variants from the novel loci from the 1000 Genomes meta-analysis in Panel A. in Panel B those 4 variants of the latter are shown in red, which are analyzed in more than 66,910 subjects. Both panel A and panel B also show the lead variants in the known associated loci with eGFRcrea. If a variant is genome wide significant in 1000 Genomes or HapMap analysis it is colored in black; it is color coded in grey, if it is not genome wide significant in 1000 Genomes or HapMap, respectively. The 80% power curves are shown for alpha=5x10^{-8} in 110,000 subjects in orange and for 140,000 subjects in blue.*

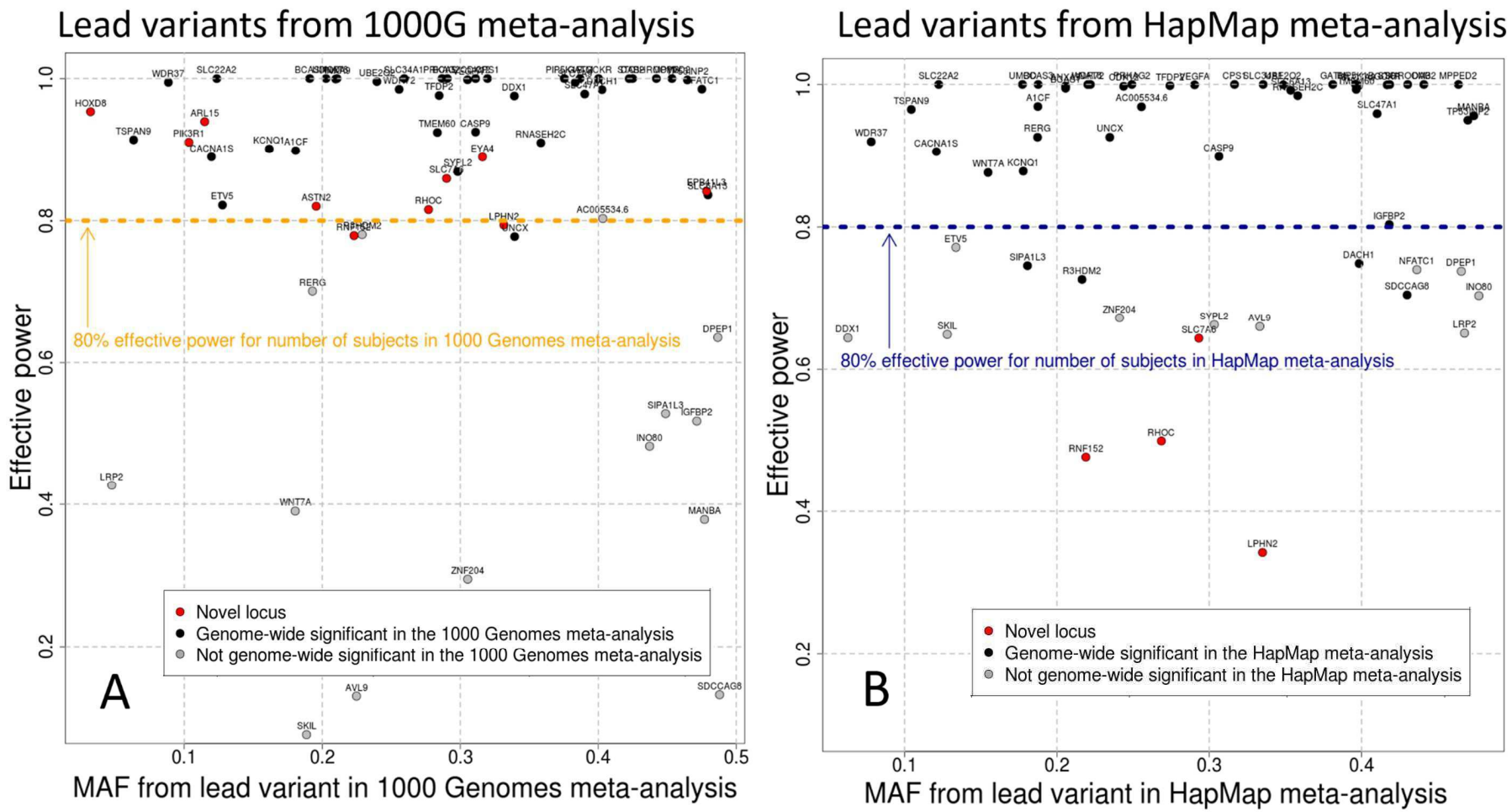

Figure 7. ***Effective power to detect no vel loci in 1000 Genomes and HapMap meta-analysis.*** *Shown are scatterplots of the effective power (x-axis) vs. MAF (y-axis) of all top variants of novel and known disease loci in the meta-analysis results based on 1000 Genomes (Panel A) and HapMap imputed genotypes (Panel B). Points are color coded in red, black and grey, if a locus is (a) novel, (b) known and genome wide significant or (c)known and not genome-wide significant in the 1000 Genomes meta-analysis (Panel A) or if it is (a) novel, (b) known and genome-wide significant or (c) known and not genome-wide significant in the HapMap meta-analysis of all subjects on eGFRcrea (Panel B). Power is calculated per locus with the effective number of subjects (number of subjects * median imputation quality of all studies, participating in the meta-analysis).*

2.2.4 Evaluating potential bias between 1000 Genomes and HapMap meta-analysis

Next I was interested if the elevated effect sizes in the 4 lead variants in *SLC7A6*, *RHOC*, *RNF152* and *LPHN2* point to a systematic bias between the CKDGen 1000 Genomes and the CKDGen HapMap meta-analysis. For this comparison the meta-analyses were restricted to variants, which were present in at least 50% of all subjects: The CKDGen 1000 Genomes meta-analysis was restricted to variants analyzed in at least 55,260 subjects and the CKDGen HapMap meta-analysis was restricted to variants analyzed in at least 66,900 subjects. Then the effect estimates were compared between the HapMap and the 1000 Genomes meta-analyses (**Figure 8**).

First, the full meta-analysis results were used (Hapmap: 133,832 subjects and 1000 Genomes: 110,517 subjects, **Panel A**). The 2,409,287 variants presents in both meta-analyses were categorized into two MAF bins (common: MAF > 5% and less frequent: MAF ≤ 5%). The effects showed a symmetric distribution and the median differences between the effects in the less frequent variants was $2x10^{-4}$ and 0 in the common variants, indicating no biased effect estimates.

Second, the effects were compared in the meta-analyses of the studies, overlapping the HapMap and 1000 Genomes meta-analysis (HapMap: 84,461 subjects 1000 Genomes: 85,088 subjects, **Panel B**). The overlap comprised 2,380,972 variants, of which 2,154,407 variants showed a MAF > 5% and 226,565 variants showed a MAF ≤ 5%. The median differences of the effects were 0 in both groups, indicating no biased effect estimates.

In summary, no systematic inflation of the effect sizes were observed between the CKDGen 1000 Genomes and the CKDGen HapMap meta-analysis.

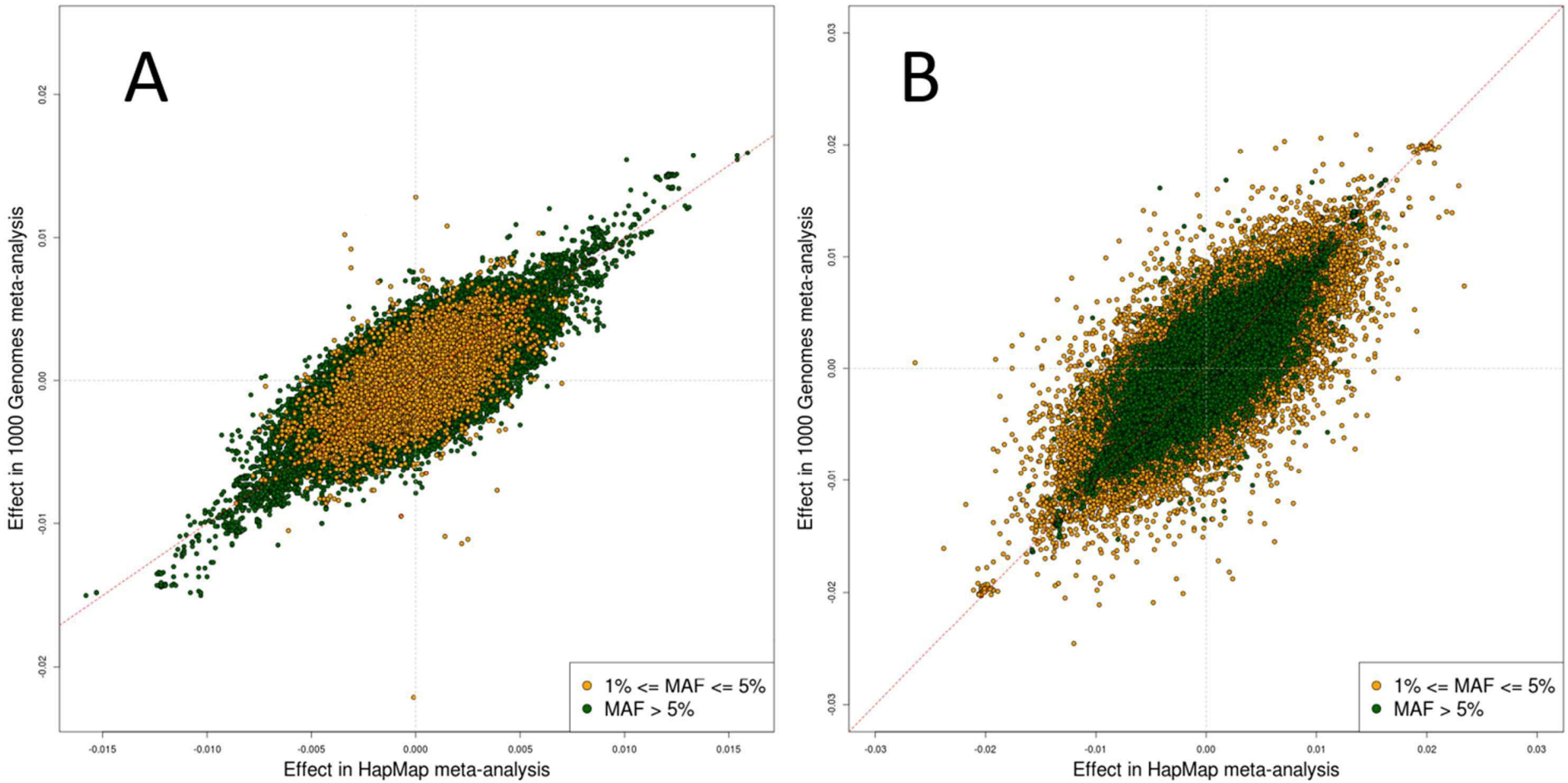

Figure 8. ***Comparison of effects, between CKDGen 1000 Genomes and CKDGen HapMap meta-analyses****. Shown are scatterplots of the effects in the full HapMap (n=133,832) and 1000 Genomes (n=110,517) meta-analyses (Panel A) and in the meta-analysis results of the overlapping studies between the HapMap (n=84,461) and the 1000 Genomes (n=85,088) meta-analyses (Panel B).*

2.2.5 Evaluating the proportion of phenotypic variance explained

Next I was interested if the CKDGen 1000 Genomes meta-analysis explained more of the proportion of phenotypic variance explained than the CKDGen HapMap meta-analysis. The proportion of phenotypic variance of eGFRcrea in the CKDGen **HapMap** meta-analysis explained by all 53 identified variants was 3.22%. The overall proportion of phenotypic variance explained in the CKDGen **1000 Genomes** meta-analysis was 3.58%: 0.46% for the 10 additional variants and 3.12% for the already known variants. Known loci that did not reach genome wide significance explained additional 0.41%. The moderate proportion of trait variance explained in the CKDGen 1000 Genomes meta-analysis is in line with findings from GWAS of other phenotypes [58].

2.2.6 Summary

In summary, this chapter shows how much can be gained by 1000 Genomes meta-analysis compared the HapMap meta-analysis. The analyses were exemplified on kidney filtration rate with data from the CKDGen consortium. 110,517 subjects were part of the CKDGen 1000 Genomes meta-analysis and 133,831 subjects were part of the CKDGen HapMap meta-analysis.

The absolute number of well imputed variants in the 1000 Genomes imputed genotypes is higher compared to the HapMap imputed genotypes. There is also a gain in imputation quality in the variants analyzed with both 1000 Genomes and HapMap reference panels. 10 additional genetic loci associated with eGFRcrea compared to the CKDGen HapMap meta-analysis were identified. But although the 1000 Genomes imputed variants yield a considerable amount of well imputed rare variants, no lead variant in novel loci was rare. 6 of those 10 additional genetic loci associated with eGFRcrea compared to the CKDGen HapMap meta-analysis are present in the 1000 Genomes reference panel and are not in the HapMap reference panel. The other 4 of the 10 additional genetic loci associated with eGFRcrea in the 1000 Genomes meta-analysis compared to the HapMap meta-analysis could be identified because the genetic effects were estimated higher in the 1000 Genomes meta-analysis compared to the CKDGen meta-analysis. One must be aware, that the true genetic effect is unknown and that the effects yielded by genome-wide meta-analysis are only estimates of the true effect. The increased effects in these four lead variants were not due to systematic inflated effects. Also the imputation qualities were comparable. Thus, the increased post-hoc power to detect these variants was due to the increased effect size, which was due to a randomly better study composition. 39 of the 53 known genetic loci of kidney function could be identified with genome-wide significance in the HapMap and 1000 Genomes meta-analysis. The number of subjects in the 1000 Genomes meta-analysis was too low or effect directions were too inconsistent to detect the remaining 14 of the 53 known genetic loci of kidney function. Finally the phenotypic variance explained could be increased by the CKDGen 1000 genomes meta-analysis compared to the CKDGen HapMap meta-analysis.

2.3 PhaseLift: An approach and software to facilitate the re-imputing of study data

Study analysts perform the time intensive tasks of phasing and imputation prior to analyzing the imputed variants for meta-analyses in consortia. Reference panel are constantly updated. Thus, study data must be re-phased und re-imputed every time a new reference panel is released. While the genotype imputation might be tractable once, study analyst can be overwhelmed by the computation burden of re-imputing study data regularly. This motivated my development of the software *PhaseLift*, which saves time in re-imputing study data with different reference panels, which is subject in this chapter.

Genotype imputation is an essential method for meta-analyzing study data genotyped on different genotyping chips for each study. It infers untyped variants in the study data with the help of external reference panels. Assuming that study and reference panel are on the same annotation, the imputation process involves (i) phasing of study genotypes (e.g. with MaCH [32], ImputeV2 [59] or ShapeIT [28, 60] and (ii) imputing of missing alleles in study haplotypes based on the information from reference haplotypes (e.g. using *minimac* [22] or *ImputeV2*). The imputed genotypes are given as dosages or probabilities and are accompanied by a measure of imputation quality (e.g. “RSQ” for minimac or “INFO” for ImputeV2).

If study and reference data are on different annotations, the imputation process will fail to match the reference SNPs correctly to the study SNPs. The annotation of study and reference data can differ for several reasons. First, study and reference data can be on different builds, which requires a so-called “lift-over” of SNP identifiers and positions from the study build to the reference build. The newer build annotation can involve SNP IDs disappearing (SNPs resolved to be monomorphic) or two SNP IDs merging into one (one ID overruled). Second, both study and reference data can contain SNPs without rs-numbers assigned (“new” SNPs). This can occur in the study data due to manufacturer-specific SNPs (annotated via the manufacturer’s annotation file) or in the reference data due to new variants detected by the sequencing of the respective reference data. This requires (i) creating a generic identifier using chromosome and position (e.g. <chr>:<pos>) for “new” SNPs in the reference panel that are not present in the study data, (ii) discarding new SNPs in the study data that are not present in the reference panel, (iii) matching new SNPs in the study data that are also present in the reference panel but possibly with different identifiers by chromosome and position.

2.3.1 Four steps of harmonizing study and reference data in the current and the novel approach of harmonization

In the case that study and reference data are on different annotation, and this is ignored, this would lead to losing genotypes in the imputation process: a SNP from the study data for which the SNP ID does not match the SNP ID in the reference data will be deleted in the imputing step and will be

imputed unnecessarily. This SNP will thus be subject to imputation uncertainty despite having been available as genotype in the study data. For example, if study data genotyped on an Affymetrix 500K SNP array is annotated on build 36 and the reference data is on build 37, the imputation would lose the genotypes of 15,000 variants due to unresolved annotation differences. Furthermore, SNPs with allelic mismatches between study and reference data will be poorly imputed or – in the case of palindromic SNPs (A/T or C/G SNPs) - will be useless after imputation. For example, for a study as described above, allele mismatches would affect about 17,000 SNPs. Therefore, a harmonization between study and reference data is highly recommended. I conceptualized four steps of harmonization: (1) Lifting rs-numbers: The rs-numbers in the study data are updated with RsMergeArch (**see Web Resources**), a remapping file for rs-numbers. (2) Lifting SNP positions: Chromosomes and positions in the study data are updated to match those in the reference data using remapping files like hg18ToHg19.over.chain (**see Web Resources**) to map build 36.3 to 37.1. This usually involves not only shifts of positions, but also changes of relative positions, which require a re-ordering of the study SNP list. (3) Harmonizing SNP IDs: Generic SNP IDs using chromosome and position are used to match SNPs without assigned rs-numbers. (4) Harmonizing alleles: Allele mismatches can often be explained by different strand annotations between study and reference data. In such cases, the information of the study SNP is corrected, otherwise the SNP is discarded in the study data. For non-palindromic SNPs (non-A/T-, non-C/G-polymorphisms), discordant strand annotations are obvious, e.g. as “A/G” polymorphism in the study and “T/C” polymorphism in the reference data. For palindromic SNPs (A/T-, C/G-polymorphisms), discordant strand annotation cannot be resolved for SNPs with minor allele frequencies (MAFs) around 0.5; allele assignment suggested by the allele frequencies is corrected if the MAF is below a certain threshold, for example 0.4, or, preferably testing for deviance from 0.5 by Fisher test.

The two-step imputation differentiates between phasing and imputing of the study data. The phasing does not – in principal - make use of the reference data. Therefore, re-phasing of the study data is not required when a re-imputation to a newer reference panel is planned (**Figure 9a**).This, however, pertains only to the case where the newer reference panel contains more subjects without updating the SNP list or SNP information (position, alleles, allele frequencies). If the reference data is on a different build than the study data or contains different SNPs or different SNP information for other reasons, the study data needs to be harmonized to the reference data. In the current pre-phasing lift-over approach, the study SNPs are harmonized with the reference SNPs on the genotype level, which then requires re-phasing before re-imputing (**Figure 9b**). This is an appropriate approach for imputing study data for the first time, but not for re-imputation due to the computational challenges for repeated phasing.

I have thus developed an approach which takes full advantage of the two-step imputation and harmonizes study and reference data on the haplotype level (post-phasing lift-over) instead of the genotype level (pre-phasing lift-over, **Figure 9c**). Thus, when the study is to be re-imputed to a newer reference panel, the study haplotypes are harmonized to the reference haplotypes and then the study data is re-imputed with no re-phasing required.

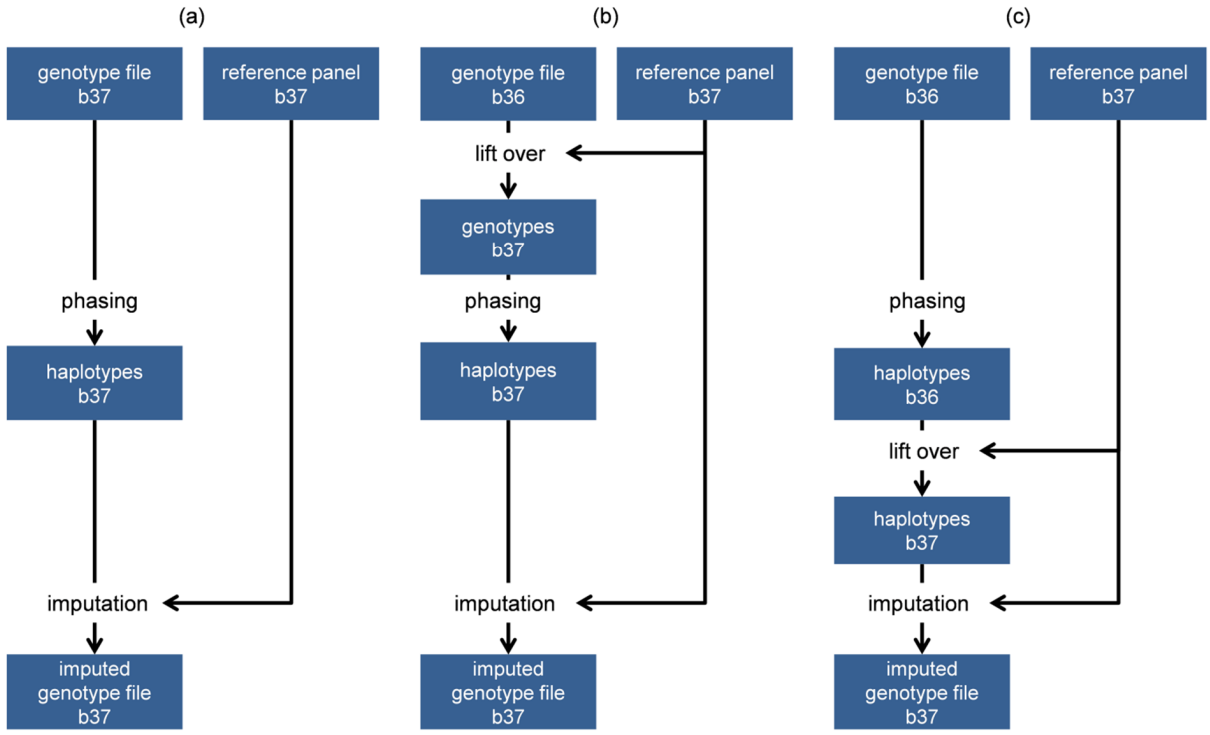

Figure 9. ***Overview of the imputation process****: (a) without lift-over (study and reference data on same annotation): Study data is phased without using the reference data and then imputed using the reference haplotypes; (b) with pre-phasing lift-over (current approach for study and reference data on different annotation): Study data is lifted/harmonized to the reference data, then phased, and finally imputed; (c) with post-phasing lift-over (novel approach for study and reference data on different annotation): study data is phased, then lifted/harmonized to the reference data, and finally imputed.*

2.3.2 Comparing pre-phasing lift over with post-phasing lift over

I compared the pre-phasing lift over with the post-phasing lift-over by imputing data from the KORA-F3 study [61] (n=1,644) with genotypes derived from an Affymetrix 500K SNP array and initially annotated on build 36.3. For imputation, I used the GIANT ALL 1000G Phase I v3 on build 37.1 as reference panel (**see Web Resources**), *MaCH* 1.0.16.b for phasing and *minimac* 2012.10.9 for imputation. The two different approaches are implemented as follows: 1) pre-phasing lift-over approach: The study data is lifted from build 36.3 to 37.1, harmonized with the specific SNP list and SNP information of the GIANT ALL reference panel, re-phased and re-imputed. 2) post-phasing lift-over approach: Study data is phased on its original build 36.3, phased study data is lifted from build 36.3 to 37.1 and harmonized to the SNP list and SNP information of the phased GIANT all reference panel, and

then the study alleles are imputed. This resulted in two sets of 30,062,047 imputed SNPs after excluding the genotyped SNPs. Most of the following is exemplified for chromosome 1 (#SNPs= 2,354,606).

As a gold standard, I used genotypes derived from the Illumina Cardio-Metabochip [62] (build 37.1) for 1,552 KORA subjects that also have the genome-wide SNP panel available. 54,599 SNPs (including 4,133 SNPs on chromosome 1) were present in both the imputed and the Metabochip typed data. For each SNP and person, imputed genotypes rounded to the nearest integer were compared to the Metabochip typed genotypes. This comparison was summarized, for each approach separately, via a 3x3 misclassification matrix (m_{ij}, ij=0,1,2) and by computing the concordance as the sum of the cells on the diagonal, m_{ii}, i=0,1,2, divided by the total number of genotypes (4,133 SNPs times 1,552 subjects = 6,414,416 genotypes).

Imputation quality was quantified using RSQ [32], which was computed for each of the imputed SNPs in the KORA data from each of the two approaches (n=1,644, #SNPs= 30,061,897, including 2,300,365 SNPs on chromosome 1). For detailed presentations, the CFH locus [63] on chromosome 1 (chr1:196,594,399 to 196,718,099 on b37.1, containing 1,428 SNPs) was chosen. This region is selected because it contained the strongest locus for age-related macular degeneration and one of the strongest genetic associations overall, the CFH locus [8].

I have implemented the novel lift-over approach in a software framework, *PhaseLift*, which includes R-, shell- and python-scripts and uses the remapping file *RsMergeArch* for lifting rs-names and the over.chain files for lifting positions (**see Web Resources**). *PhaseLift* is available on www.epi-regensburg.de/download/PhaseLift.

To highlight the extent of the harmonization problem, I provided a list of different builds released during the last five years (**Appendix 7.5.1**). I also summarized the released reference panels within builds b36.3 to b37.1 (**Appendix 7.5.2**). All reference panels not only include additional subjects, but always involve an increased number of SNPs. Specifically, when comparing b37.1 to b36.3 using the 17,479,168 SNPs contained in both data sets (b36.3: b130_SNPChrPosOnRef_36_3.bcp, b37.1: b131_SNPChrPosOnRef_37_1.bcp, **see Web Resources**), I noticed 324,866 fewer SNPs and 238,400 SNPs re-located to other chromosomes; nearly all SNPs changed their positions (**Appendix 7.5.3**). These SNPs with relative position change, and the offset of positions in a smaller region are illustrated in **Appendix 7.3**.

2.3.3 High concordances of imputed and directly typed genotypes

I imputed the KORA data twice, with the pre-phasing lift over and with the post-phasing lift-over approach. I then compared the concordances of the imputed genotypes with the Metabochip typed genotypes on the example of chromosome 1 (n=1,552, #SNPs=4,133, see Methods). I found comparable concordances for the pre- and the post-phasing approach (93.1% and 93.6%, respectively,

Table 14). The concordances were also similar between the two approaches when separated by categories of imputation quality (76.1 – 97.3% compared to 78.7 – 97.4%) or by categories of MAF (96.7 – 91.3 compared to 96.8 – 91.9). The comparability of the two approaches can be also seen using the root mean squared error (RSME, i.e. standard deviation of difference between genotype and dosage) averaged across SNPs (0.11 versus 0.11).

Table 14. ***Concordances of imputed genotypes with Metabochip typed genotypes.*** *For 4,133 SNPs on chromosome 1 in the KORA data (n=1,552) that have been imputed based on the genome-wide SNP panel and typed with the Metabochip, concordances and mean (standard deviation) of the root mean squared error (RMSE) between imputed and typed genotypes are stated by categories of RSQ and MAF.*

	#SNPs	Concordance (%) pre-phasing approach	Concordance (%) post-phasing approach	Mean RSME (SD) pre-phasing approach	Mean RSME (SD) post-phasing approach
0<=RSQ<=0.3	260	76.07	78.67	0.91 (0.50)	0.84 (0.50)
0.3<RSQ<=0.8	1,221	87.66	88.39	0.33 (0.24)	0.31 (0.23)
0.8<RSQ<=1	2,652	97.35	97.37	0.08 (0.12)	0.08 (0.12)
0<=MAF<=0.05	494	96.70	96.82	0.22 (0.33)	0.21 (0.32)
0.05<MAF<=0.2	1,529	94.48	94.68	0.17 (0.20)	0.17 (0.20)
0.2<MAF<=0.5	2,110	91.35	91.93	0.18 (0.18)	0.18 (0.17)
ALL	4,133	93.14	93.55	0.11 (0.12)	0.11 (0.12)

RSQ = imputation quality, MAF= minor allele frequency. RSQ and MAF from the pre-phasing approach imputed data.

2.3.4 Comparable imputation quality for both approaches

I was also interested to know whether the imputation qualities of the post-phasing lift-over approach are comparable to the imputation qualities of the pre-phasing lift-over approach. I therefore compared the RSQ values of the KORA-F3 data, which was imputed with each approach – again exemplified on chromosome 1 (n=1,644, #SNPs=2,300,365).

The proportion of well imputed SNPs (RSQ >0.8) and medium well imputed SNPs (0.3 ≤ RSQ = 0.8), which are usually included into GWAS, is comparable across the pre- and the post-phasing approach with 21.11% and 21.61%, respectively (**Table 15**). This proportion decreased with MAF in a comparable fashion for both approaches.

Table 15. ***Distribution of imputation quality.*** *Shown are absolute and relative frequencies of badly (RSQ ≤ 0.3), medium (0.3 < RSQ ≤ 0.8) and well imputed SNPs (0.8 < RSQ) in total and by categories of MAF (less common and rare: MAF ≤ 0.05, common: 0.05 < MAF ≤ 0.2, very common: 0.2 < MAF) in the 2,300,365 imputed SNPs of chromosome 1 in the KORA data (n=1,644).*

MAF	RSQ	Pre-phasing approach	Post-phasing approach
Total	RSQ≤0.3	1,427,895 (62.07%)	1,407,959 (61.21%)
	0.3<RSQ≤0.8	386,889 (16.82%)	395,282 (17.18%)
	RSQ>0.8	485,581 (21.11%)	497,124 (21.61%)
MAF<0.05	RSQ≤0.3	1,392,904 (77.35%)	1,376,371 (76.44%)
	0.3<RSQ≤0.8	284,103 (15.78%)	296,142 (16.45%)
	RSQ>0.8	123,661 (6.87%)	128,155 (7.12%)
0.05≤MAF≤0.2	RSQ≤0.3	22,362 (9.53%)	20,542 (8.75%)
	0.3<RSQ≤0.8	55,896 (23.81%)	54,960 (23.41%)
	RSQ>0.8	156,505 (66.67%)	159,261 (67.84%)
MAF>0.2	RSQ≤0.3	12,629 (4.77%)	11,046 (4.17%)
	0.3<RSQ≤0.8	46,890 (17.7%)	44,180 (16.68%)
	RSQ>0.8	205,415 (77.53%)	209,708 (79.15%)

RSQ = imputation quality, MAF= minor allele frequency. RSQ and MAF from the pre-phasing approach imputed data.

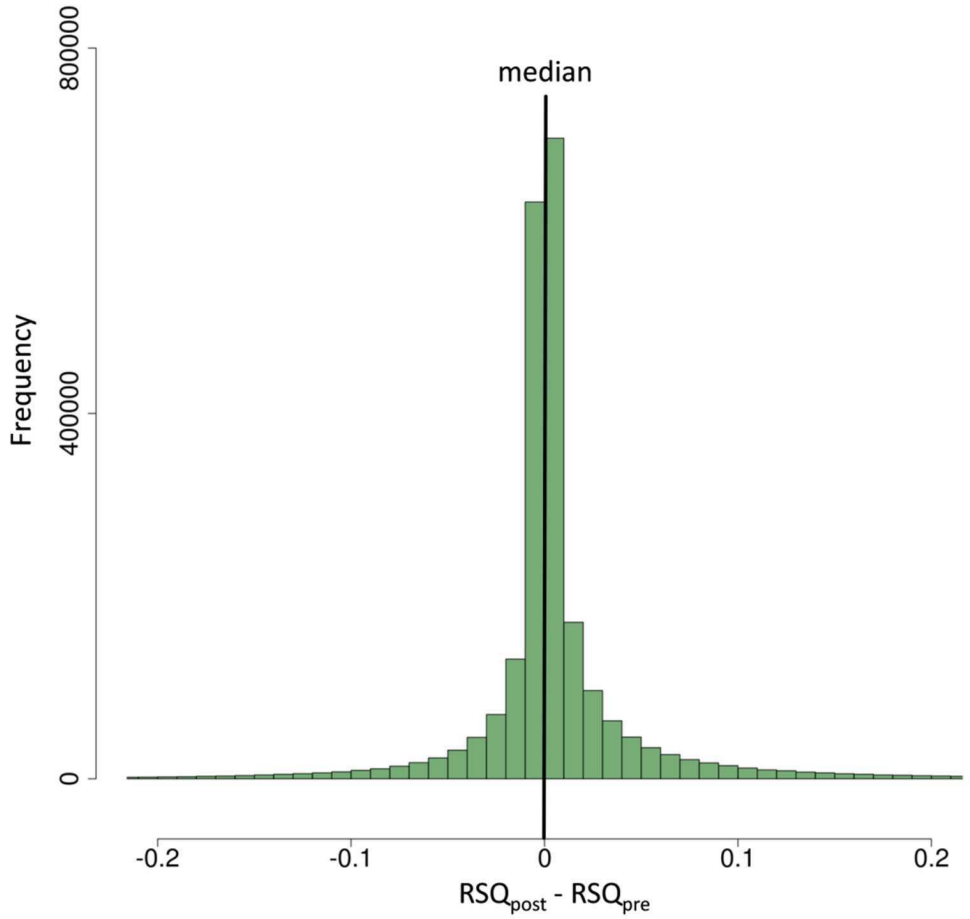

Figure 10. ***Distribution of the difference in imputation qualities between post-phasing and pre-phasing approach****. Shown is the frequency distribution of the difference of RSQs for 2,300,365 SNPs imputed with the post- and the pre-phasing approach (RSQ_{post}, RSQ_{pre}) from chromosome 1 for the KORA data (n=1,644).*

I visualized the distribution of the difference in the RSQ between the post- and the pre-phasing approach (RSQ_{post} – RSQ_{pre}) and found a symmetric distribution indicating comparable imputation qualities (mean = 0.007, sd = 0.079, **Figure 10**). Again, the symmetry remained for different categories of MAF (MAF ≤ 0.05: mean = 0.006, sd = 0.084; 0.05 < MAF ≤ 0.2: mean = 0.008, sd = 0.051; MAF > 0.2: mean = 0.001, sd = 0.065). I also found this symmetry consistently in 16 regions of 12.5 Megabases covering chromosome 1 (**Appendix 7.5.4**).

To obtain a more refined view, I further evaluated RSQ values for SNPs in one of the 16 regions (chr1:196,594,399 to 196,718,099 on b37.1, containing 1,428 SNPs). When comparing the imputation quality of the two approaches (**Figure 11a**), I found little difference for common or very common SNPs, but obvious differences for rare variants. To evaluate whether there were regional patterns for these differences, I visualized the difference of the RSQs against the chromosomal positions (**Figure 11b**). I find regional signals in both directions. When zooming into a 130 kb region directly containing the CFH gene locus (**Figure 11c**), I observed regions with little difference (around the indicated SNP rs1061170) between the RSQ of the post- and the pre-phasing approach as well as regions with large difference (around rs424535).

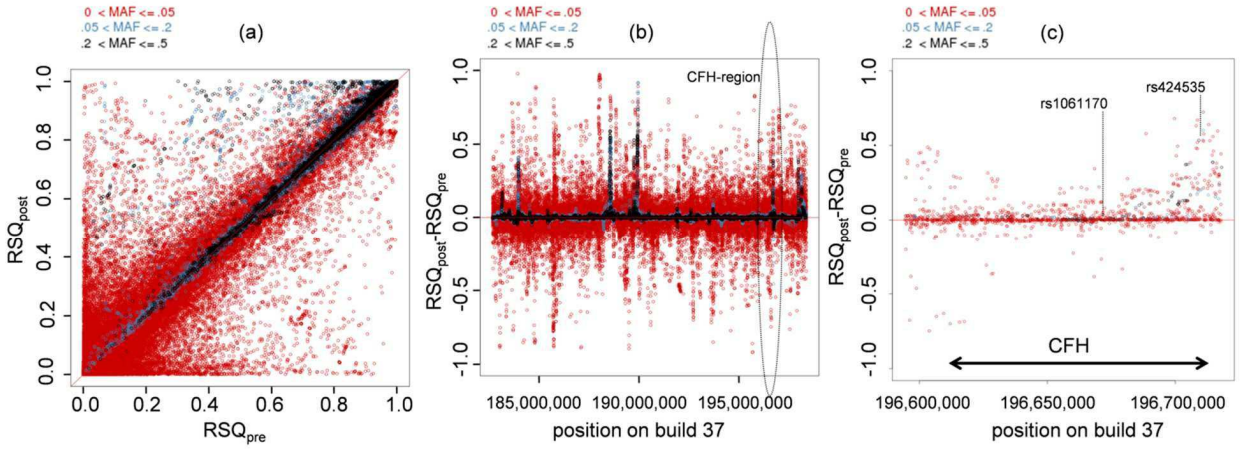

Figure 11. ***Comparison of the imputation quality between pre-phasing and post-phasing approach.*** *Shown are the RSQ values for 173,000 SNPs from a chromosome 1 region (chr1:183,000,000–198,000,000) that were imputed with the post- and the pre-phasing approach (RSQ_{post}, RSQ_{pre}) based on a genome-wide SNP panel in the KORA data (n=1,644). Color codes rare variants (red, 0 < MAF ≤ 0.05), common variants (blue, 0.05 < MAF ≤ 0.20) and very common variants (black, 0.20 < MAF ≤ 0.05). (a) RSQ_{post} versus RSQ_{pre}, (b) difference of RSQ_{post} and RSQ_{pre} versus chromosomal position, and (c) a detailed view of (b) for the CFH gene*

I was also interested whether a similarity or difference in the imputation quality (RSQ) between the two approaches also reflected similarity or difference in individuals' dosages. **Appendix 7.5.4** shows three SNPs from the CFH region example. It can be seen that perfect imputation quality (rs1061170: RSQ_{post}=0.99, RSQ_{pre}=0.99) translates also to nearly perfect agreement in dosages, while differences in the imputation quality (rs424535: RSQ_{post}=0.99, RSQ_{pre}=0.47; rs10915847: RSQ_{post}=0.51, RSQ_{pre}=0.92) translate in substantially differing dosages (**Appendix 7.5.4**).

2.3.5 Time saving by the novel post-phasing approach

Of major interest is the efficiency in terms of computation time. I compare computation time (**Table 16**) needed for the overall imputation process consisting of harmonizing, phasing, and imputing (pre-phasing approach) or phasing, harmonizing, imputing (post-phasing approach). This was performed using reference panels of different sizes (**Appendix 7.5.2**). This demonstrated that there is no difference in computation time between the two approaches for the first imputation of the study data, which was due to the fact that, in this case, both the phasing and the imputing has to be conducted. In contrast, the time savings using the post-phasing approach were considerable, when the study data is re-imputed using a newer reference panel and increases with each re-imputation. This was due to the fact that the already available phased study data can be used; only harmonization and imputing had to be repeated. Per re-imputation, the example demonstrates time savings of nearly one month (23-28 days depending on reference data) in a realistic scenario of parallelized computing on eight cores. This results in time savings of 21% (132 vs. 104 days) or 28% (197 vs. 142 days) for one re-imputation or two re-imputations, respectively.

Table 16. ***Comparison of computation time for the full imputation process.*** *Stated are the days of computation time for the full imputation process including harmonizing, phasing and imputing (pre-phasing approach) or phasing, harmonizing, and imputing (post-phasing approach). It should be noted that the pre-phasing approach requires re-phasing when repeating the imputation process using a newer reference panel. On the other side, the post-phasing approach can use the already available phased study data, re-harmonize on the phased level, and then re-impute. Thus, the omission of the computational intensive phasing results in the substantial computing time savings by the novel approach. The example given here is based on phasing and imputing the genome-wide SNP panel in the KORA data (n=1,644, SNPs=490,033 before imputation) for several reference panels. Analysts' time to process data is not taken into consideration. The imputation is processed on an eight-core-cluster. Mach is applied for phasing; Minimac is applied for imputing using ChunkChromosome with a core size of 2,500 SNPs and overlap of 500 SNPs.*

Reference panel	Average time for multiple imputations [days]					
	Number of imputations					
	1		2		3	
	pre-phasing approach	post-phasing approach	pre-phasing approach	post-phasing approach	pre-phasing approach	post-phasing approach
HAPMAP 2 (CEU)	28	28	56	28	84	28
1000G Phase I v2 (EUR)	40	40	79	51	118	63
1000G Phase I v3 (GIANT ALL)	66	66	132	104	197	142

Table 16 also highlights how the computation time increases with the size of the reference data. This underscores a further increase of computation time with newer and bigger reference data and thus an aggravation of the computing time issue in the near future.

2.3.6 Summary

Genome-wide association meta-analyses integrating imputed study data have been one of the most successful approaches to uncover the genetic basis of diseases. However, the computing time for study data imputation is steadily increasing as larger studies, larger SNP panels, and larger reference data become available. Also, study analysts are frequently facing requests for re-imputation to the latest reference data. Therefore, approaches to reduce computation time are urgently needed to meet the challenges and facilitate current and future meta-analyses.

Here, I presented a novel approach for harmonizing study data with the reference data utilized for imputing genotypes. This is an important step in the imputation process as perfect alignment of study and reference data is essential for successful imputation. My approach makes use of the idea that the harmonization is performed on a haplotype level rather than a genotype level, which avoids re-phasing when updating the imputation to a newer reference panel. I showed that my novel approach yields equally well imputed data compared to the current approach, while substantially decreasing computing time. When re-imputing study data to a newer reference panel, I estimated that up to one month in computing time can be saved by applying the novel approach.

While I saw on average equally well imputed data comparing the novel post-phasing approach to the pre-phasing approach, I found differences in imputation quality when looking at single variants. These differences were as high as perfectly imputed by one approach to very badly imputed by the other. On the one side, it is understandable when the pre-phasing approach is better: If a variant position or allele information is erroneous in the study file, it is preferable to omit this variant before phasing (as done in the pre-phasing approach). On the other side, there are usually a number of SNPs in the study file that are manufacturer-specific that are lacking in the reference file (e.g. 10,500 SNPS in a study genotyped with Affymetrix 500K array on build 36 and lacking in the reference panel 1000G Phase I v3 on build 37). These SNPs would be omitted by the pre-phasing approach before phasing. However, these SNPs can be informative during the phasing, due to their linkage disequilibrium with surrounding SNPs. The post-phasing approach keeps these variants during the phasing where they can improve imputation quality of surrounding variants. Of note, for both approaches, these variants will be omitted by the imputing step (as they are not in the reference panel), but the general route of data cleaning would re-enter such SNPs to the dosage file as genotypes after finalized imputation.

I expect that the study analysts will be more inclined to update the imputation of their study data when less computation time is involved. Therefore, my approach is clearly to be recommended as it will ultimately lead to more frequently updated imputed data and thus to improved imputed GWAS data. I anticipate that re-imputation will become even more important given the current sequencing efforts which are going to yield more and larger reference panels in the near future. Furthermore, re-imputation will be of special interest for rare variants, which are in the focus of

current analyses. While there is some work on imputation quality, comparison of different imputation software, or combining reference data for imputation [64-66], a systematic approach to lift-over and harmonization between study and reference data has not been addressed before. I provided a concept for this lift-over and harmonization process and outline how this is to be implemented.

As a summary I developed a novel approach of how to impute study data for multiple genotype imputations with different reference panels, which can be annotated on different builds to facilitate meta-analyses of several thousands of individuals and finally to help identifying susceptibility loci for complex disease.

3 On the gain of mega-imputing and mega-analyzing compared to meta-imputing and meta-analyzing individual participant data

Meta-analyses of several thousands of individuals from different studies have been very successful in identifying genetic loci associated with complex diseases [13]. The previous chapter showed, that additional genetic loci can be identified with 1000 Genomes meta-analysis compared to HapMap meta-analysis. Another approach of analyzing genome-wide data of a large number of subjects from different studies is the mega-analysis of individual participant data (IPD). IPD can be imputed jointly (including joint phasing and joint imputation, mega-imputation) or imputed separately by study (including separated phasing and separated imputation, meta-imputation). The imputed genotypes can then be analyzed jointly (mega-analysis) or separately by study (meta-analysis). Although IPD was analyzed genome-wide in mega-analyses in the past years [67, 68], there are no analyses on the influence of mega-imputation, meta-imputation, mega-analysis and meta-analysis on the ability to detect disease loci with complex disease. AMD serves me as example for comparing results from mega-imputing and mega-analyzing with meta-imputing and meta-analyzing several thousand subjects. My analyses are exemplified on data from the International Age-related Macular Degeneration Genomics Consortium (**IAMDGC)**.

In **chapter 3.1** I introduce the IAMDGC, the conducted analyses and measure to quantify the gain between two analyses. My aim in **chapter 3.2** is to analyze the gain in imputing study genotypes jointly (mega-imputation) and analyzing the imputed genotypes jointly (mega-analysis) compared to imputing study genotypes separated by study (meta-imputation), analyzing the imputed study genotypes separated by study (performing GWAs per study) and meta-analyzing the summary statistics from each GWAS. As the genotype imputation on the joint data set is computational demanding task, I show in **chapter 3.3** how genotype imputation can be parallelized and how time can be saved in a realistic scenario, using the data from the IAMDGC.

3.1 Methods and Material

In this section, first the phenotype AMD and the data from the IAMDGC is introduced. Next, it is shown how the consortium data set is analyzed and how the gain of mega-imputing and mega-analyzing versus meta-imputing and meta-analyzing this data set is quantified. Finally it is explained how to calculate power to detect susceptibility loci for binary traits.

3.1.1 Age-related Macular Degeneration

AMD is the leading cause of blindness in the elderly, which causes irreversible loss of central vision, affecting more than 10% of subjects over age 80 [69]. The heritability of AMD is estimated to vary between 46% and 70% [2]. Known risk factors for AMD are high age, smoking and sun exposure [70]. Advanced AMD is classified as wet AMD with choroidal neovascularization (CNV), when accompanied by angiogenesis or as dry AMD with geographic atrophy (GA), when angiogenesis is absent or as both CNV and GA. The genetic map of AMD was investigated for several years. The CFH locus was one of the first identified susceptibility loci for a complex disease, which was discovered by a genome-wide analysis [8]. Today, analyses of common variation have uncovered overall 21 risk loci for AMD [17, 46-51]. In the latest analysis of the International Age-related Macular Degeneration Consortium (IAMDGC, see **Web Resources**) overall 34 loci with a genome-wide wide significant lead variant (with smallest p-value in the locus) could be identified. These susceptibility loci of AMD contain variants with a wide spectrum of allele frequencies and effect sizes.

3.1.2 The IAMDGC data

The IAMDGC data consists of 26 studies and more than 50,000 subjects. These subjects were genotyped centrally with a custom-modified HumanCoreExome array by Illumina, Inc., which includes variants across the whole genome, protein-altering variants, and custom chosen variants, which were identified as susceptibility loci for complex disease by previous analyses on AMD. The genotypes and phenotypes of all subjects were aggregated into one Individual Participant Data set (IPD).

3.1.2.1 Study subjects

Overall, 52,189 unrelated subjects from 26 studies in the IAMDGC data across all ethnicities (European, African, Asian and Others) passed initial subject quality control: The data include all subjects with sufficiently high call rate (>98.5%), no discrepancies between reported gender and sex and without atypical sex chromosome compositions. Among all 52,189 subjects that pass quality control there are 16,144 AMD cases and 17,832 controls of European ancestry. The Rotterdam study included no cases and was pooled with the NHS_HPF study. A summary of the AMD status of the subjects per study, which passed quality control can be found in **Table 17**.

3.1.2.2 Genotypes

Overall, 569,645 variants were genotyped with the HumanCoreExome array in all subjects at the Center for Inherited Diseases Research (CIDR), Johns Hopkins University School of Medicine. Variants with low call rates (<98.5%), deviations from Hardy-Weinberg equilibrium with $P < 1x10^{-6}$ or mapped to multiple locations were excluded. After quality control 508,740 autosomal variants were used for the analyses. Genotypes of 4,210 subjects (from the studies BDES, Columbia, NHS_HPF and Columbia) have been derived by amplified DNA (Whole genome amplification, WGA).

3.1.3 Analysis workflow comparing meta-imputation and meta-analysis versus mega-imputation and mega-analysis of the IAMDGC study data

The subjects from the IAMGC were analyzed jointly (mega-approach) and separately (meta-approach): The **mega-imputation** (including joint phasing and joint imputing) was conducted on all 52,189 unrelated subjects. **The mega-analysis** was conducted on all 16,144 cases and 17,832 controls in one data set. The **meta-imputation** was conducted on all 52,189 unrelated subjects, separated by the 25 studies (including both phasing and imputation separated by study). **The meta-analysis** was conducted in the 16,144 cases and 17,832 controls separated by the 25 studies. The association results of those 25 studies were pooled using an inverse variance weighted method.

3.1.3.1 Genotype imputation

Identical settings were used for mega-imputation and meta-imputation for phase estimation and genotype imputation: Study haplotypes were inferred on whole chromosomes with shapeit.v2.r727.linux.x64 (states = 200 and window size = 2.5 Megabases). Genotype imputation was conducted with minimac-omp_2013_7_17 (rounds = 5) in regions, each spanning 2.5 Megabases in the genome with an overlap of 500 Kilobases to each side. For both mega-imputation and meta-imputation the 1000 Genomes Phase I version 3 (compare **chapter 2.1.2**) reference panel was used.

Table 17. ***Studies participating in the IAMDGC****. Shown are the 25 studies from all ethnicities (European, African, Asian and others) and Europeans only. The subjects from all ethnicities were used for the mega-imputation and meta-imputation, the unrelated Europeans were used for the meta-analysis and mega-analysis.*

	All ethnicities										Unrelated Europeans					
	Advanced AMD			Inter-		Un-	AMD<	Large			Advanced AMD					
Study	Mixed	CNV	GA	mediate	Controls	known	50years	Drusen	WGA	Total	Mixed	CNV	GA	Controls	WGA	Total
AREDS	229	1,202	453	868	343	3	0	1,617	0	4,715	224	1,172	445	317	0	2,158
BDES	46	46	44	134	911	47	25	466	141	1,719	41	42	39	787	53	909
Cambridge	127	579	140	0	422	0	9	0	0	1,277	127	579	139	419	0	1,264
Cologne	12	280	9	233	594	0	0	0	0	1,128	12	279	9	572	0	872
Columbia	81	326	90	0	560	0	9	260	1,103	1,326	79	292	86	483	741	940
CWRU	322	294	238	0	653	0	6	212	0	1,725	317	289	229	531	0	1,366
Edinburgh	0	180	44	147	196	0	7	0	0	574	0	179	44	193	0	416
EU_JHU	36	549	271	184	715	106	7	132	0	2,000	35	505	252	597	0	1,389
Jerusalem	0	306	1	10	207	0	0	44	0	568	0	282	1	178	0	461
Marshfield	169	15	31	0	3,273	566	127	646	0	9,595	148	10	26	2,710	0	2,894
Melbourne	51	459	68	0	412	0	5	0	0	995	49	439	67	403	0	958
Miami	71	551	124	218	406	0	6	180	0	1,556	67	518	113	358	0	1,056
Michigan	192	355	174	0	637	0	0	98	0	1,456	192	351	174	627	0	1,344
NHS_HPF/ Rotterdam	10	257	16	261	1,030	0	21	0	1,466	1,595	10	255	13	1,012	1,166	1,290
Oregon	69	435	167	0	279	0	0	0	0	950	68	425	162	272	0	927
Penn	96	460	149	73	886	3	11	361	0	2,039	88	396	133	409	0	1,026
Pitt	217	282	130	6	142	132	21	158	0	1,088	209	254	115	130	0	708
Regensburg	366	939	371	37	1,161	0	2	126	0	3,002	365	935	367	1,148	0	2,815
Southampton	17	258	65	56	615	1	2	117	0	1,131	17	253	63	580	0	913
UCSD	73	1,047	172	157	2,076	0	7	0	0	3,532	72	1,014	171	1,893	0	3,150
UMCN	12	317	35	245	477	0	10	0	0	1,096	12	312	35	427	0	786
Utah	6	730	68	88	1,347	0	3	194	1,500	2,436	6	310	18	136	228	470
UWA/LEI/Flinders	2	1,049	443	0	2,586	1	0	0	0	4,081	2	1,013	419	2,409	0	3,843
Vanderbilt	0	392	79	35	575	0	1	161	0	1,243	0	385	77	519	0	981
Westmead/Sydney	24	301	43	0	800	3	9	182	0	1,362	20	260	38	722	0	1,040
										52,189	2,160	10,749	3,235	**17,832**		**33,976**
												16,144				

3.1.3.2 *Association analysis*

Imputed genotypes from the meta-imputation and mega-imputation were analyzed with the Firth corrected logistic regression analysis on AMD status. It included DNA source and the first two ancestral principal components as covariates. Mega-analysis (on all subjects) and meta-analysis (per study) were conducted. The regression results of the 25 studies were pooled using an inverse variance weighted method with metal [21].

Let Y_{ik} be the outcome, G_{ik} the genotype, $C_{ik}^{(m)}$ several covariates (m=1,…,M), k an index for study (k=1,…,K), i an index for individual, i=1,…, n_k, and the sum of n_k = n. The statistical model for the mega-analysis assuming constant α_k and $\gamma_k^{(m)}$ across studies is:

$$f^{-1}(E[Y_i]) = \alpha + \beta G_i + \sum_{m=1}^{M} \gamma^{(m)} C_i^{(m)}$$

The mega-model includes the 5 parameters α, β, and the three covariates (two principal components and the DNA source).

The statistical model for the meta-analysis makes no assumptions, except that it assumes fixed genetic effects across studies, combining the β_k's:

$$f^{-1}(E[Y_{ik}]) = \alpha_k + \beta_k G_{ik} + \sum_{m=1}^{M} \gamma_k^{(m)} C_{ik}^{(m)}$$

The number of parameters in the meta-model is 101 (k-times α, k-times the three covariates and one for the β per variant across all subjects: 25*4 + 1).

A potential gain by the mega-analysis can derive from fewer parameters but involves model assumptions, which might cause bias or violate the type I error.

3.1.4 Measures to quantify the gain between mega-imputation and mega-analysis compared to meta-imputation and meta-analysis

3.1.4.1 *Comparing imputation quality*

Quality of genotype imputation of imputed variants was quantified with the imputation quality *RSQ* from *minimac*. The metric RSQ is the proportion of the observed (imputed) genotypes compared to the expected variance, if the genotype would be imputed without error. The imputation quality in the mega-imputation was compared to the median imputation quality across all variants of the median imputation qualities across the 25 studies from the meta-analysis.

3.1.4.2 *Power for binary traits*

Post-hoc power analysis on the estimated effect was conducted to evaluate the power to detect variants with genome wide significance in the analysis of AMD. Power estimation is calculated for two proportions with different sample sizes with a two-tailed test at the genome wide significance level of

$5x10^{-8}$ with Cohen's Power calculation for two proportions [71]. The Null-Hypothesis, that the odds in cases and controls are equal or equivalent, that the odds ratio is one (H_0: OR = 1) is tested against the alternative hypothesis, that the odds ratio is different from one (H_A: OR ≠ 1). The calculations were conducted with the R-package 'pwr' (see **Web Resources**).

3.1.4.3 Evaluation of odds ratios, standard errors and p-values

The odds ratios, standard errors and p-values were compared between the mega-imputed and mega-analyzed IAMDGC data with the original results from the IAMDGC and two other approaches: First, they were compared with the odds ratios, standard errors and p-values from the mega-imputed and meta-analyzed data and second, they were compared to the odds ratios, standard errors and p-values from the meta-imputed and meta-analyzed data. The odds ratios, standard errors and p-values were compared by scatterplots, where each point represents one variant, which was present in both analyses. The variants were color coded by the MAF in the mega-analysis (green: MAF>10%; orange: 1% ≤ MAF ≤ 10% and red: MAF < 1%). The odds ratios were compared between different analyses to examine, if the effects were biased. The standard errors and p-values were compared between different analyses to evaluate the influence of the different models used in the analyses. The differences were exemplified on chromosome 5, but can be generalized to the whole genome.

3.1.4.4 Signal detection

Signals detection was conducted to identify susceptibility loci for AMD at the genome wide significance level of $5x10^{-8}$. The association p-values of the lead variants in the 34 AMD disease loci were compared between the mega-imputed and mega-analyzed IAMDGC data with the original IAMDGC results from the IAMDGC and two other approaches: First, they were compared with the results from mega-imputed and meta-analyzed data. Second, the full tracks were compared: the results from mega-imputed and mega-analyzed data were compared to meta-imputed and meta-analyzed data. Lead variants in the 34 signal can change between analyses. Different lead variant between two analyses were identified and the change of the odds ratios and p-values were evaluated.

3.1.4.5 Evaluating the top variant in a locus

The odds ratio and p-value of the lead variant from the original IAMDGC analysis in the 34 AMD disease loci were evaluated in each analysis. The lead variant (variant with the smallest p-value in the locus) in the current analysis was possibly different from the lead variant from the original IAMDGC analysis. In this case, the lead variant from the original IAMDGC analysis and from the current analysis was shown.

3.1.4.6 Evaluating the type I error

The type I error was assessed in the mega-imputed and mega-analyzed, data on the genotyped variants with MAF >5%. First, the genomic inflation λ was calculated and second, the QQ-Plot was generated

to observe the type I error. The results were exemplified on chromosome 5 but can be generalized to the whole genome.

3.2 Results of the comparison between mega-imputing and mega-analyzing compared to meta-imputing and meta-analyzing the IAMDGC data

The aim of this chapter is to analyze the gain in mega-imputing and mega-analyzing genotypes compared to meta-imputing and meta-analyzing them. This comparison was exemplified on the IAMDGC data. First the lead variants in the 34 AMD risk loci identified by the original IAMDGC analysis were compared to those by mega-imputing and mega-analyzing the IAMDGC genotypes (**chapter 3.2.1**). Second, the mega-imputed IAMDGC genotypes were meta-analyzed and mega-analyzed. The results were compared to highlight the influence of the meta-analysis on the association results (**chapter 3.2.2**). The mega-imputed IAMDGC genotypes were compared to the meta-imputed IAMDGC genotypes to analyze the differences between the two imputation approaches (**chapter 3.2.3)**. The influence of the imputation scores on the power to detect disease loci was subject in **chapter 3.2.4**. The results of the meta-imputed and meta-analyzed IAMDGC data was compared with the results of the mega-imputed and mega-analyzed IAMDGC data to evaluate the differences between the classical meta-approach in consortia with the mega-approach as applied in the IAMDGC (**chapter 3.2.5**). Finally, the type I error rate in the meta-imputed and meta-analyzed IAMDGC data was investigated in **chapter 3.2.7**).

3.2.1 Susceptibility loci previously identified by the IAMDGC mega-analysis

The IAMDGC identified 34 independent genetic loci associated with AMD [17]. Among them are common and less common variants (in *COL8A1* and *ACAD10*) and one rare lead variant (in gene *C9)*. They include 29 SNPs, two insertions and two deletions. These loci were identified in the original IAMDGC analysis, by a sequential forward selection (SFS) approach on mega-imputed and mega-analyzed genome wide data from the IAMDGC. In the sequential forward selection approach first single variant association was computed for all imputed variants. Then the variant with the smallest p-value and its flanking ±5 Megabases region was selected, repeating the process until no genome-wide significant variant ($P \leq 5 \times 10^{-8}$) was left yielding a number of 10 Megabase regions. Within each of these large regions each variant was re-analyzed conditioning on the top variant. This was repeated by adding the previously identified genome-wide significant variant(s) within the respective 10 Megabases region. This yielded one or more independently associated genome-wide significant variant(s) per 10 Megabases region.

Table 18. ***Results from the mega-imputation and mega-analysis in the IAMDGC GWA analysis in the 34 lead variants identified by the IAMDGC SFS analysis****. Overall, 34 lead variant were identified in the IAMDGC SFS analysis. Shown are the results of these identified 34 lead variant of the IAMDGC GWA analysis. Results are sorted by chromosome and position.*

Variant ID	Chr	Position (bp)	Closest Gene	Ref./ Non- Ref. ($MAF_{controls}$)	OR	P-value
rs10922109	1	196,704,632	*CFH*	A/C(0.426)	0.38	**$1.7x10^{-615}$**
rs11884770	2	228,086,920	*COL4A3*	T/C(0.278)	0.90	**$2.5x10^{-8}$**
rs62247658	3	64,715,155	*ADAMTS9-AS2*	C/T(0.433)	1.14	**$1.3x10^{-14}$**
rs140647181	3	99,180,668	*COL8A1*	C/T(0.016)	1.60	**$5.9x10^{-12}$**
rs10033900	4	110,659,067	*CFI*	T/C(0.477)	1.15	**$5.7x10^{-17}$**
rs114092250	5	35,494,448	*PRLR/SPEF2*	A/G(0.022)	0.70	**$2.9x10^{-8}$**
rs62358361	5	39,327,888	C9	T/G(0.00889)	1.79	**$1.6x10^{-14}$**
rs116503776	6	31,930,462	*C2/CFB/SKIV2L*	A/G(0.148)	0.57	**$8.2x10^{-104}$**
rs943080	6	43,826,627	*VEGFA*	C/T(0.497)	0.88	**$1.1x10^{-14}$**
rs7803454	7	99,991,548	*PILRB/PILRA*	T/C(0.19)	1.13	**$5.0x10^{-9}$**
rs1142	7	104,756,326	*KMT2E/SRPK2*	T/C(0.346)	1.11	**$1.3x10^{-9}$**
rs79037040	8	23,082,971	*TNFRSF10A*	G/T(0.479)	0.90	**$2.5x10^{-11}$**
rs71507014	9	73,438,605	*TRPM3*	GC/G(0.405)	1.10	**$3.2x10^{-8}$**
rs10781182	9	76,617,720	*MIR6130/RORB*	T/G(0.306)	1.11	**$2.5x10^{-9}$**
rs1626340	9	101,923,372	*TGFBR1*	A/G(0.209)	0.88	**$3.9x10^{-10}$**
rs2740488	9	107,661,742	*ABCA1*	C/A(0.275)	0.90	**$1.7x10^{-8}$**
rs12357257	10	24,999,593	*ARHGAP21*	A/G(0.223)	1.11	**$3.9x10^{-8}$**
rs3750846	10	124,215,565	*ARMS2/HTRA1*	C/T(0.208)	2.81	**$1.5x10^{-735}$**
rs3138141	12	56,115,778	*RDH5/CD63*	A/C(0.207)	1.17	**$1.6x10^{-9}$**
rs61941274	12	112,132,610	*ACAD10**	A/G(0.018)	1.50	**$1.7x10^{-9}$**
rs9564692	13	31,821,240	*B3GALTL*	T/C(0.299)	0.89	**$3.1x10^{-10}$**
rs61985136	14	68,769,199	*RAD51B*	C/T(0.384)	0.90	**$1.5x10^{-10}$**
rs2043085	15	58,680,954	*LIPC**	C/T(0.381)	0.88	**$6.7x10^{-14}$**
rs5817082	16	56,997,349	*CETP*	CA(0.264)	0.84	**$5.1x10^{-19}$**
rs72802342	16	75,234,872	*CTRB2/CTRB1**	A/C(0.08)	0.80	**$9.7x10^{-12}$**
rs11080055	17	26,649,724	*TMEM97/VTN**	A/C(0.486)	0.91	**$1.1x10^{-8}$**
rs6565597	17	79,526,821	*NPLOC4/TSPAN10*	T/C(0.381)	1.13	**$2.4x10^{-12}$**
rs67538026	19	1,031,438	*CNN2*	T/C(0.498)	0.90	**$4.4x10^{-8}$**
rs2230199	19	6,718,387	*C3*	G/C(0.208)	1.43	**$1.6x10^{-69}$**
rs429358	19	45,411,941	*APOE*	C/T(0.135)	0.70	**$2.2x10^{-42}$**
rs142450006	20	44,614,991	*MMP9**	TTTTC/T(0.141)	0.85	**$3.1x10^{-10}$**
rs201459901	20	56,653,724	*C20orf85**	T/TA(0.07)	0.76	**$3.3x10^{-16}$**
rs5754227	22	33,105,817	*SYN3/TIMP3*	C/T(0.137)	0.77	**$9.8x10^{-25}$**
rs8135665	22	38,476,276	*SLC16A8**	T/C(0.195)	1.14	**$9.3x10^{-11}$**

Chr = Chromosome; Position (bp) is the chromosomal position given based on NCBI build 37, Ref./ Non-Ref. ($MAF_{controls}$)= reference allele, nonreference allele and MAF in controls, Closest gene is gene(s) nearest to the variant. OR = odds ratio; genome wide significant p-values are in bold. * indicates, that the lead variant in the locus identified by the IAMDGC GWAS analysis was different from the lead variant identified by the IAMDGC SFS analysis.

3.2.1.1 *Evaluating the lead variants in the 34 susceptibility loci of AMD in the mega-imputed and mega-analyzed data compared to the original IAMDGC analysis*

I repeated a regular GWA analysis testing each of the variants separately (mega-imputation and mega-analysis) and compared the odds ratios and p-values of the lead variants in the 34 susceptibility loci for AMD of my IAMGC GWA analysis (**Table 18**) with those from the original IAMDGC SFS analysis. The p-values, standard errors and odds ratios were equivalent and all 34 loci were genome-wide significant in my GWA analysis. A direct comparison of the odds ratios, standard errors and p-values were illustrated in **Figure 12**.

There were small differences repeating mega-imputation and mega-analysis due to random noise by the software. These differences were observed, although all parameters and seeds, which can be configured, were identical. Overall I conducted four mega-imputations and four mega-analyses (**Appendix 7.6)**.

In seven loci the lead variant identified by the mega-imputed and mega-analyzed IAMDGC data was different compared to the lead variant identified by the IAMDGC SFS analysis (in loci *ACAD10*, *LIPC*, *CTRB2/CTRB1*, *TMEM97/VTN*, *MMP9*, *C20orf85* and *SLC16A8*). These different lead variants were proxies (R^2>0.8) of the lead variants identified by the IAMDGC SFS analysis. The results in these seven loci were discussed in **chapter 3.2.6**. The identification of different lead variants was due to random noise in the statistical analysis of these variants. The odds ratio were equivalent in those seven lead variants and the p-values were only marginally smaller in the IAMDGC GWA analysis compared to the IAMDGC SFS analysis.

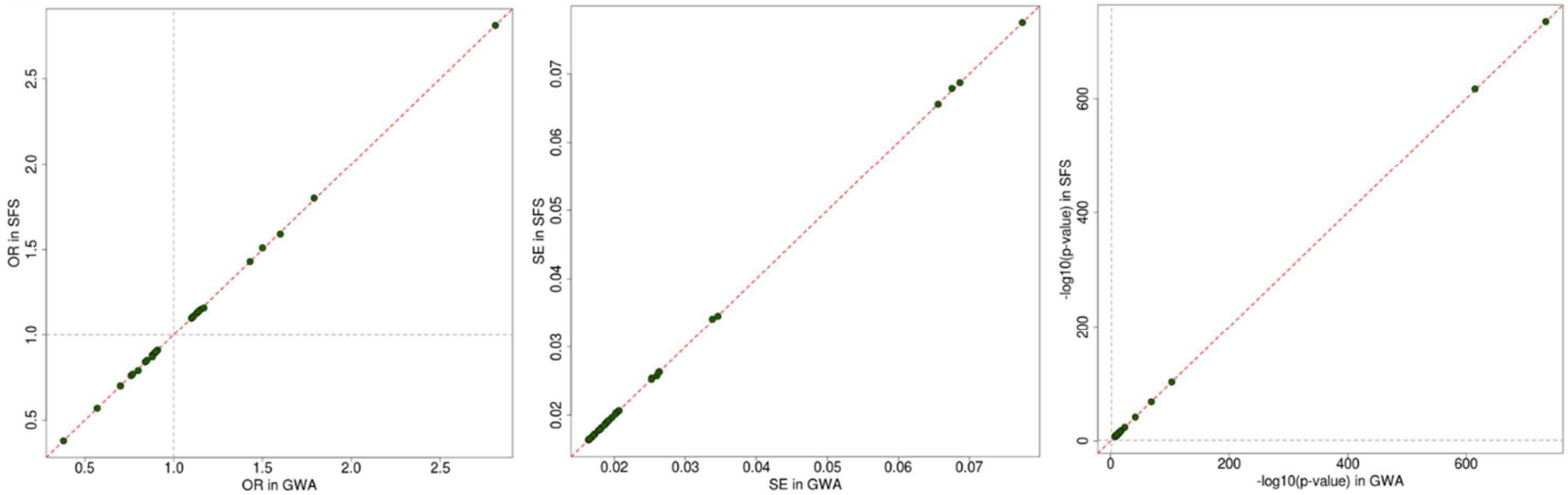

Figure 12. ***Comparison of odds ratios, standard errors and p-values between the IAMDGC GWA analysis with the IAMDGC SFS analysis in the 34 AMD disease loci.*** *The IAMDGC sequential forward selection (SFS) analysis identified 34 lead variants in the 34 AMD disease loci. Shown are the scatterplots of those lead variants comparing the results of the IAMD GWA-analysis (X-axis) with the results from the IAMDGC SFS analysis (Y-axis). Panels A, B and C compare the odds ratios (OR), standard errors (SE) and –log10 p-values, respectively.*

3.2.1.2 *Power to detect the lead variants in the 34 AMD disease loci for AMD*

Next, I was interested in the power to detect the lead variants in the 34 disease loci for AMD. For this analysis no imputation quality is considered. Post-hoc power for these 34 variants were calculated for the analysis of 16,144 cases and 17,832 controls with the estimated effect per variant. **Figure 13** visualizes, that there is more than 80% power for 28 of the 34 loci (power: min=68.68%, 25th percentile=89.00%, median=99.08%, 75th percentile=99.95%, max=100%). The estimated power for these 34 variants and the power for less frequent and common variants dependent from the estimated odds ratio can be found in **Appendix 7.7**.

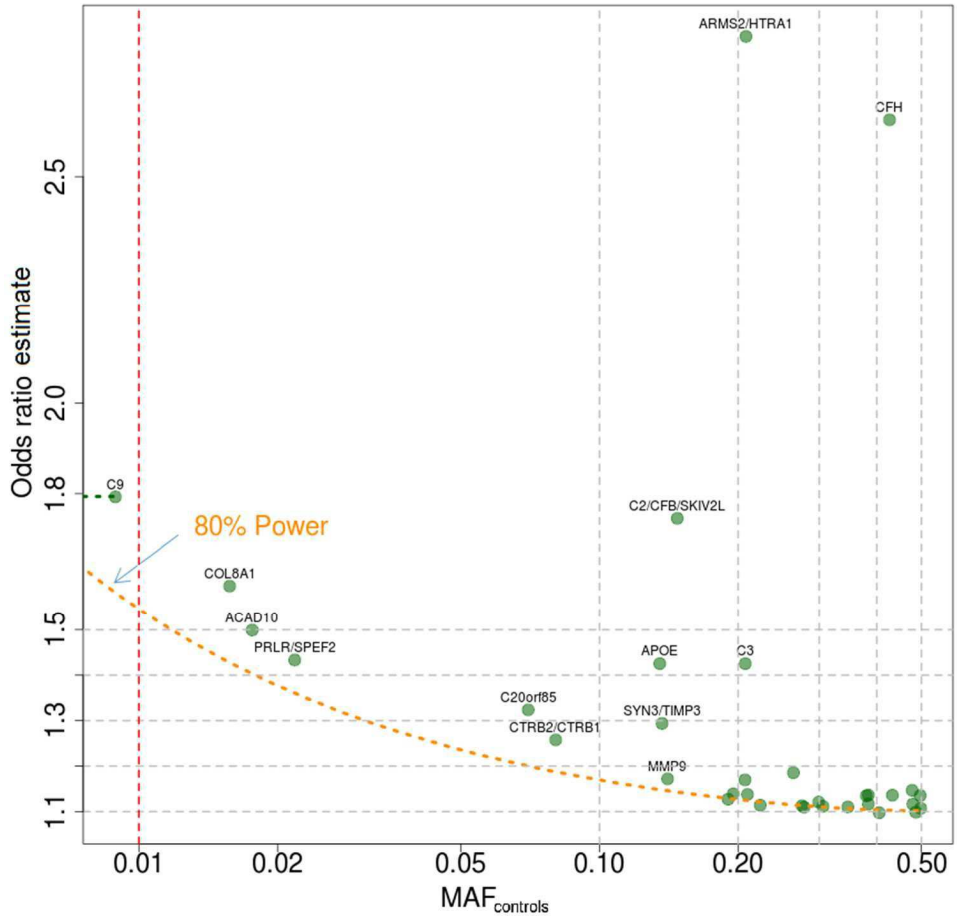

Figure 13. ***Odds ratios versus MAF in the lead variants of the 34 genetic loci associated with genome wide significance with AMD in the mega-analysis****. Shown is a scatterplot of the MAF in controls (X-axis) and the odds ratio (Y-axis). The x-axis is on a logarithmic scale. The orange curve indicates 80% power for the analysis of AMD in 16,144 cases and 17,821 controls as in the IAMDGC data. The threshold of MAF=1% for rare variants is shown as red vertical line. Imputation quality is not considered in this scatterplot.*

3.2.2 Evaluating the gain of mega-analysis compared with meta-analysis of mega-imputed genotypes

Next I was interest how much can be gained by mega-analyzing compared to meta-analyzing imputed genotypes. For this comparison the mega-imputed genotypes were used for both mega-analysis and meta-analysis.

3.2.2.1 Comparing meta-analysis with mega-analysis in the 34 susceptibility loci identified by the IAMDGC

The p-values and odds ratios of the lead variants of the 34 susceptibility loci identified by the IAMDGC were compared between meta-analysis and mega-analysis of the mega-imputed genotypes (**Table 18**, **Table 19)**. Among these lead variants identified by the IAMDGC SFS analysis, nine of the 34 signal were not genome-wide significant in the meta-analysis. The odds ratios, standard errors and p-values in these variants were illustrated in **Figure 14**. The odds ratios were comparable across the 34 variants, indicating, that there is no systematic inflation. The standard errors were higher in the mega-imputed and meta-analyzed data and the p-values were smaller in the IAMDGC GWA analysis compared to the mega-imputed and meta-analyzed data, reflecting the increased number of parameters in the meta-analysis model compared to the number of parameters in the mega-analysis model.

In 19 loci, lead variants in the mega-imputed and meta-analyzed analysis were identified, which showed smaller p-values compared to the lead variant in the IAMDGC SFS analysis. These different lead variants were proxies (R^2>0.8) to the lead variants from the IAMDGC SFS analysis. Considering also these, six of the overall 34 loci identified by the IAMDGC SFS analysis were missed by the mega-imputation and meta-analysis. These six loci were reported for the first time in the current analysis of the IAMDGC.

In summary, the meta-imputed and meta-analyzed data yielded different lead variants in the 34 AMD disease loci and missed six loci, identified in the IAMDGC SFS analysis.

Table 19. ***Meta-analysis results of mega-imputed data from the IAMDGC in the 34 AMD disease loci.*** *Shown are the odds ratios, p-values and heterogeneity in the 34 lead variants identified by the IAMDGC SFS analysis. Odds ratios, p-value and heterogeneity are shown in those loci, where the lead variant in the mega-imputed and meta-analyzed analysis was different from the lead variant in the IAMDGC SFS analysis.*

SFS variant	Closest gene	OR	P-value	I^2%	Lead variant from GWA if different	OR	P-value	I^2%
rs10922109	*CFH*	0.49	0	58.4	rs1089033	0.49	**8.62×10^{-303}**	56.6
rs11884770	*COL4A3*	0.87	2.76×10^{-07}	11.9	rs11296554	0.87	**2.99×10^{-08}**	10.4
rs62247658	*ADAMTS9-AS2*	1.12	**2.09×10^{-11}**	47	rs6775974	1.12	**1.67×10^{-11}**	46.7
rs140647181	*COL8A1*	1.14	**3.07×10^{-08}**	0	rs56339461	1.14	**1.50×10^{-08}**	0
rs10033900	*CFI*	1.13	**2.72×10^{-13}**	0	---	---	---	---
rs114092250	*PRLR/SPEF2*	0.84	4.27×10^{-05}	0	rs35559912	0.84	4.19×10^{-06}	0
rs62358361	*C9*	1.64	**2.10×10^{-09}**	0	---	---	---	---
rs116503776	*C2/CFB/SKIV2L*	0.57	**3.08×10^{-96}**	45.5	---	---	---	---
rs943080	*VEGFA*	0.88	**2.28×10^{-13}**	0	---	---	---	---
rs7803454	*PILRB/PILRA*	1.18	**3.29×10^{-09}**	11.5	rs11761306	1.18	**1.43×10^{-09}**	19.9
rs1142	*KMT2E/SRPK2*	1.09	2.75×10^{-07}	21	rs4727618	1.09	2.68×10^{-07}	0
rs79037040	*TNFRSF10A*	0.90	**2.45×10^{-10}**	0	---	---	---	---
rs71507014	*TRPM3*	1.09	1.62×10^{-06}	0	---	---	---	---
rs10781182	*MIR6130/RORB*	1.11	1.24×10^{-06}	0	rs10781159	1.11	**3.86×10^{-08}**	3
rs1626340	*TGFBR1*	0.87	**8.60×10^{-11}**	0	rs401186	0.87	**8.48×10^{-11}**	0
rs2740488	*ABCA1*	0.89	**1.26×10^{-09}**	52	---	---	---	---
rs12357257	*ARHGAP21*	1.11	6.51×10^{-07}	11	---	---	---	---
rs3750846	*ARMS2/HTRA1*	2.38	0	63.8	rs58649964	2.38	**1.70×10^{-317}**	45.8
rs3138141	*RDH5/CD63*	1.17	**3.96×10^{-09}**	0	rs56108400	1.17	**3.13×10^{-09}**	0
rs61941274	*ACAD10*	1.44	2.95×10^{-07}	29.7	rs61941272	1.44	2.89×10^{-07}	29.7
rs9564692	*B3GALTL*	0.90	**1.74×10^{-08}**	23	---	---	---	---
rs61985136	*RAD51B*	0.85	1.42×10^{-07}	22.4	rs11624933	0.85	**2.34×10^{-09}**	10.2
rs2043085	*LIPC*	0.88	**7.46×10^{-12}**	0	rs2414577	0.88	**9.54×10^{-13}**	0
rs5817082	*CETP*	0.84	**6.08×10^{-18}**	0	---	---	---	---
rs72802342	*CTRB2/CTRB1*	0.83	**4.29×10^{-08}**	0	rs55993634	0.83	**1.79×10^{-08}**	---
rs11080055	*TMEM97/VTN*	0.91	**3.29×10^{-08}**	0	rs704	0.91	**1.71×10^{-08}**	0
rs6565597	*NPLOC4/ TSPAN10*	1.12	**1.64×10^{-09}**	0	rs62075723	1.12	**1.13×10^{-09}**	0
rs67538026	*CNN2*	0.91	2.18×10^{-06}	0.1	---	---	---	---
rs2230199	*C3*	1.38	**1.38×10^{-52}**	0	---	---	---	---
rs429358	*APOE*	0.73	**1.18×10^{-31}**	26.2	---	---	---	---
rs142450006	*MMP9*	0.85	**1.42×10^{-10}**	0	rs1888235	0.85	**1.20×10^{-10}**	0
rs201459901	*C20orf85*	0.76	**2.41×10^{-14}**	0	---	---	---	---
rs5754227	*SYN3/TIMP3*	0.77	**6.35×10^{-23}**	0	---	---	---	---
rs8135665	*SLC16A8*	1.15	**1.04×10^{-09}**	29.1	rs11089861	1.15	**3.10×10^{-11}**	0

Chr = Chromosome; Position (bp) is the chromosomal position given based on NCBI build 37, Closest gene is gene(s) nearest to the variant, Ref./ Non-Ref. ($MAF_{controls}$)= reference allele, nonreference allele and MAF in controls. OR = odds ratio; I^2% is the heterogenity as reported by the meta-analysis software [%]. ”Novel” are those lead variants, which were identified in the current IAMDGC analysis for the first time, all others are ”Known”.

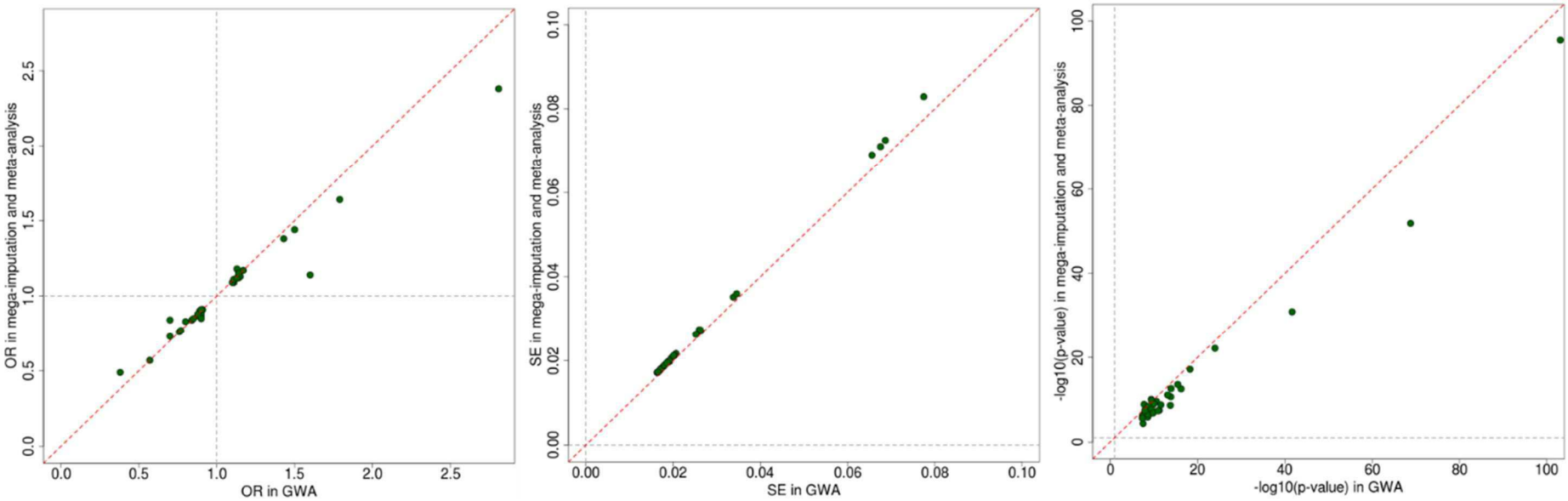

Figure 14. ***Comparison of odds ratios, standard errors and p-values between the mega-imputed and meta-analyzed IAMDGC data with the IAMDGC GWA analysis in the 34 AMD disease loci.*** *The IAMDGC SFS analysis identified 34 lead variants in the 34 AMD disease loci. Shown are the scatterplots of those lead variants comparing the results of the IAMDGC GWA analysis (X-axis) with the results from the mega-imputed and meta-analyzed IAMD analysis (Y-axis). Panels A, B and C compare the odds ratios (OR), standard errors (SE) and –log10 p-values, respectively.*

3.2.2.2 Comparing meta-analysis with mega-analysis on chromosome 5

The results from the mega-analysis were also compared to the results from the mega-analysis on chromosome 5.

The meta-analyzed genotypes comprise 1,807,108 variants and the mega-analyzed imputed genotypes yielded overall 1,800,995 with an overlap of 1,800,995 variants in both results. The 6,113 variants in the meta-analysis results but not in the mega-analysis results have a MAF ≤ $1x10^{-4}$ and are present in less than 14 studies.

Variants with a minor allele count (MAC) < 100 (equivalent with a MAF < 0.15%) were excluded, because they showed extreme differences in the estimated odds ratios between mega-analysis and meta-analysis. The models could not derive comparable results for these variants. The MAC and imputation quality per variant was determined in mega-imputed and mega-analyzed variants. Also variants with imputation quality < 0.4 were excluded, which is in line with the current practice in meta-analyses in consortia. Overall, there were 728.156 variants with RSQ ≥ 0.4 and MAC ≥ 100. Among those variants there were 163,805 rare variants (MAF < 1%), 235,195 less common variants (MAF between 1% and 10%) and 329,156 common variants (MAF > 10%).

The odds ratios, standard errors and p-values were visualized in **Figure 15A-C**. The odds ratios were comparable in both analyses: half of the variants showed an increase and a decrease, respectively (median decrease of $6.7x10^{-4}$), indicating no biased effects. The scatter of odds ratios is highest in the rare variants. The standard errors were higher in the meta-analysis: 93.5% of the standard errors were greater in the meta-analysis compared to the mega-analysis (median increase of $3.6x10^{-2}$). Approximately 62.8% of all variants showed increased of p-values and no variant were identified with genome wide significance additional to the mega-analysis. These increased p-values and standard errors reflected the loss of power due to the increased number of parameters in the underlying model of the meta-analysis compared to the mega-analysis.

The subset of 564,345 variants with MAF ≥ 1% and imputation quality ≥ 0.4 consisted of 156,920 less common and 407,425 common variants. The odds ratios, standard errors and p-values were illustrated in (**Figure 15D-F)**. Here the odds ratio were also comparable between the mega-analysis and the mega-analysis. The standard errors and p-values were higher in the meta-analysis compared to the mega-analysis indicating the loss of power in the meta-analysis compared to the mega-analysis. The p-values of the *C9* lead variant was decreased, but is genome wide significant in both mega-analysis and meta-analysis. The lead variant in the *PRLR* locus was genome-wide in the mega-analysis but not in the meta-analysis.

In summary, it is indicated, that the observed increase of standard errors and p-values between mega-analysis and meta-analysis were due to the increased number of parameters of the meta-analysis compared to the mega-analysis. No biased odds ratio were observed, as they were comparable between the mega-analysis and the meta-analysis. No genome wide significant variants were identified in the meta-analysis additional to the mega-analysis. The results exemplified with odds ratios, standard errors and p-values on chromosome 5 can be generalized to the other chromosomes as well (data not shown).

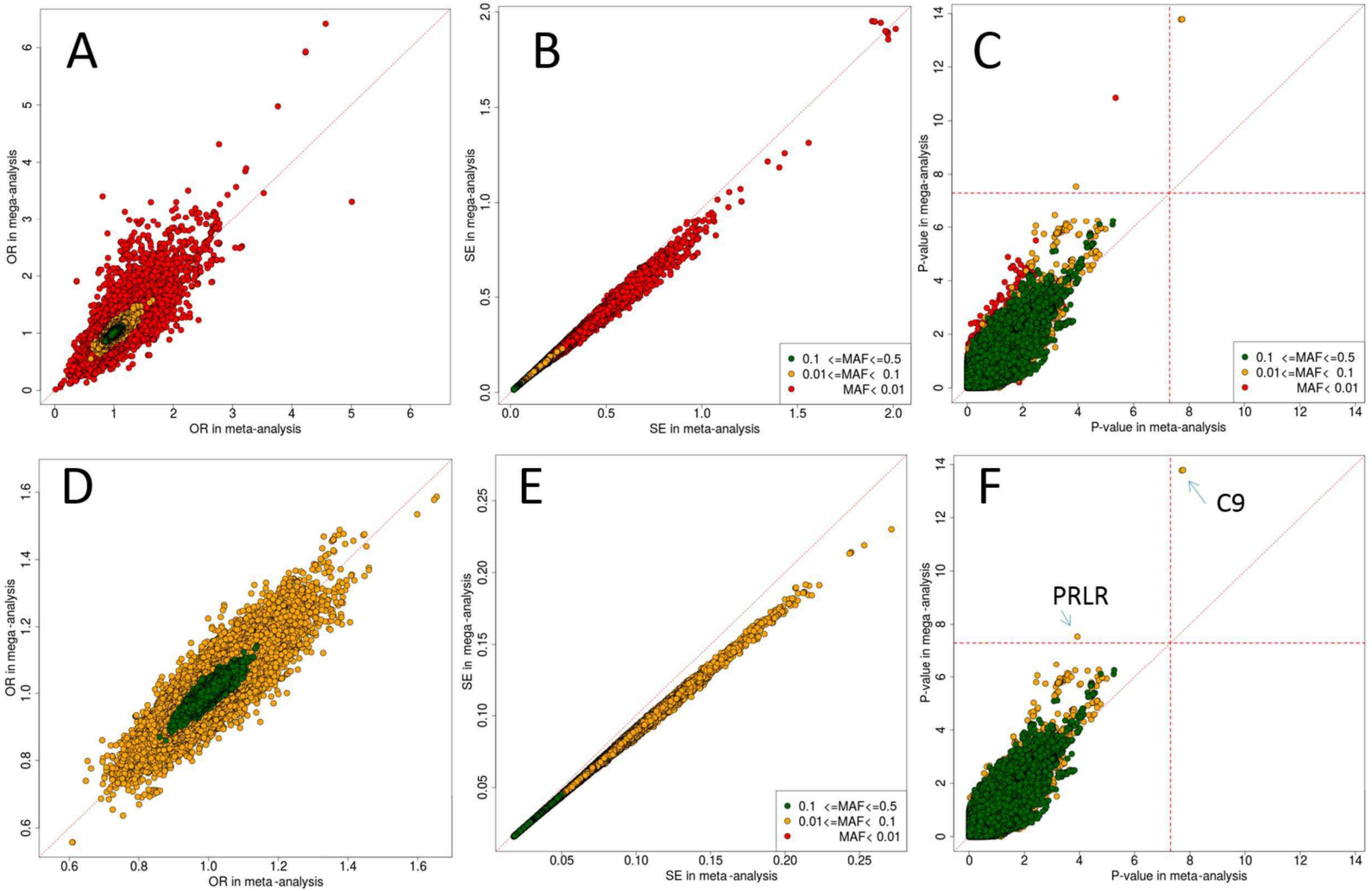

Figure 15. ***Comparison of odds ratios, standard errors and p-values between mega-analysis and meta-analysis on mega-imputed genotypes on chromosome 5.*** *Shown are the odds ratios (OR, panel A and D), standard errors (SE, Panel B and E) and p-values (Panel C and F on a negative logarithmic scale) from meta-analysis (X-axis) compared to mega-analysis (Y-axis). The dashed line in Panel C and F indicate the genome wide significance threshold Panels A, B, C show all variants on chromosome 5 with MAC ≥ 100 and imputation quality > 0.4, whereas panels D, E, F show high quality variants with MAF>1% and RSQ >0.4.*

3.2.3 Comparing imputation qualities between mega-imputation and meta-imputation

Next I was interested in the influence of the genotype imputation on the imputation quality across the whole genomes and in the lead variants in the 34 susceptibility loci of AMD.

3.2.3.1 Comparison of imputation qualities in all variants between mega-imputed and meta-imputed genotypes

The comparison between the imputation qualities is exemplified on chromosome 5. This comparison can be generalized to the other chromosomes as well (data not shown). The imputation quality of the mega-imputed variants and the median imputation quality across the 25 studies in the meta-imputation are shown in categories of MAF and imputation quality in **Table 20**.

More well imputed variants (imputation quality ≥ 0.8) were in the mega-imputation compared to the meta-imputation (36.73% vs. 28.71% in the mega-imputation and the meta-imputation, respectively). The common and less frequent variants were equally well imputed (common: 50.79% vs. 49.81%; less frequent: 87.98% vs. 87.75%). Thus the mega-imputation compared to the meta-imputation yielded no gain for common and less frequent variants. The mega-imputation yielded a gain in imputation quality in rare variants. There was a higher number of well and medium well imputed rare variants in the meta-analysis compared to the meta-analysis (17.03% vs. 5.29% and 40.68% vs. 15.63%).

*Table 20. **Distribution of imputation quality on chromosome 5**. Shown are the absolute numbers and relative frequencies of badly (RSQ < 0.4), medium (0.4 ≤ RSQ < 0.8) and well imputed variants (0.8 ≤ RSQ) in total and by categories of MAF (rare: MAF < 0.01, less frequent: 0.01 ≤ MAF ≤ 0.05 and common: 0.05 < MAF).*

MAF	RSQ	#variants by Mega-imputation	#variants by Meta-imputation
Total	RSQ<0.4	538,711 (29.94%)	986,083 (54.8%)
	0.4<=RSQ<0.8	599,839 (33.33%)	296,807 (16.49%)
	0.8<=RSQ	660,883 (36.73%)	516,543 (28.71%)
MAF<0.01	RSQ<0.4	510,385 (42.29%)	954,310 (79.08%)
	0.4<=RSQ<0.8	490,875 (40.68%)	188,599 (15.63%)
	0.8<=RSQ	205,493 (17.03%)	63,844 (5.29%)
0.01<=MAF<=0.05	RSQ<0.4	20,674 (11.64%)	23,579 (13.28%)
	0.4<=RSQ<0.8	66,719 (37.57%)	65,561 (36.91%)
	0.8<=RSQ	90,208 (50.79%)	88,461 (49.81%)
0.05<MAF	RSQ<0.4	7,652 (1.84%)	8,194 (1.97%)
	0.4<=RSQ<0.8	42,245 (10.18%)	42,647 (10.28%)
	0.8<=RSQ	365,182 (87.98%)	364,238 (87.75%)

MAF = Minor allele frequency; RSQ = Ratio of observed variance to expected variance, as reported by minimac. MAF is taken from the mega-imputed genotypes.

3.2.3.2 Comparison of imputation qualities in rare variants between mega-imputed and meta-imputed genotypes

We have seen that the imputation qualities per variant among the rare variants on chromosome 5 were smaller in the meta-imputation compared to the imputation qualities in the mega-imputation. Because it was unclear, why rare variants yielded higher imputation qualities in the mega-analysis compared to the meta-analysis, I was interested in the influence of the MAC on the imputation qualities.

This influence was evaluated in the imputation qualities of all 1,206,753 rare variants (MAF < 1%, **Figure 16A** and **Figure 16D)**, in the 377,524 variants with MAF < 1% and MAF ≥ 0.01% (or equivalent with MAC ≥ 68%, **Figure 16B** and **Figure 16E**) and in 829,229 variants with MAF <0.01% (**Figure 16C** and **Figure 16F).** In all rare variants is the median imputation quality in the mega-analysis 0.48 and the median imputation quality across variant (of the median imputation quality across the 25 studies) in the meta-analysis is 0.084. While the median imputation qualities are comparable between meta-imputed and mega-imputed variants with MAF between 0.01% and 0.1% (0.43 vs. 0.31), the difference of the median imputation qualities in the variants with MAF <0.01% is high (0.50 vs. 0.04), indicating, that the meta-imputation could not impute variants with MAF < 0.01%.

In summary, the mega-imputation yielded comparable imputation qualities compared to the meta-imputation of common and less common variants. Comparable imputation qualities were not derived from variants with MAF < 0.01%, where the mega-imputation showed higher imputation qualities compared to the meta-imputation.

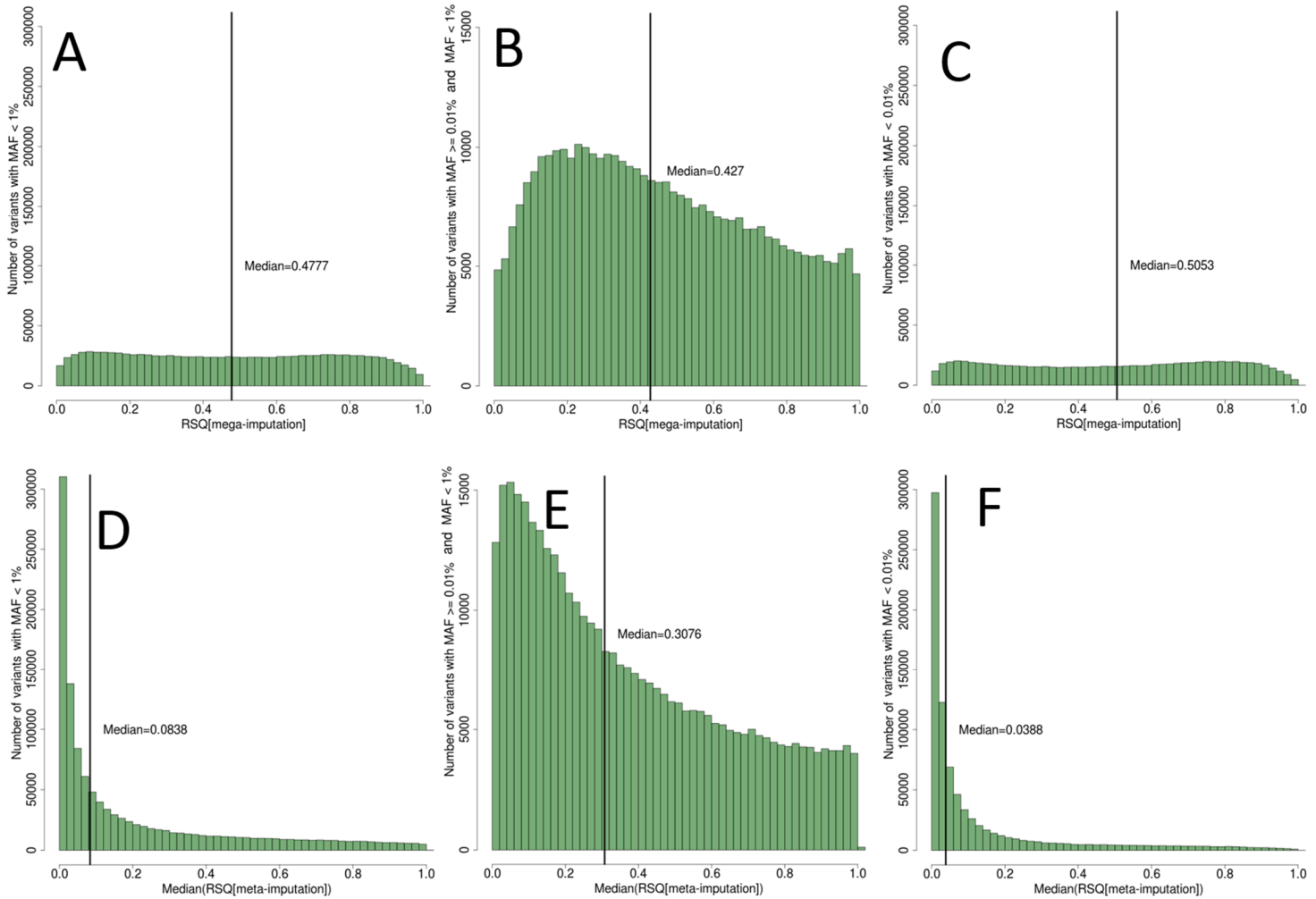

Figure 16. ***Median imputation quality of rare variants on chromosome 5 in mega- vs. meta-imputed genotypes****. Shown are the histograms of the imputation qualities among mega-imputed (Panels A-C) and meta-imputed rare variants (panels D-F) as reported by minimac. Panel A and D show all 1,206,753 rare variants (MAF<1%), panels B and E show the 377,524 variants with MAF ≥ 0.01% and MAF <1%; Panels C and F show the 829,229 variants with MAF < 0.01%. The median imputation quality for all variants in the mega-imputation and the median imputation quality across all variants of median imputation qualities across all 25 imputed studies from the meta-analysis are given as black horizontal lines. Different axis scaling was used in panels B and E, due to a smaller number of variants.*

3.2.4 Comparing the power of mega-imputed compared to meta-imputed genotypes

Next, I was interested in the influence of the imputation quality on the power to detect variants in the IAMDGC with genome-wide significance.

3.2.4.1 Comparing power in the 34 AMD disease loci

Given the similarities in imputation quality from **chapter 3.2.3**, I was interested in the change of power in the lead variants from the 34 AMD disease loci, comparing the meta-imputed and mega-imputed IAMDGC data. Power depends on the cases and controls analyzed, as well as the MAF, odds ratio and the imputation quality of the imputed variants. But only the imputation qualities differs between the mega-analysis and meta-analysis of the meta-imputed and mega-imputed IANDGC variants. The meta-analysis yield less power compared to the mega-analysis, due to the higher number of parameters in model. As this loss of power was not modelled in the evaluation of the power, I evaluate the differences in power in comparing the imputation qualities of the meta-imputed with the mega-imputed 34 lead variants. The imputation qualities of the 34 loci, known to be associated with AMD (see **Figure 17**) were generally comparable. Higher or equivalent imputation qualities were observed in 15 variants in the mega-imputation compared to the meta-imputation. The absolute difference is only small (min=0%, 25th percentile=0.02%, median=0.06%, 75th percentile=0.52%, max=3.77%, standard deviation=0.73%, compare **Appendix 7.8**). A complete summary of imputation qualities from the mega-analysis and all 25 studies in all 34 loci can be found in **Appendix 7.8**.

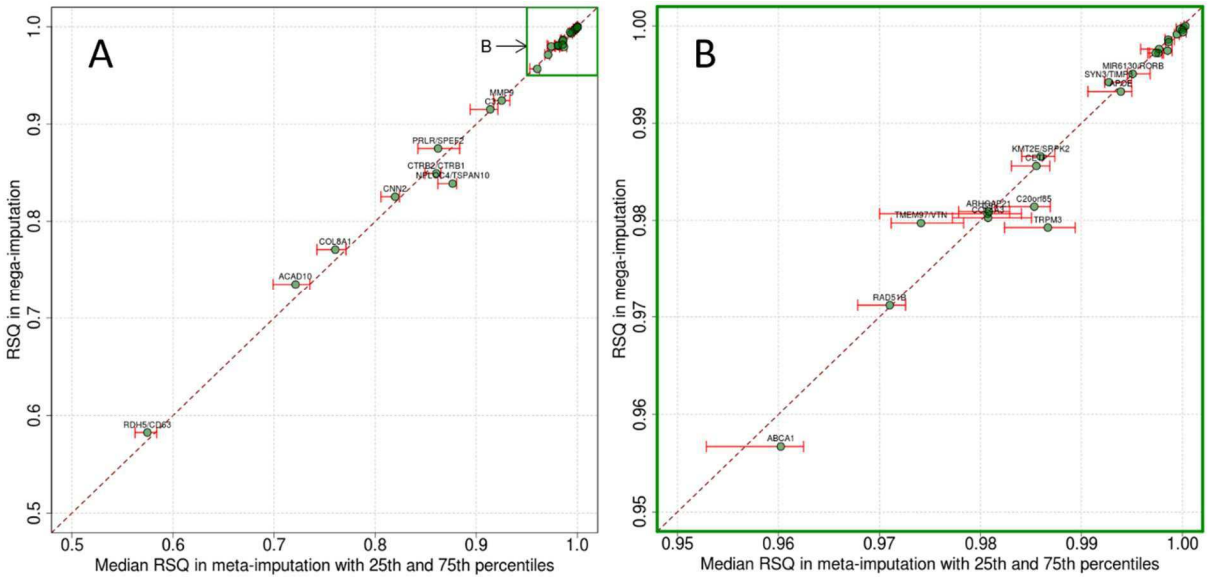

Figure 17. ***Imputation quality of the 34 loci known to be associated with AMD****. Shown are the imputation qualities (=RSQ) from the mega-analysis (Y) versus the median imputation qualities from the 25 studies including 25th and 75th percentiles as red whiskers (X). Panel A shows all variants, panel B shows only those variants with imputation quality > 0.95 in either imputation.*

In summary, both meta-imputed and mega-imputed IAMDGC data yielded high statistical power to detect the lead variants in the 34 AMD disease loci, as the differences between the imputation qualities between the meta-imputed and mega-imputed lead variants in the 34 AMD disease loci were small.

3.2.4.2 Evaluating power in rare variants in the IAMDGC data

Next I was interested how much power mega-analyses yield to detect rare variants. The change of power to detect a variant with genome-wide significance is visualized in **Figure 18** .Here the α-level of $5x10^{-8}$ (reflecting genome-wide significance), the number of subjects and case-control ratio from the IAMDGC data were assumed. The IAMDGC data yielded at least 80% power for variants with a MAF of 10% and an odds ratio of at least 1.17 (**Appendix 7.7.1**). Additionally, the IAMDGC data yielded at least 80% power for a perfectly imputed (imputation quality = 1) variant with a MAF ≥ 0.51% (**Appendix 7.7.27.7**), and that 27 of the 34 lead variants in the IAMDGC data yielded at least 80% power to be associated with genome-wide significance in a wide spectrum of MAFs and effect sizes (**Appendix 7.7.3**).

I evaluated power curves for a variant with $MAF_{controls}$=0.09% and a moderate estimated effect size (OR=1.8). These were taken from the rare lead variant in the C9 locus, identified by the IAMDGC SFS analysis. The data yielded sufficient power to detect this variant with an imputation quality = 0.8 for a MAF ≥ 0.63%. The median imputation qualities in rare variants in **Table 20** (on all 1,206,753 variants) were $MedianRSQ_{mega\text{-}imputed}$=0.478 and $MedianRSQ_{meta\text{-}imputed}$=0.084. For these imputation qualities the IAMDGC data yielded at least 80% to detect a variant with these imputation qualities with a MAF ≥ 1.07% and MAF ≥ 6.7%, respectively.

In summary, the IAMDGC data set is well power for the detection of rare variants with a MAF ≥ 0.63% for well imputed variants with an imputation quality ≥ 0.8.

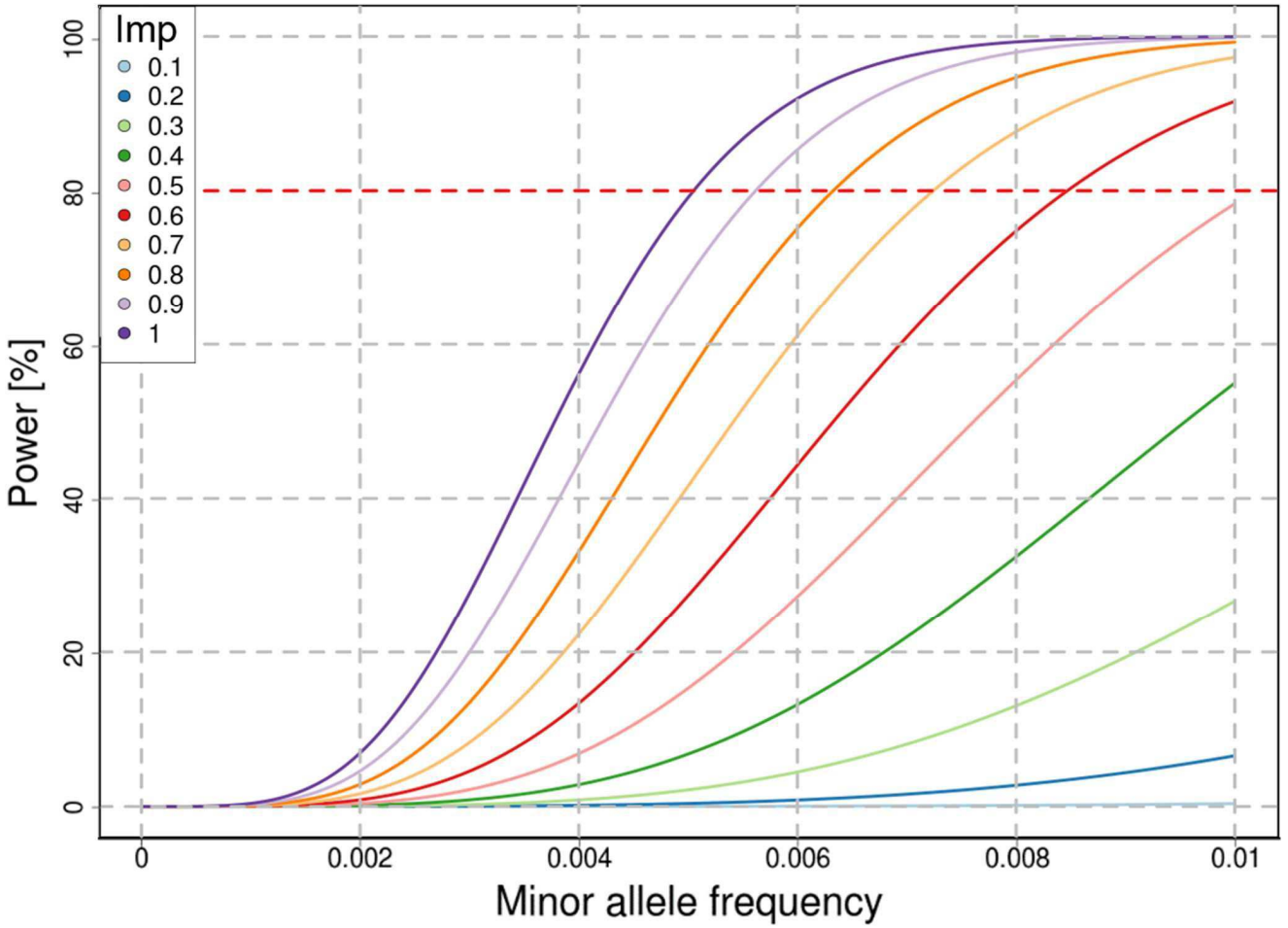

Figure 18. ***Power to detect rare variants dependent by imputation quality****. Shown is the MAF (x-axis) vs. power [%] (y-axis) for variant with an odds ratio of 1.8 in 16.144 cases and 17.832 controls. Shown is the power as function of imputation quality (Imp).*

3.2.5 Evaluating the gain of mega-imputing and mega-analyzing with meta-imputing and meta-analyzing IPD

Next, I was interested in comparing the meta-imputed and meta-analyzed IAMDGC data (the workflow commonly applied in consortia) with the mega-imputed and mega-analyzed IAMDGC data.

3.2.5.1 Comparing meta-analysis with mega-analysis in the 34 susceptibility loci identified by the IAMDGC

To evaluate the gain between these two approaches, I first compared the p-values, odds ratios and standard errors in the 34 lead variants in the susceptibility associated with AMD. Among these lead variants, ten of the 34 signal were not genome-wide significant in the meta-analysis of meta-imputed data (**Table 21**). The odds ratios, standard errors and p-values in these variants were illustrated in **Figure 19**. The odds ratios were comparable across the 34 variants, indicating, that there is no systematic inflation. The standard errors were higher in the meta-imputed and meta-analyzed data and the p-values were smaller in the IAMDGC GWA analysis compared to the meta-imputed and meta-analyzed data, reflecting the increased number of parameters in the meta-analysis model compared to the number of parameters in the mega-analysis model.

In 21 loci, lead variants in the meta-imputed and meta-analyzed analysis were identified, which showed smaller p-values compared to the lead variant in the IAMDGC GWA analysis. These different lead variants were proxies (R^2>0.8) to the lead variants from the IAMDGC SFS analysis. Considering also these, eight of the overall 34 loci identified by the IAMDGC SFS analysis were missed by the meta-imputation and meta-analysis. These eight loci were reported for the first time in the current analysis of the IAMDGC. The study specific effects and standard errors in the 34 susceptibility loci for AMD across all 25 studies is shown in **Appendix 7.9**, indicating homogeneity of the study effect across the 25 studies. No further evaluation of homogeneity across the studies was conducted in this work.

In summary, the meta-imputed and meta-analyzed data yielded different lead variants in the 34 AMD disease loci and missed eight loci, identified in the IAMDGC SFS analysis.

Table 21. ***Lead variants in the 34 AMD disease loci from the meta-imputed and meta-analyzed IAMDGC data****. Shown are association statistics of mega- and meta-analysis of the lead variants from the 34 loci associated with AMD identified by the mega-analysis. A lead variant is shown, if it was different from the reported lead variant in one of the 34 disease loci.*

SFS lead variant	Closest gene	OR	P-value	I^2%	Lead variant from GWA if different	OR	P-value	I^2%
rs10922109	*CFH*	0.40	0	58.7	rs1089033	0.48	**1.16×10^{-306}**	57.4
rs11884770	*COL4A3*	0.90	2.18×10^{-07}	13.3	rs112103000	0.87	6.74×10^{-08}	4.8
rs62247658	*ADAMTS9-AS2*	1.12	**2.46×10^{-11}**	47.1	rs6775974	1.12	**1.48×10^{-11}**	47.3
rs140647181	*COL8A1*	1.49	**4.90×10^{-08}**	0	rs56339461	1.14	**1.39×10^{-08}**	0
rs10033900	*CFI*	1.13	**3.02×10^{-13}**	0	---	---	---	---
rs114092250	*PRLR/SPEF2*	0.75	4.13×10^{-05}	0	rs35559912	0.84	2.32×10^{-06}	0
rs62358361	*C9*	1.64	**2.18×10^{-09}**	0	---	---	---	---
rs116503776	*C2/CFB/SKIV2L*	0.57	**4.51×10^{-96}**	46.2	---	---	---	---
rs943080	*VEGFA*	0.88	**2.29×10^{-13}**	0	---	---	---	---
rs7803454	*PILRB/PILRA*	1.14	**2.33×10^{-09}**	11	rs11761306	1.18	**6.50×10^{-10}**	22.6
rs1142	*KMT2E/SRPK2*	1.10	3.23×10^{-07}	19.3	rs6950894	1.09	2.45×10^{-07}	0
rs79037040	*TNFRSF10A*	0.90	**1.56×10^{-10}**	0	---	---	---	---
rs71507014	*TRPM3*	1.09	1.86×10^{-06}	1.9	---	---	---	---
rs10781182	*MIR6130/RORB*	1.09	1.26×10^{-06}	0	rs144700666	1.10	6.78×10^{-08}	0
rs1626340	*TGFBR1*	0.87	**7.95×10^{-11}**	0	rs44466	0.86	**6.22×10^{-11}**	0
rs2740488	*ABCA1*	0.88	**5.66×10^{-10}**	51.3	---	---	---	---
rs12357257	*ARHGAP21*	1.11	7.33×10^{-07}	10.6	---	---	---	---
rs3750846	*ARMS2/HTRA1*	2.71	0	63.7	rs2248799	1.80	**3.80×10^{-244}**	35.3
rs3138141	*RDH5/CD63*	1.16	**5.24×10^{-08}**	0	rs56108400	1.17	**8.07×10^{-09}**	0
rs61941274	*ACAD10*	1.44	4.54×10^{-07}	28.1	rs61941272	1.44	4.47×10^{-07}	27.7
rs9564692	*B3GALTL*	0.90	**1.62×10^{-08}**	23.1	---	---	---	---
rs61985136	*RAD51B*	0.91	1.20×10^{-07}	21.9	rs11624933	0.85	**9.40×10^{-10}**	6.1
rs2043085	*LIPC*	0.88	**7.54×10^{-12}**	0	rs2414577	0.88	**1.03×10^{-12}**	0
rs5817082	*CETP*	0.84	**4.00×10^{-18}**	0	---	---	---	---
rs72802342	*CTRB2/CTRB1*	0.83	8.12×10^{-08}	0	rs55993634	0.83	**1.40×10^{-08}**	0
rs11080055	*TMEM97/VTN*	0.91	**3.19×10^{-08}**	0	rs704	0.91	**1.02×10^{-08}**	0
rs6565597	*NPLOC4/ TSPAN10*	1.12	**2.03×10^{-09}**	8.8	rs62075723	1.12	**4.21×10^{-10}**	0
rs67538026	*CNN2*	0.91	3.66×10^{-06}	0	rs58369307	0.87	2.10×10^{-06}	0
rs2230199	*C3*	1.37	**3.18×10^{-50}**	0	---	---	---	---
rs429358	*APOE*	0.72	**8.27×10^{-32}**	22.5	---	---	---	---
rs142450006	*MMP9*	0.84	**1.29×10^{-10}**	0	rs1888235	0.85	**1.12×10^{-10}**	0
rs201459901	*C20orf85*	0.76	**2.88×10^{-14}**	0	rs117739907	0.76	**2.78×10^{-14}**	0
rs5754227	*SYN3/TIMP3*	0.77	**1.17×10^{-22}**	0	---	---	---	---
rs8135665	*SLC16A8*	1.14	**6.86×10^{-10}**	30.2	rs11089861	1.14	**2.76×10^{-10}**	0.9

The gene closest to the variant is given. OR = odds ratio, The P-value is zero in the CFH and ARMS2/ HTRA loci as the meta-analysis software did not provide a p-value. I^2% is the heterogeneity obtained by the meta-analysis software.

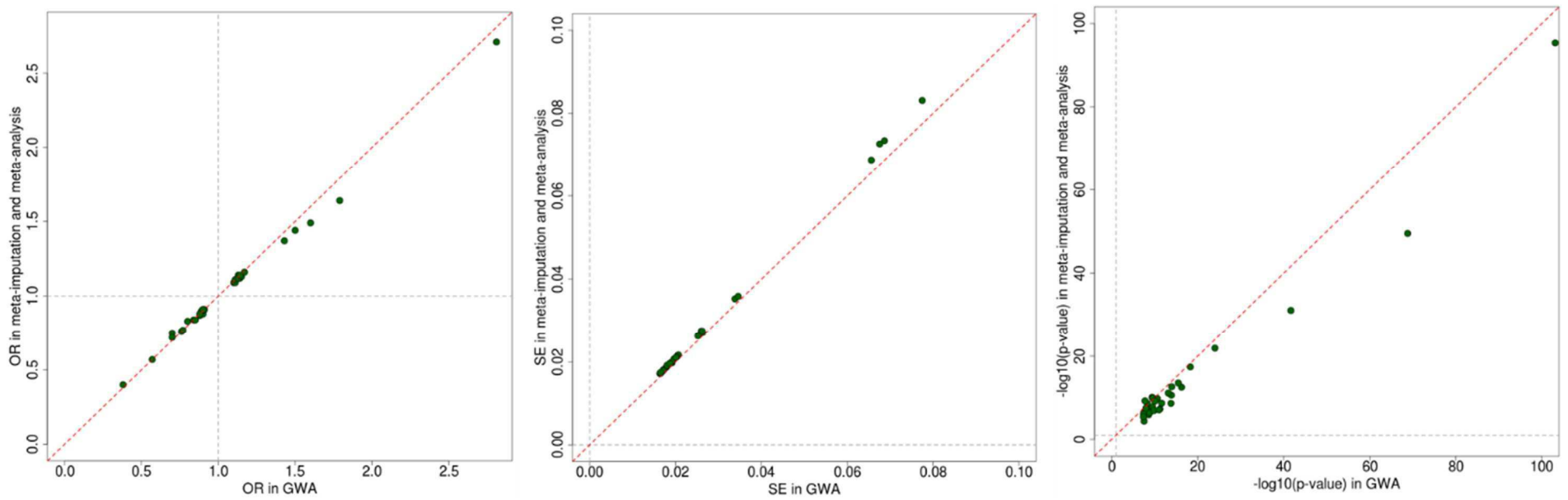

Figure 19. ***Comparison of odds ratios, standard errors and p-values between the meta-imputed and meta-analyzed IAMDGC data with the IAMDGC GWA analysis in the 34 AMD disease loci.*** *The IAMDGC SFS analysis identified 34 lead variants in the 34 AMD disease loci. Shown are the scatterplots of those lead variants comparing the results of the IAMDGC GWA analysis (X-axis) with the results from the meta-imputed meta-analyzed IAMD analysis (Y-axis). Panels A, B and C compare the odds ratios (OR), standard errors (SE) and –log10 p-values, respectively.*

3.2.5.2 Comparing meta-analysis with mega-analysis on chromosome 5

The results from the meta-imputed and meta-analyzed IAMDGC data were compared to the results from the mega-imputed and mega-analyzed IAMDGC data on chromosome 5.

The number of variants in this comparison is identical to those in **chapter 3.2.2.2,** as the MAC and imputation quality was determined in mega-imputed and mega-analyzed variants: Variants with a minor allele count (MAC) < 100 (equivalent with a MAF < 0.15%) and RSQ <0.4 were excluded.

The odds ratios, standard errors and p-values were visualized in **Figure 20A-C**. The odds ratios were comparable in both analyses: half of the variants show an increase and a decrease, respectively (median decrease of $8.8x10^{-4}$), indicating no biased effects. The scatter of odds ratios is highest in the rare variants. The standard errors are higher in the meta-analysis: 94.9% of the standard errors are higher in the meta-imputed and meta-analyzed data compared to the mega-imputed and mega-analyzed data (median increase of $2.2x10^{-2}$). Approximately 61.4% of all variants show increased of p-values and no variant can be identified with genome wide significance additional to the mega-imputed and mega-analyzed data.

The subset of variants with MAF ≥ 1% and imputation quality ≥ 0.4 was also identical to the number of variants in **chapter 3.2.2.2** in **Figure 20D-F)**. Here the odds ratio were also comparable between the both analyses. The standard errors and p-values were higher in the meta-imputed and meta-analyzed data compared to the mega-imputed and mega-analyzed data, indicating the loss of power in the meta-analysis compared to the mega-analysis. Again, the p-values of the *C9* lead variant was decreased, but was genome wide significant in both analyses. The lead variant in the *PRLR* locus was genome-wide in the mega-analysis but not in the meta-analysis.

In summary, the results of this comparison are in line with the results in **chapter 3.2.2.2:** No biased odds ratio were observed; no additional genome wide significant variants were identified; the increased standard errors and p-values reflect the loss of power between the models. The results exemplified with odds ratios, standard errors and p-values on chromosome 5 can be generalized to the other chromosomes as well (data not shown).

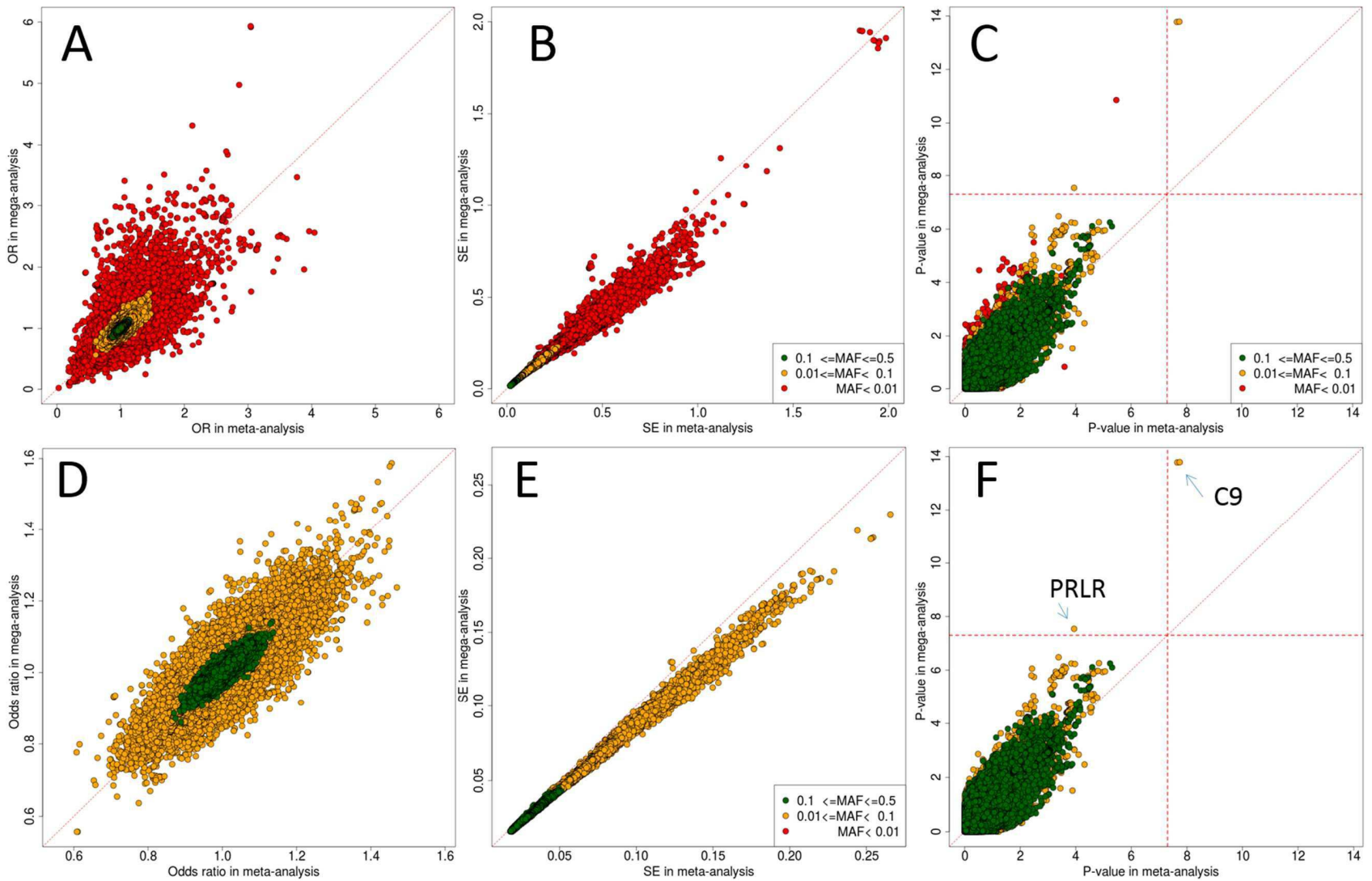

Figure 20. ***Comparison of odds ratios, standard errors and p-values between mega-imputed and mega-analyzed versus meta-imputed and meta-analyzed genotypes on chromosome 5****. Shown are the odds ratios, standard error s and p-values on a negative logarithmic scale of all variants in the meta-imputed and meta-analyzed genotypes (x-axis) compared to the mega-imputed and mega-analyzed genotypes (y-axis). Panels A-C show all variants on chromosome 5, whereas panels D, E, F show high quality variants with MAF>1% and RSQ >0.4.*

3.2.6 Evaluating the change of the top variants of the mega-imputed and mega-analyzed data in the 34 AMD disease loci

In seven loci, the lead variants in the mega-imputation and mega-analysis (GWA analysis) of the IAMGC data was different from the lead variant identified by the IAMDGC SFS (in the loci *ACAD10*, *LIPC*, *CTRB2/CTRB1*, *TMEM97/VTN*, *MMP9*, *C20orf85* and *SLC16A8,* **Table 22***)*. The odds ratios, standard errors and p-values in those seven variants were compared in the mega-imputation/analysis, mega-imputation/ meta-analysis and meta-imputation/ analysis. Six variants were genome wide significant in all three analyses, whereas the p-values were lowest in the mega-imputed and mega-analyzed data. The lead variant in the *ACAD10* locus was only genome wide significant in the mega-imputed and mega-analyzed data. The odds ratios were at comparable levels in all three analyses. The standard errors were higher in the mega-imputation/ meta-analysis and meta-imputation/ analysis, due to the increased number of parameters in the model. These results are consistent with previous results.

Table 22. ***Evaluation of the 6 lead variants from the mega-imputed and mega-analyzed data, which are different from the lead variants in the IAMGDC SFS.***

				Mega-imputed and mega-analyzed			Mega-imputed and meta-analyzed			Meta-imputed and meta-analyzed		
Variant ID	Closest Gene	Chr	Position (bp)	OR	Standard error	P-value	OR	Standard error	P-value	OR	Standard error	P-value
rs61941287	*ACAD10*	12	112132610	1.53	0.0711	1.60×10^{-09}	1.46	0.0748	3.80×10^{-07}	1.46	0.0771	7.70×10^{-07}
rs2414577	*LIPC*	15	58680954	0.87	0.0172	4.77×10^{-15}	0.88	0.0180	9.54×10^{-13}	0.88	0.0180	1.03×10^{-12}
rs55993634	*CTRB2/CTRB1*	16	75234872	0.81	0.0311	7.80×10^{-12}	0.83	0.0324	1.79×10^{-08}	0.83	0.0325	1.40×10^{-08}
rs4795433	*TMEM97/VTN*	17	26649724	0.91	0.0163	1.11×10^{-08}	0.91	0.0171	4.00×10^{-08}	0.91	0.0171	4.27×10^{-08}
rs1888235	*MMP9*	20	44614991	0.86	0.0247	2.31×10^{-10}	0.85	0.0257	1.20×10^{-10}	0.85	0.0258	1.12×10^{-10}
rs117739907	*C20orf85*	20	56653724	0.76	0.0345	2.96×10^{-16}	0.76	0.0358	2.42×10^{-14}	0.76	0.0358	2.78×10^{-14}
rs11089861	*SLC16A8*	22	38476276	1.14	0.0202	3.51×10^{-11}	1.15	0.0212	3.10×10^{-11}	1.14	0.0213	2.76×10^{-10}

The closest gene to the variant is reported. Chr= chromosome. Position (bp) is give in base positions on GRCh37. OR = odds ratio.

3.2.7 Evaluating the type I error

Finally, it was the question, if the type I error was increased in the meta-imputed and meta-analyzed IAMDGC data. To evaluate this, the genomic inflation was derived from the 14,939 genotyped variants with MAF > 5% on chromosome 5 to avoid potentially biased results introduced by imputed variants and less common or rare variants. The genomic inflation λ = 1.05 indicates no systematic inflation of the p-values. Also the QQ-plot (**Figure 21)** of these common variants indicates no systematic inflation. This evaluation was conducted on chromosome 5 and can be generalized to the whole genome.

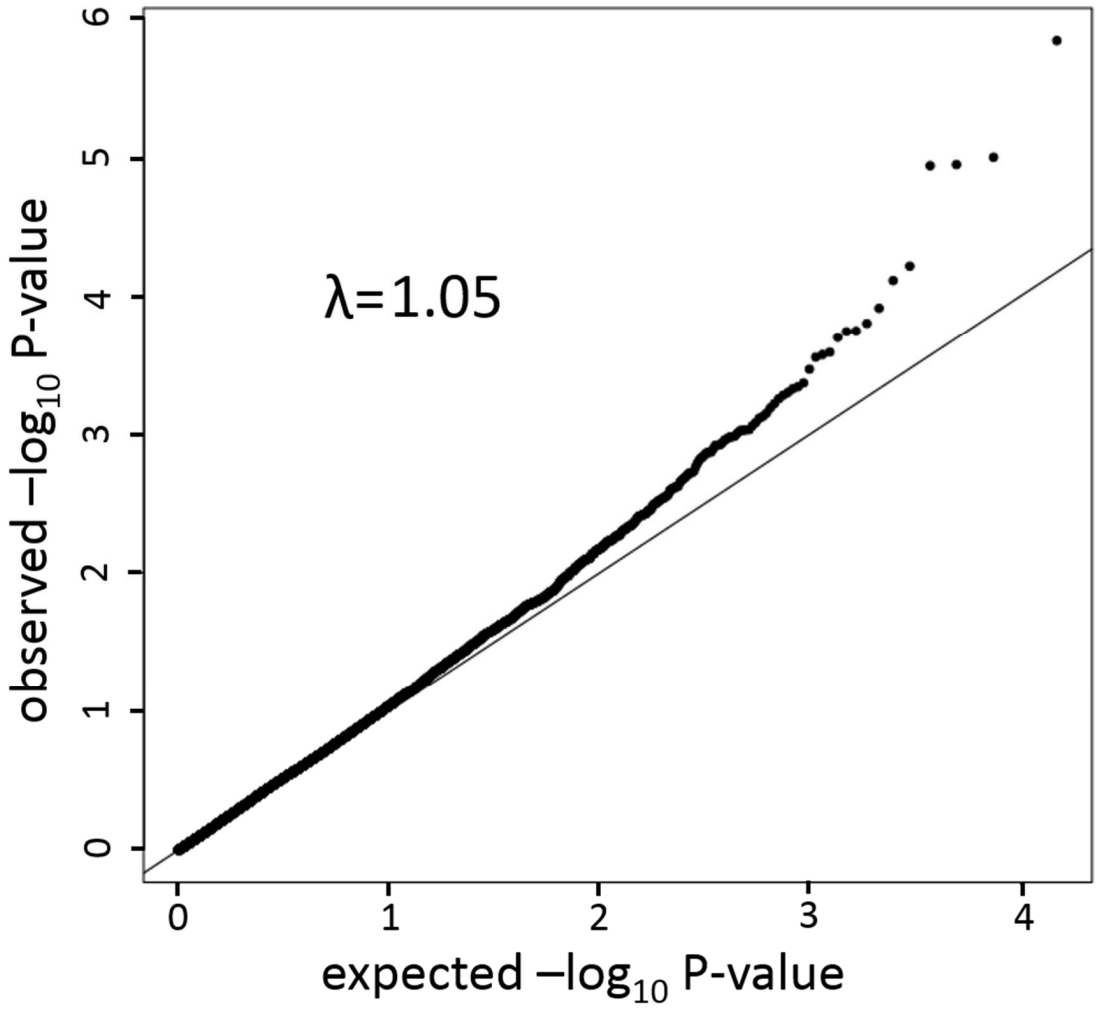

Figure 21. ***QQ-Plot of p-values from the meta-imputation and meta-analysis of the IAMDGC data on chromosome 5.*** *Shown is the expected –log₁₀ P-value vs. the observed –log₁₀ p-values of the 14,939 genotyped common variants (MAF>5%) on chromosome 5.*

3.2.8 Summary

Mega-imputation and mega-analysis proves to be a powerful approach to identify genetic susceptibility loci for AMD. I showed that mega-imputing and mega-analyzing the IAMDGC data identified the same AMD disease loci compared to the IAMDGC SFS analysis. Meta-analyzing compared to mega-analyzing the same mega-imputed IAMDGC data missed six (of the 34) AMD disease loci. Mega-imputing a large number of subjects across several studies results yielded comparable imputation qualities of less common and common variants, but higher imputations qualities in variants with MAF <0.01% compared to meta-imputing subjects by study. I evaluated that the genome wide data of the IAMDGC yields high power to detect disease loci with low to moderate effect estimates for AMD with genome wide significance and that the data yielded sufficient power to detect even rare variants with moderate effect estimates with genome-wide significance. The commonly applied approach of meta-imputing and meta-analyzing missed eight (of the 34) AMD disease loci. Overall, the mega-analysis of the IAMDGC data is a power full approach to identify disease loci for complex disease due to the increased power in the statistical model, which includes less parameters compared to the meta-analysis approach.

3.3 Optimizing computing resources for mega-imputation

Imputing untyped variants in study data in several thousands of subjects is a computational demanding task. But with the introduction of the 1000 Genomes reference panels this computational burden has even risen. Then the 2 step approach of imputation was introduced to allow the study analysts to separate the haplotype estimation from the genotype imputation, facilitating genotype imputation. Now, the pooled IPD of more than 50,000 subjects yield yet another challenge, increasing the computational burden again. Genotype imputation of all subjects from 25 studies with the high density 1000 Genomes reference panel is a challenge, which makes the efficient use of computation resources necessary. It is a challenge to impute such a large data set in reasonable time, even on modern multi-core server systems. This chapter aims at identifying the computational burden for mega-imputing of large scale genome wide data sets and at solving these challenges in parallelizing computations on a multi core server cluster. I exemplify my analysis with data from the IAMDGC. First I identify the technical aspects for mega-imputation (**chapter 3.3.1**) and second I suggest how to solve them (**chapter 3.3.2**).

3.3.1 Technical aspects for parallelizing mega-imputation

It is a mandatory requirement for the software used for phase estimation and genotype imputation, that it can split the genome wide data into smaller parts and to process the parts in parallel. For the mega-imputation in my thesis I focus on the software *ShapeIT* for phase estimation and on *minimac* for genotype imputation, as both can work in parallel on multiple cores.

To optimize the computing resources for the mega-imputation, I **first** evaluated the optimal strategies for separating genome wide data into smaller parts where possible. **Second** I optimized imputation accuracy by evaluating optimal parameters for both phasing and imputation. **Third** I analyzed the time needed for phase estimation and imputation separately. To minimize the overall time to imputation I proposed a pipelining concept that minimizes the overall time to imputation. **Fourth** I investigated the needed volatile memory for both steps and **fifth** sheded light on needed data storage to buffer imputed genotypes on hard disk.

3.3.2 Parallelizing mega-imputation

3.3.2.1 Splitting genome wide data into smaller parts

The aim of this chapter is to evaluate how the complexity of phase estimation and genotype imputation can be reduced by splitting the genetic data into smaller parts. For phase estimation it is recommended by the authors to estimate haplotypes in whole chromosomes [60]. This approach yields highest accuracy but phase estimation is very time consuming (see also **chapter 3.3.2.3**). Genotype imputation can theoretically also be done on whole chromosomes, but as imputed data of a whole chromosome may extend volatile memory even on modern server clusters (see also **chapter 3.3.2.4**), splitting chromosomes into smaller parts is highly recommended. Furthermore a significant time gain can be achieved in splitting chromosomes into smaller parts (also called *chunks*). The loss in accuracy will be small, except for populations where haplotype sharing extends to several Megabases (see imputation cookbook of minimac in the **Web Resources**). This is typical for studies from exclusively more isolated populations (for example from specific regions in Finland or Sardinia), which is not the case in a typical consortia setting. Two possible strategies can be implemented: The so called *Target based chunking* divides the chromosome into parts, where the number of variants from study data is equal in all parts. On the other hand the so called *Reference based chunking* divides the chromosome into parts, spanning an equal number of base pairs (compare **Figure 22** and the imputation cookbook in the **Web Resources**). I apply the reference based chunking for three reasons: **First**, both approaches yield comparable accuracy. **Second** reference based chunking is a less complex task, because (in contrast to the target based chunking) it does not utilize external files, defining start and stop variants of the parts. This ultimately means, that reference based chunking can be integrated into scripts quite easy. **Third**

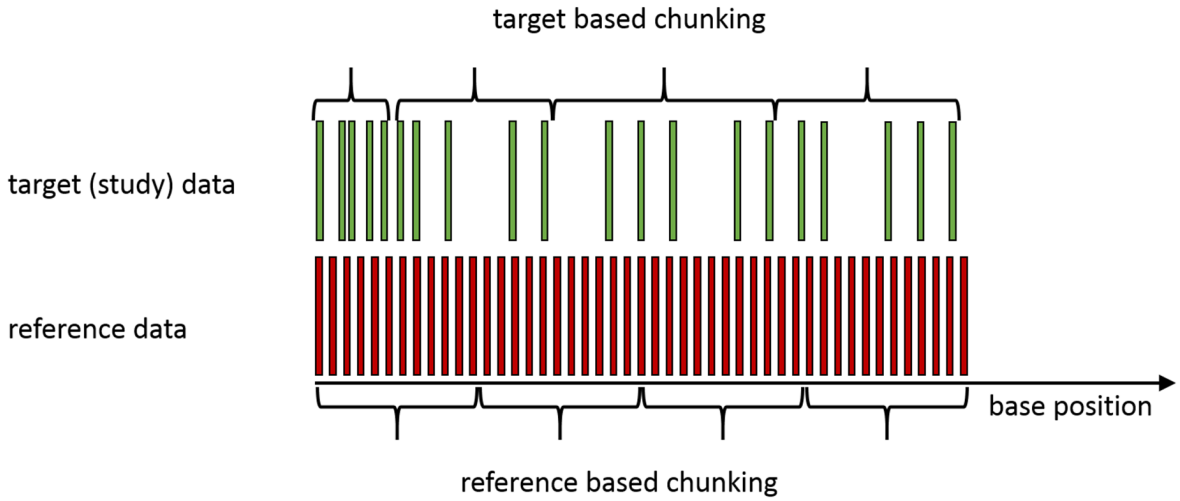

*Figure 22. **Approaches of splitting chromosomes into parts for genotype imputation**. Shown are the target based chunking, where (in this example) five variants (green bars) are used per part. The length of each part (in base pairs) may vary. In contrast the reference based chunking utilizes all variants in a region of equal base pair length independent from the coverage of the number of variants in the region. The length of all parts (in base pairs) is equal.*

the imputed genotypes are subdivided into equally sized files (in terms of base positions), which make look-ups and extraction of specific region (for example gene sets) after imputation much easier.

3.3.2.2 Deriving optimal program parameters

Next I identified optimal parameters for phase estimation and genotype imputation, which are then utilized for calculating time constraint and deriving an optimized schedule for imputation. Different parameters can be used to phase and impute untyped variants (compare **chapter 1.3.1**). The Hidden Markov Model underlying the algorithm for phase estimation utilized a number of subjects, which is specified by a *states* parameter. The set of variants used at a time is specified as *window* size. For genotype estimation a so called *chunk* defines the core region (in chromosomal positions), where the variations are inferred. To avoid problems with imputing variants at the ends of the chunk, an *overlap* to each side is specified (compare **chapter 1.3.2**).

*Table 23. **Evaluating optimal parameters for phase estimation and imputation**. Shown are the median imputation qualities resulting from phasing and imputation with permutations of states and window size for phase estimation as well as chunk size and overlap for genotype imputation on the example of chromosome 21 in the IAMD data set.*

Parameter				Median imputation quality of variants by category of MAF								
Phasing		Imputing										
states	window [MB]	chunk [MB]	overlap [Kb]	Total	<.1%	<1%	<5%	<10%	<20%	<30%	<40%	<50%
200	1	2.5	500	0.8842	0.57	0.80	0.91	0.93	0.94	0.92	0.94	0.92
200	1	2.5	250	0.8840	0.57	0.81	0.91	0.93	0.94	0.92	0.94	0.92
200	1	5	250	0.8837	0.57	0.80	0.91	0.93	0.94	0.92	0.94	0.92
200	1	2.5	100	0.8837	0.57	0.81	0.91	0.93	0.94	0.92	0.94	0.92
200	1	5	100	0.8835	0.57	0.81	0.91	0.93	0.94	0.92	0.94	0.92
100	1	2.5	500	0.8817	0.56	0.80	0.91	0.93	0.94	0.92	0.94	0.92
100	1	2.5	250	0.8815	0.55	0.80	0.91	0.93	0.94	0.92	0.94	0.92
100	1	5	500	0.8815	0.55	0.80	0.91	0.93	0.94	0.92	0.94	0.92
100	1	5	250	0.8814	0.55	0.80	0.91	0.93	0.94	0.92	0.94	0.92
100	1	2.5	100	0.8810	0.55	0.80	0.91	0.93	0.94	0.92	0.94	0.92
100	1	5	100	0.8809	0.55	0.80	0.91	0.93	0.94	0.92	0.94	0.92
200	0.5	2.5	500	0.8804	0.57	0.80	0.91	0.93	0.94	0.92	0.94	0.91
200	0.5	5	500	0.8802	0.57	0.80	0.91	0.93	0.94	0.92	0.94	0.91
200	0.5	2.5	250	0.8801	0.56	0.80	0.91	0.93	0.94	0.92	0.94	0.91
200	0.5	5	100	0.8799	0.56	0.80	0.91	0.93	0.94	0.92	0.94	0.91
200	0.5	5	250	0.8799	0.56	0.80	0.91	0.93	0.94	0.92	0.94	0.91
200	0.5	2.5	100	0.8798	0.56	0.80	0.91	0.93	0.94	0.92	0.94	0.91
100	0.5	2.5	250	0.8774	0.55	0.80	0.90	0.93	0.94	0.92	0.94	0.91
100	0.5	5	250	0.8773	0.55	0.79	0.90	0.93	0.94	0.92	0.94	0.91
100	0.5	2.5	500	0.8770	0.55	0.79	0.90	0.93	0.94	0.92	0.94	0.91
100	0.5	5	500	0.8770	0.55	0.79	0.90	0.93	0.94	0.92	0.94	0.91
100	0.5	2.5	100	0.8768	0.55	0.79	0.90	0.93	0.94	0.92	0.94	0.91
100	0.5	5	100	0.8767	0.55	0.79	0.90	0.93	0.94	0.92	0.94	0.91

For phase estimation: The parameters states are given as absolute numbers 100 and 200 and window size is categorized into 0.5 and 1 MB. For genotype imputation: Chunk size can be 2.5 and 5 Megabases and the overlap is categorized by 100, 250 and 500 Kb. The median imputation quality is exemplified in categories of all variants and by variants subdivided into several MAF categories (<.1%, <1%, <5%, <10%,<20%,<30%,<40% and <50%).

The evaluation of *states*, *window size*, *chunk size* and *overlap* shows only minor differences in the median imputation quality. There are nearly identical imputation qualities for all categories of common variants (MAF>5%: 0.91, MAF<10%: 0.93, MAF<20%: 0.94, MAF<30%: 0.92, MAF<40%: 0.94, MAF<50%: 0.92). Highest imputation qualities across all variants can be obtained for 200 states and 1 Megabase window size for phase estimation and 2.5 Megabases chunk size and 500 Kilobases overlap in genotype imputation (median imputation quality = 0.8844, see **Table 23**).

3.3.2.3 Finishing mega-analysis in reasonable time

To evaluate the time constraints, I investigated phase estimation and genotype imputation separately. To achieve most accurate estimated haplotypes in phase estimation, it is recommended to phase whole chromosomes. The overall 508,740 variants were split by chromosome. Full chromosomes were phased at once. **Table 24** shows how long each autosome needs to be phased with 8 parallel threads on a server (min ~= 2 days for chromosome 21, max ~= 16 days for chromosome 1). Thus, phase estimation of all 22 autosomes needs 163 hours * 8 cores ~= 3.5 core-years.

Untyped variants are estimated in parts (chunks). The time to imputation of one chunk depends on the number of variants in the chunk and varies between 4 and 16 hours (median = 10 hours). Overall genotype imputation is done in 1,064 [chunks] * 4 cores * 10 hours ~= 4.9 core-years. Thus, the overall computation burden for phase estimation and genotype imputation is approximately 8.4 core-years.

Modern server architectures and the programs used for phase estimation and genotype estimation allow the parallelization of these computations. It was my aim to minimize the overall time to phasing and imputation with pipelining the single computations: first all 22 autosomes are phased in parallel. Assuming phasing on 8 cores per chromosomes, this would require a server cluster of 22 * 8 = 176 cores. Then phasing of the chromosome with the least variants (chromosome 21 in **Table 24**) is completed after 2 days and the phasing of the chromosome with the most variants (chromosome 1 in **Table 24**) is completed after 16 days. Since the phasing in the smaller chromosomes is completed, genotype imputation can be done in these chromosomes before the phasing of the larger chromosomes is completed. Thus the limiting factor is the time needed for imputing all chunks of the largest chromosome. Assuming 176 cores, 44 chunks can be imputed in parallel and the whole chromosome 1 can be finished in 3 rounds, considering a small time buffer for big chunks. Thus, both phase estimation and genotype imputation of the IAMDGC data set can be finished in a reasonable time with parallel computation on a modern server cluster with 176 cores in less than 3 weeks (16 + 2 = 18 days), mastering an overall computational burden of 8.4 core-years.

Table 24. ***Main statistics for data separation and computation resources for phase estimation and imputation of the IMDGC data****. Shown are the resources needed for phasing on an 8 core server cluster.*

chr	#variants	time [days]	memory [GB]	start [bp]	stop [bp]	#chunks	#chunks (merged/ no gaps)
1	47,359	15.83	40	10,583	249,239,465	100	89
2	39,735	13.87	39	10,140	243,185,846	98	94
3	34,472	11.43	33	60,157	197,946,621	80	78
4	27,697	10.22	29	10,240	191,043,593	77	75
5	27,934	10.50	29	11,956	180,876,273	73	71
6	34,102	9.95	33	73,924	171,051,269	69	67
7	25,950	8.92	27	16,161	159,128,574	64	62
8	22,958	7.86	25	11,880	146,303,866	59	57
9	21,894	7.62	23	10,023	141,132,999	57	45
10	23,521	7.68	24	60,523	135,523,864	55	52
11	28,468	8.28	25	70,855	134,946,451	54	52
12	25,604	7.80	23	61,107	133,841,510	54	53
13	14,564	5.05	17	19,020,013	115,109,852	39	38
14	16,840	5.17	16	19,002,084	107,289,453	36	35
15	16,804	4.84	15	20,001,200	102,520,965	34	32
16	18,629	5.35	19	60,054	90,292,811	37	30
17	20,258	5.44	17	56	81,194,907	33	31
18	11,752	4.10	14	10,644	78,017,128	32	29
19	21,277	4.91	16	80,840	59,118,838	24	22
20	12,691	3.82	13	60,479	62,965,028	26	24
21	6,513	2.01	8	9,411,243	48,119,751	16	14
22	9,718	2.36	10	16,050,408	51,243,297	15	14
SUM	508,740	163.00				1132	1,064

Chr = chromosome, #variants = number of variants, time [days] =time needed for phasing in days, memory [GB] = memory used for imputation in Gigabytes (GB), Start and stop position are given as base position on GRCh37, # chunks overall gives the number of chunks when imputing the chromosome in chunks of 2.5 Megabases size, does not account for gaps (f.e. centromer) in the reference data, # chunks (merged/ no gaps) gives the number of chunks imputed, when considering gaps and merging regions with small number of variants.

3.3.2.4 Deriving volatile memory constraints

Next I evaluated how much memory is needed for phasing and imputation. Phase estimation is implement with Hidden Markov Models, for which states and transition probabilities must be stored for all variants on the whole chromosome and for all subjects in the study sample. This means, that phasing programs require a high amount of volatile memory, which depends on the study sample size. In the IAMDGC data the memory for phase estimation varies between 8 and 40 Gigabytes (GB) per chromosome (for chromosome 21 with 6,513 variants and chromosome 1 with 47,359 variants, respectively).

Genotype imputation is not conducted on the whole chromosome, but in chunks. The memory used for genotype imputation depends on the size of the chromosome and on the number of variants in the respective chunk. Thus, the largest chunks from chromosome 1 utilize the most volatile memory. They need about 6 GB volatile memory, whereas for chunks in smaller chromosomes (for example chromosome 14 to 22) not more than 3 GB are needed. Nevertheless, volatile memory in genotype imputation is (compared to the memory consumption of the phase estimation) not the limiting factor, when the imputation is performed separately in chunks.

3.3.2.5 Evaluating storage needed for mega-imputation

Finally I investigated how much storage is need for phasing and imputation for a mega-imputation. To save storage, all data sets were compressed to zip-files. The input data set for phase estimation, consisting of 508,740 genotypes in the 52,189 subjects needs 3.7 GB disk space. The phased data consists of the same genotypes per subjects and thus needs comparable disk space of 4.2 GB. The recombination map used for phase estimation has a size of about 100 Megabytes (MB). The imputed data set requires 814 GB storage (**Table 25**). As a consequence **each subject** needs about **16 MB** data storage.

*Table 25**. Storage needed for the mega-imputation of the IAMDGC data set by chromosome**. Shown is the size in Megabytes of the unphased, phased and imputed data sets. For the latter the number of chunks is given. The needed storage on hard disk drive is quantified by the minimum, maximum, median and the sum of all files.*

				imputed			
	unphased	phashed		min	max	median	overall
chr	[MB]	[MB]	#chunks	[MB]	[MB]	[MB]	[MB]
1	304	352	89	401	1,138	676	62,223
2	307	350	94	353	1,417	680	66,443
3	264	301	78	345	1,295	694	55,763
4	237	271	75	323	1,679	743	58,049
5	231	263	71	370	1,348	717	51,356
6	266	302	67	353	1,293	681	48,898
7	207	236	62	457	1,415	760	48,907
8	199	227	57	376	2,002	686	43,974
9	169	193	45	460	1,345	749	34,989
10	191	218	52	355	1,458	732	39,835
11	191	217	52	445	1,549	749	39,695
12	182	206	53	404	1,351	686	38,747
13	139	159	38	460	1,507	702	28,451
14	124	141	35	509	1,089	705	25,972
15	116	132	32	322	1,269	772	25,197
16	122	139	30	482	1,691	875	28,644
17	109	124	31	498	1,283	793	25,760
18	109	125	29	515	1,497	779	24,441
19	94	105	22	517	1,916	991	22,720
20	93	106	24	300	1,476	806	19,870
21	54	62	14	212	1,158	891	12,223
22	61	69	14	457	1,540	862	12,772
	3,771	4,298					814,929

Chr = chromosome, MB = Megabyte, #chunks = number of chunks of the chromosome, min = minimum, max = maximum, overall = storage needed for all chunks in that chromosome.

3.3.3 Summary

Mega-imputation of big IPD sets are a challenge in genetic epidemiology. In this chapter I exemplified how to conduct a mega-imputation with the IMAGC data set that consist of 52,189 subjects genotyped at 508,740 variants, each. In using state-of-the-art techniques, such as parallel haplotype reconstruction on whole chromosomes with *ShapeIT* and parallel genotype imputation in parts of 2.5 Megabases with *minimac*, I exemplified how this computational demanding task can be finished within less than 3 weeks. Although technical requirements are high – a multi-core server cluster consisting of about 176 cores with up to 40 GB volatile memory and 814 GB storage is needed – the process is feasible.

4 Discussion

4.1 Summary of main results

In the following, I summarize the main results for each of the initial objectives. First I show how much we gain in using high density versus low density reference panels for imputation in a meta-analysis (**objective 1**; summary in **chapter 4.1.1**). Second I quantified the gain in mega-imputing and mega-analyzing compared to meta-imputing and meta-analyzing (**objective 2**; summary in **chapter 4.1.2**). Finally I summarize how computational methods and software support study analysts to re-imputed study with different and how a large scale mega-imputation can be conducted in reasonable time (**objective 3**; summary in **chapter 4.1.3**).

4.1.1 Identifying genetic loci associated with complex disease by meta-analysis of variants imputed with high and low density reference data

In my work, I evaluated the gain of using high density reference panels for genotype imputation compared to using low density reference panels for genotype imputation in a meta-analysis setting. For this evaluation I utilized data from the CKDGen consortium, which aimed at identifying risk loci for kidney function with the trait eGFRcrea. The low density HapMap reference panel and the high density 1000 Genomes reference panel were used for genotype imputation in CKDGen.

There are mainly two differences between these meta-analyses: First, the reference panels infer a different number of well imputed variants suitable for meta-analysis (HapMap: ~2.4 Mio, 1000 Genomes: ~11 Mio variants). Second, the number of analyzed subjects differ (HapMap: 133,318, 1000 Genomes: 110,517 subjects). As expected, more genotypes were imputed with the 1000 Genomes reference panels compared to the number of genotypes imputed with the HapMap reference panels. This increase was due to the increased number of variants comparing the 1000 Genomes with the HapMap reference panels. The genotypes imputed with 1000 Genomes reference panels showed higher imputation qualities. The overall number of well imputed variants imputed with 1000 Genomes reference panels was higher compared to the overall number of variants imputed with HapMap reference panels. Also in the overlap of all variants present in both 1000 Genomes and HapMap imputed variants, the 1000 Genomes imputed variants showed higher imputation qualities.

Surprisingly, the signal detection of genetic loci for kidney function revealed ten genetic loci, which have not been discovered by the HapMap meta-analysis, although the overall number of subjects was smaller in the 1000 Genomes meta-analysis compared to the HapMap meta-analysis. Among those ten loci, six loci are in the 1000 Genomes reference panel but not in the HapMap reference panel, pinpointing a main advantage of the 1000 Genomes reference panels compared to the HapMap reference panels: The 1000 Genomes reference panels include genetic regions, which were not included in the HapMap reference panels and thus allows the detect of novel loci associated

with complex disease in yet uncharted genetic regions. Thus, it is highly recommended to use high density reference panels instead of using low density reference panels for genotype imputation in a meta-analysis. I showed, that the identification of the other four (of the ten) loci detected by the CKDGen 1000 Genomes meta-analysis, not identified by the CKDGen HapMap meta-analysis was due to a better study composition in the studies exclusively meta-analyzed in the 1000 Genomes meta-analysis, compared to the studies exclusively meta-analyzed in the HapMap meta-analysis. However, the genotype imputation with 1000 Genomes reference panels is a computational demanding task. I provide computational methods and software to help study analysts to overcome this burden (see **chapter 3.3**).

4.1.2 On the gain from mega-analysis compared to meta-analysis

To evaluate the gain in mega-imputing and mega-analyzing IPD compared to meta-imputing and meta-analyzing genome wide association studies, I used the data from the IAMDGC consortium that consists of 52,189 quality controlled subjects, genotyped at 508,740 variants. The data is imputed jointly (mega-imputation) and separately by study (meta-imputation).

The IAMDGC detected 34 AMD disease loci with mega-imputing variants and a sequential forward selection approach on the imputed variants. The 34 AMD disease loci were identified in my work with the mega-imputation and mega-analysis.

To identify the influence of imputing the IAMDGC data jointly or by study, I compared the mega-imputation with the meta-imputation. The mega-imputation showed generally higher imputation qualities. There was a higher number of well imputed variants in the mega-imputation compared to the meta-imputation. The mega-imputation also yielded a gain in imputation quality in rare variants: There was a higher number of well and medium well imputed variants in the meta-analysis compared to the meta-analysis. But the mega-imputation yielded no gain for common and less frequent variants as the common and less frequent variants were equally well imputed.

To identify the influence of imputing and analyzing the IAMDGC data jointly or by study on signal detection, I conducted two comparisons: First the mega-imputed and mega-analyzed data of the IAMGGC data was compared to the meta-imputed and meta-analyzed data from the IAMGC. Second, I compared the mega-imputed and mega-analyzed data from the IAMDGC with the mega-imputed and meta-analyzed data from the IAMDGC. Overall, there was a gain in mega-imputing and mega-analyzing the data from the IAMDGC: I found that the meta-imputation/analysis missed eight loci, identified by the mega-imputation/analysis. These eight loci were identified as novel signals in the IAMDGC SFS analysis. Mega-imputing and meta-analyzing the data missed six loci. These six were among the eight loci missed by the meta-imputation/analysis.

This was about signal detection and was not a prove that the mega-analysis is always better than the meta-analysis, as mega-imputation and mega-analysis need to be evaluated in more detail:

Particularly in the meta-analysis it is clear, that it is a gold standard analysis in the sense, that it does not apply any model assumptions. But type I error and a bias of the estimated effects can occur. If model assumptions, as study homogeneity are valid, then the meta-analysis would yield less power compared to the mega-analysis. But I found that the empirical type I error was not increased. I also identified no biased effect estimates. The mega-imputation yielded higher imputation qualities compared to the meta-imputation. As mega-imputation is a computational demanding task, study analysts were supported to overcome this high computation burden (see **chapter 3.3**).

4.1.3 Software and approaches to accelerate genotype imputation

As a consequence to the high computational burden of mega-imputing large scale IPD, I provided methods and software to support study analysts to re-impute study data with high density reference panels for the detection of disease loci associated with genome-wide significance with complex disease.

When analysts prepare study genotypes of a single study, they exclude subjects and variants during quality control. The study data is also annotated to the genome build of the reference panel, which will be used subsequently for the time intensive phasing and imputation (**pre-phasing approach**). When a novel reference panels is released, it can be annotated on a more recent genome build compared to the study data. Then it becomes necessary that the study analyst re-annotates, re-phases and re-imputes the study data with this current build.

I evaluated which steps are needed to prepare genotype imputation with a reference panel. This was exemplified with HapMap and 1000 Genomes reference panels. I showed how a study analyst integrates changes between the build in the study data and the reference panel and which steps are needed to re-annotate, re-phase and re-impute the data, each time a novel reference panel is released. To facilitate this process and to allow the study analyst to quickly re-impute the study genotypes, I introduced the **post-phasing approach** of lift over for imputation. I exemplify these processes with the data of the KORA study. I show, that re-imputing the KORA data twice with current 1000 Genomes reference panels saves about one month of computing time (197 vs. 142 days), when conducting the post-phasing lift over approach (compared to using the pre-phasing lift over approach).

The computational burden of genotype imputation in large scale IPD, which consist of several studies (**mega-imputation**), is even higher compared to imputing the data of a single study. I showed how to overcome this computational demanding task with the data from the IAMDGC, which consists of 52,189 subjects, genotyped at 508,740 variants each. First the parameters used for phase estimation and genotype imputation (with the programs *ShapeIT* and *minimac*) were optimized to impute the genotypes with highest imputation qualities. Second the parallelization and pipelining of the phasing and imputation steps broke the overall burden of 8.5 CPU-years down to less than 3 weeks. I achieved this with an efficient pipeline on a server cluster.

These approach enables study analysts to quickly utilize the latest reference panels and thus allows study analysts to detect genetic loci of complex disease with the latest reference data in reasonable time.

4.2 Comparison to literature

4.2.1 Identifying genetic loci associated with complex disease by meta-analysis of variants imputed with high and low density reference data

My observations are in line with previous observations (see **chapter 2.2**): More variants were available in the analysis in 1000 Genomes imputed data compared to the HapMap imputed data. The number of well imputed common variants was higher in 1000 Genomes imputed data compared to the HapMap imputed data. Common variants present in both 1000 Genomes and HapMap imputed data were imputed with comparable imputation qualities. However, the 1000 Genomes imputed data yielded a high number of well imputed rare variants, which were missing in HapMap imputed data. Overall, the superiority of the 1000 Genomes reference panel compared to the HapMap reference panel was underpinned with the higher number of well imputed common variants and additional well imputed rare variants.

I conducted a systematic literature research using *Pubmed* (see **Web Resources**) to identify comparable work identifying susceptibility loci for complex diseases and comparing the imputation qualities between 1000 Genomes and HapMap imputed data. I used the keywords *1000 Genomes imputation*, *HapMap imputation*, *comparison of imputation qualities*, *rare variant association*, *complex disease*, *genome-wide association meta-analysis*, *high density* and *low density* for my comparison to literature. Overall, only a few articles, including comparisons between HapMap and 1000 Genomes imputations and analyses, were available. An overview over current work is given in **Table 26**. A comparison between HapMap and 1000 Genomes analyses was conducted in four articles. All analyses identified susceptibility loci additional to either previous meta-analyses based on HapMap imputed genotypes or compared to previously reported variants from other work. The lead variants in the additional loci were mainly not rare. Only one work identified rare variants associated with complex disease [72], where a single study was used for both identifying susceptibility loci and comparison of imputation qualities. In one article, the identification of susceptibility loci and comparison of imputation qualities was conducted like I did in the kidney function data, while the number of subjects used to evaluate the influence of HapMap versus 1000 Genomes imputation was conducted in a different number of subjects compared to the detection of susceptibility loci [73]. The study subjects were used for comparing the detection of susceptibility loci for complex disease between 1000 Genomes and HapMap analyses in two works [72, 73], while a comparison to literature was conducted in two other articles [74, 75].

Other work compared state of the art reference panels regarding the absolute number of variants and variant density, emphasizing the increased number of rare variants in the 1000 Genomes compared to the HapMap reference panels [76, 77]. In line with the here presented results, the overall higher density and the increased number of variants (especially rare variants) enabled the evaluation of additional 1000 Genomes imputed data compared to HapMap imputed data.

Table 26. ***Comparison of current literature identifying novel susceptibility loci for complex disease on 1000 Genomes imputed data****. Shown are four articles, which have published novel loci associated with complex diseases. A systematic literature search using PubMed yielded these four articles, which include the identification of novel disease loci for complex disease. If available, also the number of subjects used for comparing imputation qualities between HapMap and 1000 genomes imputed genotypes are given.*

			Number of subjects in association analysis		Additional loci by 1000 Genomes imputed data		Number of subjects to evaluate imputation qualities	
Paper	Outcome	Study/consortium	HapMap	1000 Genomes	Rare variants [MAF<=1%]	Others [MAF>1%]	HapMap	1000 Genomes
Wood et al.	93 quantitative circulating factors	InCHIANTI	1,210	1,210	4	9	1,210	1,210
Nikpay et al.	Coronary heart disease	Cardiogram	63,751 cases 130,681 controls	60,801 cases 123,504 controls	0	10	194,432	184,305
Horikoshi et al	Glycaemic and Obesity-Related	Engage	Literature search	Up to 87,048	0	4	8,078	8,078
Kinnersley et al.	Glioma	SU.VI.MAX, KORA, POPGEN & Heinz Nixdorf Recall, 1958 Birth Cohort, CGEMs	Literature search	5,637 cases 9,158 controls	0	5	--	--

Outcome = outcome analyzed in the literature, study/ consortium: this study or consortium analyzed the data, Number of subjects in association analysis = number of subjects in the 1000 Genomes and HapMap meta-analysis, Additional loci by 1000 Genomes imputed data specifies the number of loci detected by meta-analysis of 1000 Genomes imputed genotypes, which could not be identified in previous meta-analysis on HapMap imputed genotypes, Number of subjects to evaluate the imputation qualities = the number of subjects used for the comparison of the imputation quality in the paper. "—"indicates that differences of imputation quality were not evaluated.

4.2.2 On the gain from mega-analysis compared to meta-analysis

In the comparison of mega-analyzing compared with meta-analyzed study data, I evaluated the imputation qualities between both approaches in **chapter 3.2.3**. I showed that the mega-imputation is superior to the meta-imputation. Work exist to optimize genotype imputation for study specific ethnicities. This was done by imputing genotypes with different reference panels. The imputation qualities were subsequently compared. The performance of different reference panels for genotype imputation were compared in small studies, for example in 443 unrelated individuals from 29 worldwide populations [78] or in 464 Mexican Americans [79] with HapMap reference data. More recent work compare imputation quality of genotypes imputed with multiple 1000 Genomes reference panels (also genotyped on different platforms) in 153 European subjects [77], in 595 African Americans [80] or in 665 Latinos [81]. A comparison between HapMap and 1000 Genomes imputed genotypes in a comparably large set of subjects (~50,000 subjects) was not yet performed.

In my work I illustrated the gain of mega-analyzing data compared to meta-analyzing it. In the following I summarize examples (of numerous work) of literature related to this comparison between mega-analysis and meta-analysis. Mega-analyses investigating the genetics of complex disease, pooling phenotypic and genotypic data across studies, were reported for other phenotypes, such as bipolar disorder. 3 studies were imputed separately (meta-imputed) with a HapMap reference panel and the imputed genotypes were pooled for the mega-analysis in 2,836 cases and 2,744 controls [67]. **But no mega-imputation across all study subjects and a comparison between meta-imputed with mega-imputed genotypes was performed**. Other work on bipolar disorder identified additional susceptibility loci by mega-analyses of 10,596 and 18,190 subjects in a case-control setting, also without comparing imputation qualities [82, 83]. Another phenotype analyzed by an IPD analyses was depressive disorder. A mega-analysis in 32,050 subjects identified additional loci associated with disease [68]. While meta-analyses of GWAS for kidney function, renal disease and eye disease exist, there were no published mega-analyses to investigate the genetic background of health and disease for kidney or eye related phenotypes.

The differences between the results of meta-analysis and mega-analysis were investigated in the statistical literature. It was shown in theory that regression analysis on IPD compared to meta-analysis of study specific regression results are equivalent [84, 85]. It was furthermore shown, that the effect estimate from meta-analysis is a good estimate for analyzing data jointly in theory [86] and in small studies with 4,792 and 9,791 subjects [87, 88].

Finally, there was also work on non-genetic analyses: There are other IPD analyses for complex outcomes such as obesity [89], pregnancy [90], celial disease [91] or kidney function [92] in non-genetic analyses. A comparison of effect estimates obtained from a meta-analysis of published data compared to an mega-analysis using IPD showed excellent quantitative agreement between the summary effect

estimates of the risk for ovarian cancer from the meta-analysis compared to the mega-analysis in a non-genetic analysis [93].

In summary, in the aforementioned literature the statistical theory of meta-analysis and mega-analysis is investigated and exemplifies the differences of their results in a small number of subjects. But there is no investigation on comparing the influence of meta-imputed compared to mega-imputed genotypes on the results of meta-analysis and mega-analysis in a large number of subjects, was conducted in this thesis.

4.2.3 Software and approaches to accelerate genotype imputation

I introduced software and approaches to accelerate genotype imputation for re-imputing single studies and to parallelize genotype imputation for large scale IPD.

The *PhaseLift* software lifts the genome build of haplotypes prior to genotype imputation and the approach of massively parallelizing the phasing and imputation enables the imputation of large scale genome wide data. For the lift over of genomic builds, there are online tools from the University of California (*liftOver*) and the National Center for Biotechnology Information (*NCBI remapping service*, see **Web Resources**). The latter offers an offline version, which can be installed on servers to lift genetic data from one build to another. This offline version was also part of my published software *PhaseLift*. Additionally the web-based platform for data intensive biomedical research tool *Galaxy* also offers a lift over functionality (for these three tools please see **Web Resources**). With my software *PhaseLift* the build of the study data is lifted to the build of the reference panel without the use of online tools in a fast and comprehensive way.

Current literature provided practical guide lines to facilitate genotype imputation with low density and high density reference panels for study analysts, which included the workflow from quality control, phase estimation, genotype imputation to association analysis. Study analysts can use check lists for study quality control of genotypes and phenotypes and it was shown how to prepare the genotypes for phase estimation and genotype imputation with HapMap reference data [94], on meta-analysis mastering the challenges of 1000 Genomes imputed genotypes [95] or for example on guidelines for the most commonly used programs for phase estimation and genotype imputation *ImputeV2* and *minimac* [96, 97]. In these articles, guidelines and check lists also consisted of harmonizing the genotype build in the study data to the genotype build of the reference panel. However, the task of harmonizing study data with reference panel was not addressed in detail. On the other hand, the lift over of genotype annotation is included in the imputation pipeline *Molgenis-impute*, which automates the set up and running of all the steps of the imputation process [98].

While I concentrate in my thesis mainly on *ShapeIT* for phase estimation and *minimac* for genotype imputation, current existing programs, used for phase estimation and genotype imputation, were compared in the literature. The performance of different approaches for phase estimation and

genotype imputation was compared for subjects genotyped on different genotyping platforms [99] or to establish a novel software. (for example *ShapeIT* [28] or *HapiUR* [24]) or genotype imputation (*ImputeV2* [59] and *Beagle* [23]).

The used computational resources used by the software *minimac* was optimized to conduct genotype imputation even faster on multiple cores in parallel [100]. Other work recommended to parallelize the GWAS analyses with the R-package *doPar* in parallel, but did not consider time or memory constraints [101]. With protocols for genotype imputation it was illustrated how to conduct quality control, phase estimation and genotype imputation also covering the harmonization of builds between study data and reference panel. The focus in these work was not on optimizing the time to imputation with detailed investigations on time and memory constraints [31, 96, 102].

In contrast, in my work, I showed how time can be saved in re-imputing study data with different panels and I illustrated how much time can be saved. Additionally, I showed how the time to imputation of large scale data can be minimized. Here I explicitly analyzed the hardware resources needed for this computationally demanding task.

4.3 Relevance of my work to complex disease genetics

Generally it is the strength of my work, that I analyzed the gain of the genotype imputation methodology. This cost effective method allows study analysts to maximize the available resources, as it utilizes software for phasing, imputation and reference panels, which are free of charge.

4.3.1 Identifying genetic loci associated with complex disease by meta-analysis of variants imputed with high and low density reference data

The identification of susceptibility loci was conducted in data from the CKDGen consortium in one of the largest genetic data sets on kidney function worldwide with more than 100,00 subjects.

I showed, that in the 1000 Genomes imputed study data yield a higher number of well imputed variants and well imputed rare variants, which are not available in HapMap imputed data. This is due to the higher density of variants in the 1000 Genomes reference panels compared to the HapMap reference panels. More specifically my findings add novel knowledge about the complex genetic architecture of kidney function. I showed here that the map of susceptibility loci for kidney function is enhanced by meta-analyzing 1000 Genomes imputed data compared to meta-analyzing HapMap imputed data. The 1000 Genomes panel covers genetic regions, which were not covered by the HapMap reference data, which lead to the discovery of six novel susceptibility loci for kidney function. I additionally identified four risk loci for kidney function, due to a better study composition compared to the HapMap study composition. These loci were missed by previous HapMap meta-analyses.

4.3.2 On the gain from mega-analysis compared to meta-analysis

The comparison between the mega-analyzed and meta-analyzed data was conducted in one of the largest IPD sets worldwide with more than 50,000 subjects from the IAMDGC.

The study data was imputed jointly and per study. The comparison of the imputation qualities from the jointly imputed were compared with the imputation qualities from the data set imputed by study. The mega-imputation contained a higher number of well imputed variants compared to the meta-imputation. The mega-imputation yielded no gain for common variants compared to the meta-imputation. There was a gain in imputation quality in rare variants in the mega-imputed data compared to the meta-analysis. No comparison between mega-imputed with meta-imputed genotypes in such large scale study data set has been performed, yet.

I compared the commonly used approach of meta-imputing and meta-analyzing study data in consortia with mega-imputing and mega-analyzing the data jointly for disease loci detection. The mega-imputation/ analysis approach identified 34 disease loci for AMD with genome-wide significance. Here I show, that eight of these 34 loci were missed, when the same data was meta-imputed/ analyzed. When mega-analyzing the mega-imputed variants, six (of the eight) loci were identified. These result showed, that the mega-approach was essential for the identification of the 34 AMD disease loci.

In summary, my work showed, that the mega-imputation of a large scale data set can generally increase the overall number of well imputed variants and can specifically increase the number of well imputed rare variants. The mega-approach was essential to identify the 34 AMD disease loci.

4.3.3 Software and approaches to accelerate genotype imputation

As novel reference panels are constantly released, there is a need to constantly re-impute study data to the latest reference panel. I have shown here, that novel reference panels allow the detection of novel susceptibility loci for complex disease, compared to using previous reference panels. Thus it is essential, that study analysts can (re-) impute study subjects to the latest reference panel in a fast and efficient way. This is especially true for large study data sets, which yield an additional high computation burden for genotype imputation. Thus, I introduced software and approaches to accelerate genotype imputation for re-imputing single studies and to parallelize genotype imputation for large scale IPD. This allows study analysts to unravel novel loci for complex disease.

First, the technological advances will result in constantly released reference panels, which will be annotated on different builds. *PhaseLift* applies all changes of any novel build to estimated study haplotypes very fast. This makes it very relevant for study analysts as it save time when re-imputing study haplotypes with any novel reference panel, annotated on any novel build.

Second, the computational burden for the analysts will rise, because more variants will be genotyped per study subjects and because reference data will consist of more subjects and variants.

This will ultimately lead to bigger data sets and thus to higher computational burdens for genotype imputation. My optimization of the imputation process allows to diminish the computational burdens of genotype imputation of single studies and for mega-imputation of multiple studies, which underscores the timeliness and relevance of my work.

4.4 Strength and limitations

The major strength of this work is, that it fills an important gap for the analysis of complex disease with genome wide data: It is the first structured work that quantifies the gain of using 1000 Genomes imputed study data in meta-analysis as well as in mega-analysis. My investigations furthermore include both a continuous trait (kidney function) as well as a binary trait (AMD) in real large scale data. Here I identified disease loci associated with complex disease with real data and developed methods and software to accelerate genotype imputation.

4.4.1 Identifying genetic loci associated with complex disease by meta-analysis of variants imputed with high and low density reference data

It is a strength of my investigations that I analyzed real large scale data on kidney function, which is an important measure for kidney disease. My analyses consisted of one of the largest genetic data set on kidney function worldwide. Genotypes of more than 130,000 subjects imputed with HapMap reference data and of more than 110,000 subject imputed with 1000 Genomes reference data were analyzed. My results adds novel knowledge on the genetic architecture of kidney function, which is ultimately needed in diagnosis, treatment and prevention of chronic kidney disease. My analysis is also the first large scale comparison of HapMap and 1000 Genomes imputed meta-analyses on kidney function. There are only few works known including a comparable number of subjects in the analysis of other complex disease traits. In fact there is only one, which compares the HapMap with the Genomes meta-analysis in analyses of comparable number of subjects.

A limitation of my work is, that a direct comparison of the exact number of overlapping subjects between both meta-analyses was not conducted. Due to study specific reasons, the number of analyzed subjects may differ in a study between HapMap and 1000 Genomes GWAS. The logistic effort to re-impute with both HapMap and 1000 Genomes reference panels, re-analyze per study and re-meta-analyze across all overlapping subjects, would be a burden for study analysts and is beyond the scope my work. Also fine mapping of the known loci could be improved to get more information about each locus. But my focus was more on signal detection than on fine mapping.

4.4.2 On the gain from mega-analysis compared to meta-analysis

As privacy restriction mostly prohibit to pool the phenotypes and variants across several studies, large scale IPD is rare. It is thus the strength of my work that the investigations are done in one of the largest

IPD in the world with more than 50,00 subjects, which yields highly associated variants with a large variety of allele frequencies and effect sizes. Furthermore, there is no work that compares the mega-imputation with the meta-imputation of a large scale data, as I did in this work. An important question for future study designs is how all data are collected from all studies. Privacy restrictions of a study can prohibit to share the phenotype and genotype of the study participants with study analysts in a consortium. This means, that not all studies could contribute to the mega-analysis.

It is a limitation of my work, that the meta-analysis has finer statistical modelling aspects, which needs a lot of software, for example to detect interaction and to perform stratified analyses. Statistical approaches are missing in meta-analyses, which represent the complete model. Signal detection and type I error was my focus, where I found no bias. Model assumptions should be checked if the homogeneity of covariate effects across studies are given. If yes, it would be of interest if they would be large enough to have any effect on the effect estimates. One must be aware, that the model does not accounted for the α-estimate of the case-control-ratio. The impact of this modelling on association results were not evaluated in my work. A limitation of my work is furthermore, that quality control on study level is not included in my analyses, which would reflect a realistic scenario in consortia. Also, genotype imputation of the IAMDGC data set allows for the analysis of haplotypes. I did not analyze haplotypes to identify haplotypes associated with AMD across the whole genome. But my focus was on signal detection, analysis of single variant analysis and genotype imputation.

4.4.3 Software and approaches to accelerate genotype imputation

The implemented software *PhaseLift* is a very practical solution for study analysts that reflects the current and the future technological advances. It is a strength of my work that it does not only account for build changes of current data, but is furthermore suitable for data released in the future. *PhaseLift* is an easy-to-use software package that accelerates the re-imputation of study data with novel reference panels. Already today the *GRCh* build 38 build has been introduced. It is a strength of *PhaseLift* that it is applicable to any novel released build (as the *GRCh* build 38) and emphasizes the relevance and strength of my work. It is limitation of my work, that *PhaseLift* does not include a framework for quality control of variants and subjects or a pipeline for phase estimation and genotype imputation, but it was the clear focus of my work to provide a solution for the specific problem of unharmonized study data with the reference panel prior to genotype imputation.

My optimization of the mega-imputation was conducted on a high number of subjects in real data. Although the results were very specific for the variants from the IAMGC, it may pinpoint the essential limiting constraints and can thus guide study analysts for future mega-imputations.

It is a limitation of my work, that my investigations presume, that a multi core server cluster with 176 cores and at least 40 GB of volatile memory are available. But mega-imputation cannot be conducted on desktop PCs in reasonable time, which makes the presence of high throughput

computation facilities mandatory. If study partners do not have sufficient resources to cope with the high computational burdens of mega-imputation, it might look reasonable to conduct genotype imputation on commercial server solutions (for example Google or Amazon cloud). Again, privacy restrictions play a crucial role to decide if this can be done. This option is not discussed in my work.

4.5 Conclusion and Outlook

In conclusion, I showed in my work that our knowledge of the genetic architecture of complex diseases can be enhanced with methods and software for genotype imputation with high density reference panels.

In my investigations of the CKDGen data on kidney filtration rate I find that highly associated disease loci can be identified additional to genetic loci previously reported in analyses of even more subjects. Although my analyses add valuable results to a well-studied topic, meta-analyses of variants imputed with high density reference panels will help scientists to reveal even more of the genetic architecture of complex diseases in the future. The ongoing technological progress in genotyping und sequencing technologies will result in decreasing costs of genotyping or sequencing genetic variants of subjects. As a consequence novel reference data will be constantly released.

For example the 1000 Genomes project released a novel reference panel with even more subjects and more variants (the 1000 Genomes Phase III version 5 panel). Also the UK10K group and the HRC have generated reference data of increasing size for all reference panels (see **Web Resources)**. The latter is momentarily only accessible by uploading sensible patient data to a central imputation server. This process violates current privacy restrictions of many studies. Thus, it will be important in the future to ensure at the same time, that the latest advances in genetic epidemiology are available for scientists around the world and to maintain the right of privacy of sensible patient data.

On the other hand it is worthwhile to overthink the classical approach in consortia to meta-impute and meta-analyze study specific data, but to pool the data and mega-impute and mega-analyze them. As shown with the data of the IAMDGC, additional susceptibility can be identified by replacing the commonly used approach of meta-imputing and meta-analyzing study data with mega-imputing and mega-analyzing it. This is also only possible, if the privacy of the generated individual participant data is guaranteed. Despite this requirement, my work may guide other consortia to overcome the commonly used practice of meta-analyzing data and to pool study specific data into one large data set to leverage the discovery of the genetic background of complex traits.

The provided software and approaches in my work supports study analysts to save time in re-imputing study data for association analysis and to facilitate the computational demanding task of imputing large scale genome wide data. My research pinpoints the needed resources of a large scale data set and may guide study analyst to infer the needed resources for their own analyses. Advances in sequencing and genotyping techniques will ultimately lead to both more subjects and more variants in both study data and reference panels. Mega-imputation will then need even more resources, such as at least 100 GB volatile memory and more than 2 TB storage. Phase estimation and genotype imputation can then be conducted in 1 or 2 months, depending on the resources for parallelization

(number of cores and the amount of volatile on the server cluster) and technical advances in the methodologies (better programs and methods).

My work focuses mainly on 1000 Genomes reference panels. While these panels are cross sectional data of the general population, the next methodological steps may include the generation and use of disease based reference panels, to investigate the role of disease specific mutations in disease specific population in more detail. Additionally, my work mainly concentrates on the analysis of the cross sectional quantification of kidney function and eye disease and can be enhanced in the future by analyses of progression of disease, which can further help to elucidate our understanding of the mechanisms, which cause physiological processes of complex disease.

Overall, my work shows how the insights into the genetic architecture of complex disease can be enhanced with imputation methodology. My results are applicable to future meta-analyses as well as mega-analyses. In identifying additional susceptibility loci for complex disease, the biological fundamental of health and disease can inseminate personal health care and pharmaceutical research, improving diagnosis, treatment and prevention of complex disease, which will finally lead to an increased health in the general population.

5 Summary

Genotype imputation infers variants, which are not directly assayed in study subjects, by matching inferred study haplotypes with those in external reference panels. Inferred variants are subsequently used to identify genetic loci associated with complex disease. Technological advances in genotyping and sequencing technologies have created a novel generation of high density reference panels, which enable to shed additional light on the genetic architecture of complex traits. But the gain in using these novel high density reference panels for genotype imputation and association analysis is still missing. Thus this work focused on identifying the gain in analyzing variants imputed with high density reference panels compared to analyzing variants imputed with low density reference panels in large scale genome wide data.

I showed in my work how high density reference panels increase our knowledge of the genetic maps on kidney function and AMD. I further developed a tool to assist study analysts for imputing genome wide data for meta-analyses in consortia and I optimized the imputation of untyped variants in individual participant data of large scale.

First, I compared a meta-analysis of variants imputed with HapMap reference panels with variants imputed with 1000 Genomes reference panels in data from the CKDGen consortium. The comparison of imputation qualities evidenced the overall superiority of the imputation with the 1000 Genomes reference panels and illustrates the increased possibility to detect rare variants associated with complex disease. The meta-analysis on kidney filtration rate of the variants imputed with the 1000 Genomes reference panel confirm the majority of previously reported susceptibly loci for kidney function and furthermore allow the identification of 10 additional loci.

Second, I quantified the gain in mega-imputing and mega-analyzing individual participant data compared to meta-imputing the same data per study and meta-analyzing study specific effect estimates. For this analysis I used one of the world's largest individual participant data set from the IAMDGC. I illustrated that the imputation quality of untyped variants imputed jointly across all studies is superior to the imputation quality of variants imputed separated by study and showed that there is a gain of mega-analyzing imputed variants compared to meta-analyzing the same imputed variants. This gain is even bigger, when mega-analyzing variants imputed jointly is compared to mimicking a realistic scenario in consortia of meta-analyzing variants imputed per study.

Third, I facilitate the computational demanding task of genotype imputation with the software *PhaseLift*, which harmonizes phased study haplotype with any reference panel on any build. This enables study analysts save time in re-imputing study data. Study analysts perform the computational intensive phase estimation once and re-impute the study haplotypes with any novel reference panel on any novel genomic build, without repeating the tedious phase estimation. In optimizing the mega-

imputation of large scale genome wide variants across several studies and identifying parameter and constraints for genotype imputation, I assist study analysts to overcome the computational demanding task of imputing large genome wide data.

In summary, genome wide association analyses on variants imputed with high density reference panels further chart the genetic map of complex traits, which will ultimately lead to an increased understanding of the biological mechanisms in the health and disease and to improve diagnosis, treatments and prevention of complex disease for patients.

6 Zusammenfassung

Genotyp-Imputation schätzt genetische Varianten in Teilnehmern einer Studie, die nicht direkt gemessen wurden, indem sie übereinstimmende Abschnitte der Haplotypen in den Studienteilnehmern mit denen aus externen Referenz-Haplotypen zur Deckung bringt. Diese geschätzten Varianten können anschließend genutzt werden um genetischer Regionen zu identifizieren, die hoch assoziiert sind mit komplexen Erkrankungen. Technologische Fortschritte in der Genotypisierungs- und Sequenzierungstechnik haben eine neue Generation von Referenzdaten mit hoher Dichte hervorgebracht, die versprechen die genetische Architektur komplexer Erkrankungen näher zu beleuchten. Der Gewinn, den diese neuen Referenzdaten mit sich bringen wurde jedoch noch nicht eingehend untersucht. Deswegen ist es das Ziel meiner Arbeit aufzuzeigen, welchen Gewinn die Verwendung dieser neuen Generation von Referenzdaten mit sich bringt.

Ich zeige in meiner Arbeit, wie die Verwendung dieser neuen Generation von Referenzdaten unsere Kenntnisse über die genetische Architektur, insbesondere der Nierenfunktion und der AMD erweitert. Des Weiteren entwickelte ich ein Programm, das den Prozess der Genotyp-Imputation beschleunigt und ich optimierte die Genotyp-Imputation sehr großer genomweiter Datensätze.

Anhand von Daten aus dem CKDGen Konsortium verglich ich zunächst die Meta-Analyse von Varianten, die mit HapMap Referenzdaten imputiert wurden mit der Meta-Analyse von Varianten, die mit 1000 Genomes Referenzdaten imputiert wurden. Der Vergleich der Imputationsgüte unterstreicht die Überlegenheit der Imputation mit 1000 Genomes Referenzdaten gegenüber der Imputation mit HapMap Referenzdaten und zeigt auf, dass es mit Hilfe der 1000 Genomes Referenzdaten mit höherer Wahrscheinlichkeit genomweite Assoziationen mit seltenen genetischen Varianten geben kann. Die Meta-Analyse bezüglich Nierenfiltration konnte die Mehrzahl der bekannten genetischen Regionen zur Nierenfiltration verifizieren und erlaubte darüber hinaus auch die Identifizierung 10 zusätzlicher Regionen.

Des Weiteren quantifizierte ich den erzielten Gewinn wenn ein großer genomweiter Datensatz zusammen imputiert und analysiert wird im Gegensatz dazu, dass die Imputation und die Assoziationsanalyse nach Studien getrennt berechnet wird und die Effektschätzer anschließend in einer Meta-Analyse vereint wurden. Für diese Analyse stand mir einer der weltweit größten Datensätze aus Einzelpersonen aus dem IAMDGC zur Verfügung. Zunächst zeigte ich, dass die Imputation aller Studienteilnehmer zusammen höhere Imputationsqualität generiert, verglichen mit der Imputation getrennt nach Studie. Ich zeigte, dass man bessere Assoziationsergebnisse erhält, wenn man gemeinsam imputierte Varianten auch gemeinsam auswertet, verglichen damit, dass man die gemeinsam imputierten Varianten getrennt nach Studie auswertet und danach meta-analysiert. Noch bessere Ergebnisse erzielt man, wenn man zunächst alle undetektierten Varianten zusammen

imputiert und auch zusammen auswertet im Gegensatz dazu, dass man den die undetektierten Varianten getrennt nach Studien imputiert und auswertet, um die studienspezifischen Effektschätzer danach zu meta-analysieren.

Mit der Software *PhaseLift* beschleunigte ich die Genotyp-Imputation. *PhaseLift* harmonisiert die geschätzten Haplotypen der Studienteilnehmer mit Referenzdaten, die auf einer beliebigen Annotation sein können. Dadurch entfällt die rechenintensive, wiederholte Haplotypschätzung, wenn Studiendaten mit mehreren Referenzdaten imputiert werden sollen. Durch die Optimierung der Imputation von hochdimensionalen genomweiten Datensätzen auf großen Server-Clustern und durch die Identifizierung der benötigten Rechenressourcen dafür, ermöglichte ich es Studienanalysten die recheninstensive Genotyp-Imputation auch auf hochdimensionale Daten anzuwenden.

Zusammenfassend haben meine Untersuchungen ergeben, dass die Genotyp-Imputation mit der neuen Generation von Referenzdaten nicht nur unsere Kenntnisse über die genetische Architektur komplexer Erkrankungen erhöht, sondern auch, dass durch verbesserte Analysemethoden und Software die genetische Epidemiologie zu einem besseren Verständnis von Krankheiten und krankheits-relevanten Merkmalen und letztendlich auch zu einem besseren Verständnis der Ursachen und Entstehung von komplexen Erkrankungen beiträgt.

7 Appendix

7.1 Effective number of subjects in novel and known loci

The power analysis in **chapter 2.2.3** shows that ten lead variant can be identified in the 1000 Genomes meta-analysis, which were not identified in the HapMap meta-analysis. Of them, six were available in the 1000 Genomes reference panel, but not in the HapMap reference panel. This was underpinned when comparing the effective number of subjects (number of subjects per variant * imputation quality) between the HapMap and 1000 Genomes meta-analysis, which is illustrated in **Figure 22**. The **six lead variants in the ten additional loci** (in genes *HOXD8*, *ARL15*, *PIK3R1*, *EYA4*, *ASTN2* and *EBP41L3*) had a very low effective number of subjects in the HapMap meta-analysis (number of subjects range between 1,871 and 13,031) compared to the effective number of subjects in the 1000 Genomes meta-analysis (red dots in the upper triangle). The 4 other loci (in genes *LPHN2*, *RHOC*, *SLC7A6* and *RNF152*) are present in both reference panels. Their effective number of subjects is higher in the HapMap meta-analysis (red dots in lower triangle).

The effective number of subjects in the lead variants from the HapMap meta-analysis (green dots in **Figure 23)** is in all but three known loci higher in the HapMap meta-analysis due to their higher number of subjects. Their median imputation qualities range from 0.50 to 0.99**.**

Among the **53 lead variants from the 1000 Genomes meta-analysis (in the 53 loci identified by the HapMap meta-analysis,** blue dots in **Figure 23**), there are 32 variants with higher effective number of subjects (in the upper triangle) in the 1000 Genomes meta-analysis, as they are not tagged in the HapMap reference panel. The other 21 (of 53) top variants have higher effective number of subjects (due to the higher number of subjects) in the HapMap meta-analysis (in the lower triangle).

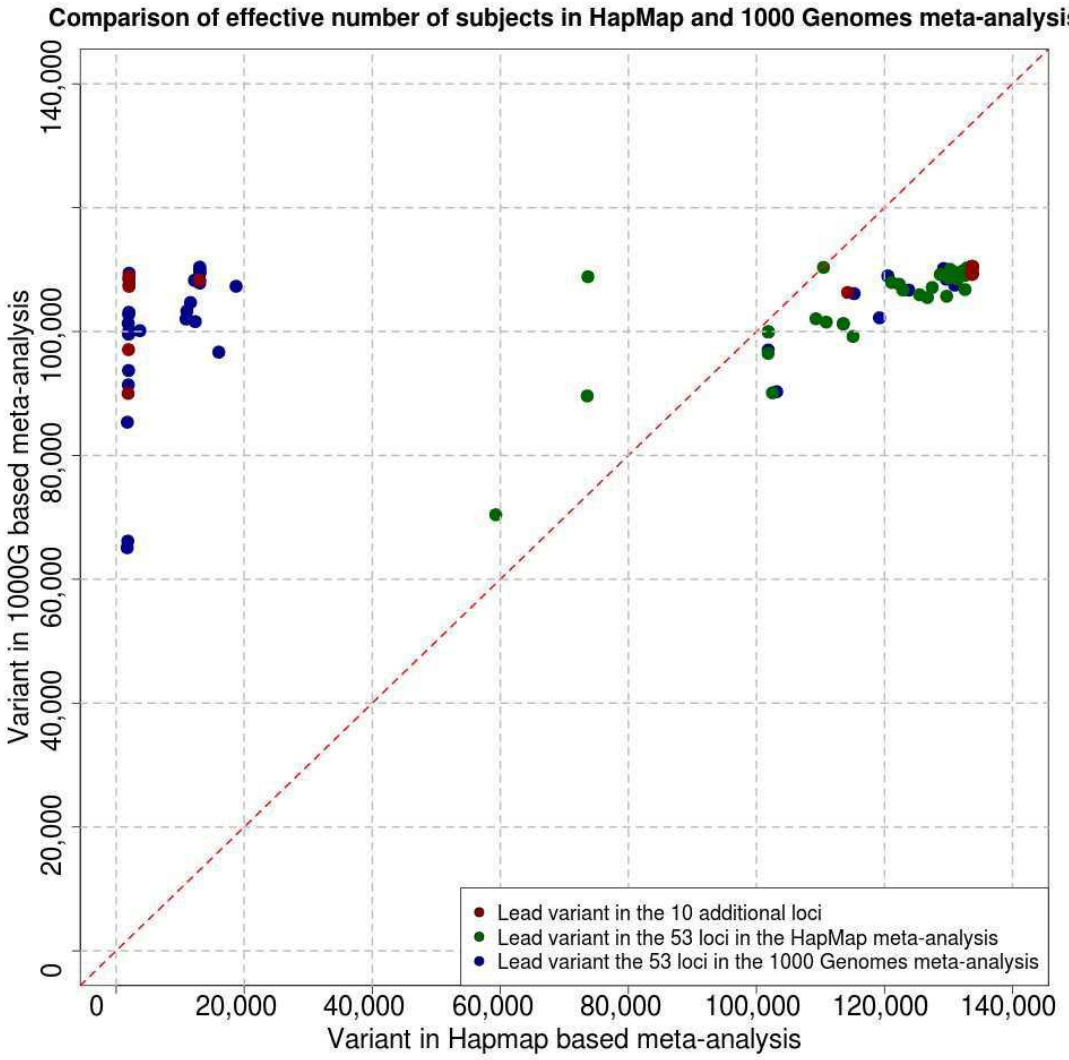

Figure 23. ***Comparison of effective number of subjects in the 1000 Genomes and HapMap meta-analysis results.*** *The effective number of subjects (number of subjects * imputation score) is compared between the HapMap (x) and the 1000 Genomes (y) based meta-analysis results. Variants are color coded by category: 10 top variants in novel loci (red), 53 reported variants in known loci (green) and 53 top variants in known loci (blue).*

A complete over view of all lead variants with their median imputation quality, number of studies that contributed to the meta-analysis and the effective number of subjects in the analysis separated by HapMap meta-analysis and 1000 Genomes meta-analysis can be found in **Table 27**.

Table 27. ***Overview over the median imputation qualities and effective number of subjects of the top variants in the 53 known and 10 novel kidney function disease loci from the HapMap and 1000 Genomes meta-analysis.*** *Shown are the median imputation qualities, the number of analyzed studies and the effective number of subjects of all top variants in novel loci, reported variants in known loci and top variants in known loci from the HapMap meta-analysis and the 1000 Genomes meta-analysis results.*

				HapMap meta-analysis			1000 Genomes meta-analysis		
VariantID	Chr	Position (bp)	Closest Gene	Median imp quality	# stu dies	Effective num subjects	Median imp quality	# stu dies	Effective num subjects
Top variant in locus identified additional to loci identified by HapMap meta-analysis									
rs10874312	1	82,944,571	*LPHN2*	0.9985	46	133,602	1.0000	29	109,424
rs12144044	1	113,248,791	*RHOC*	0.8552	48	114,238	0.9620	33	106,317
rs187355703	2	176,993,583	*HOXD8*	0.9560	1	1,934	0.8886	31	97,085
rs111366116	5	53,295,546	*ARL15*	0.9930	1	2,009	0.9709	33	107,301
rs113246091	5	67,739,274	*PIK3R1*	0.9896	1	2,002	0.9850	32	108,451
rs7764488	6	133,812,872	*EYA4*	0.9964	1	2,016	0.9849	33	108,849
rs13298297	9	119,264,108	*ASTN2*	0.9249	1	1,871	0.8145	33	90,016
rs1111571	16	68,363,181	*SLC7A6*	0.9998	46	133,779	1.0000	31	110,515
rs9962915	18	5,593,171	*EPB41L3*	0.9956	3	13,031	0.9801	33	108,312
rs12458009	18	59,350,507	*RNF152*	0.9994	48	133,686	0.9982	29	109,220
Variant with smallest p-value in locus identified by HapMap meta-analysis									
rs12124078	1	15,869,899	*CASP9*	0.9974	49	133,376	0.9991	33	110,413
rs12136063	1	110,014,170	*SYPL2*	0.9986	49	133,536	0.9983	33	110,326
rs267734	1	150,951,477	*ANXA9*	0.9929	48	132,772	0.9896	33	109,369
rs3850625	1	201,016,296	*CACNA1S*	0.9895	46	132,264	1.0000	29	109,422
rs2802729	1	243,501,763	*SDCCAG8*	0.8614	50	115,094	0.8978	33	99,220
rs6431731	2	15,863,002	*DDX1*	0.5954	47	73,571	0.8202	31	89,613
rs1260326	2	27,730,940	*GCKR*	0.9907	48	132,529	0.9967	29	109,057
rs13538	2	73,868,328	*NAT8*	0.9897	48	128,710	0.9883	33	109,220
rs4667594	2	170,008,506	*LRP2*	0.9939	49	132,897	0.9940	33	109,849
rs7422339	2	211,540,507	*CPS1*	0.7675	49	101,866	0.9044	33	99,954
rs2712184	2	217,682,779	*IGFBP2*	0.9997	47	133,755	1.0000	29	109,424
rs6795744	3	13,906,850	*WNT7A*	0.9701	50	129,713	0.9565	33	105,705
rs347685	3	141,807,137	*TFDP2*	0.9999	48	133,710	1.0000	33	110,511
rs9682041	3	170,091,902	*SKIL*	0.9977	48	133,506	0.9998	29	109,398
rs10513801	3	185,822,353	*ETV5*	0.9930	50	132,783	0.9881	33	109,201
rs17319721	4	77,368,847	*SHROOM3*	0.9878	49	132,073	0.9904	33	109,457
rs228611	4	103,561,709	*MANBA*	0.9984	48	133,567	0.9899	31	109,393
rs11959928	5	39,397,132	*DAB2*	0.9725	50	130,029	0.9831	33	108,651
rs6420094	5	176,817,636	*SLC34A1*	0.9936	46	130,232	0.9960	31	110,071
rs7759001	6	27,341,409	*ZNF204*	0.9947	50	133,016	0.9897	33	109,379
rs881858	6	43,806,609	*VEGFA*	0.8298	50	110,907	0.9187	33	101,529
rs2279463	6	160,668,389	*SLC22A2*	0.9840	50	131,571	0.9851	33	108,870
rs10277115	7	1,285,195	*UNCX*	0.4977	46	59,291	0.6372	33	70,425
rs3750082	7	32,919,927	*AVL9*	0.9955	49	126,711	0.9545	33	105,493
rs6465825	7	77,416,439	*TMEM60*	0.9998	48	133,775	0.9998	29	109,403
rs7805747	7	151,407,801	*PRKAG2*	0.5592	48	73,681	0.9850	32	108,852
rs6459680	7	156,258,568	*AC005534.6*	0.9783	50	130,790	0.9923	33	109,671
rs10109414	8	23,751,151	*STC1*	0.9906	50	132,419	0.9939	33	109,846
rs4744712	9	71,434,707	*PIP5K1B*	0.9972	50	133,343	0.9980	33	110,290
rs10794720	10	1,156,165	*WDR37*	0.9984	47	133,565	0.9996	31	110,473
rs10994860	10	52,645,424	*A1CF*	0.9189	50	122,863	0.9655	33	106,705
rs163160	11	2,789,955	*KCNQ1*	0.9802	50	131,074	0.9875	33	109,132
rs3925584	11	30,760,335	*MPPED2*	0.9901	50	132,354	0.9938	33	109,833
rs4014195	11	65,506,822	*RNASEH2C*	0.9936	49	132,867	0.9935	33	109,797
rs10774021	12	349,298	*SLC6A13*	0.9982	48	133,558	0.9995	29	109,365
rs10491967	12	3,368,093	*TSPAN9*	0.9791	48	131,004	0.9987	29	109,280
rs7956634	12	15,321,194	*RERG*	0.9938	49	132,898	0.9980	33	110,294
rs1106766	12	57,809,456	*R3HDM2*	0.8266	48	110,481	0.9986	33	110,364
rs626277	13	72,347,696	*DACH1*	0.9971	49	133,416	0.9981	31	110,305
rs2928148	15	41,401,550	*INO80*	0.9981	50	133,457	0.9954	33	110,012

rs2453533	15	45,641,225	*GATM*	0.9997	50	133,682	0.9999	33	110,511
rs491567	15	53,946,593	*WDR72*	0.9928	50	132,755	0.9929	33	109,731
rs1394125	15	76,158,983	*UBE2Q2*	0.9383	48	125,495	0.9682	29	105,940
rs12917707	16	20,367,690	*UMOD*	0.9532	50	127,453	0.9692	33	107,117
rs164748	16	89,708,292	*DPEP1*	0.9914	50	132,572	0.9663	33	106,787
rs2453580	17	19,438,321	*SLC47A1*	0.7728	49	102,504	0.8154	33	90,115
rs11078903	17	37,631,924	*CDK12*	0.9212	49	122,311	0.9740	33	107,642
rs9895661	17	59,456,589	*BCAS3*	0.8507	49	113,560	0.9162	33	101,259
rs8091180	18	77,164,243	*NFATC1*	0.7807	48	101,825	0.8900	31	96,479
rs12460876	19	33,356,891	*SLC7A9*	0.9849	50	131,678	0.9901	33	109,426
rs11666497	19	38,464,262	*SIPA1L3*	0.9516	48	121,111	0.9766	33	107,926
rs6088580	20	33,285,053	*TP53INP2*	0.9894	49	131,363	0.9815	33	108,465
rs17216707	20	52,732,362	*BCAS1*	0.8177	49	109,291	0.9236	33	102,067
Variant with smallest p-value from 1000 Genomes meta-analysis in locus identified by HapMap meta-analysis									
rs7546668	1	15,855,123	*CASP9*	0.9986	3	13,069	0.9923	33	109,663
rs10127790	1	109,891,133	*SYPL2*	0.9018	50	120,524	0.9862	33	108,990
rs267738	1	150,940,625	*ANXA9*	0.9989	47	133,645	1.0000	29	109,425
rs2783971	1	243,474,536	*SDCCAG8*	0.9786	50	130,842	0.9771	33	107,991
rs807601	2	15,793,014	*DDX1*	0.9795	48	130,972	0.9726	33	107,491
rs780093	2	27,742,603	*GCKR*	0.9984	50	133,495	0.9973	33	110,219
rs4500972	2	73,767,897	*NAT8*	0.9702	1	1,963	0.9010	33	99,576
rs35472707	2	169,995,581	*LRP2*	0.8439	3	11,047	0.9454	31	103,293
rs1047891	2	211,540,507	*CPS1*	0.9384	1	1,898	0.9044	33	99,954
rs2541381	2	217,683,836	*IGFBP2*	0.9969	3	13,050	0.9757	33	107,828
3:13918234 :INDEL	3	13,918,234	*WNT7A*	NA	0	0	0.9383	31	86,823
rs7640665	3	141,813,172	*TFDP2*	0.9675	1	1,957	0.9300	33	102,781
rs6770214	3	171,006,768	*SKIL*	0.9596	4	18,732	0.9710	33	107,312
rs6809651	3	185,814,642	*ETV5*	0.9991	48	133,675	0.9996	31	110,474
rs13146355	4	77,412,140	*SHROOM3*	0.9664	50	129,234	0.9969	33	110,174
4:103573122 :INDEL	4	103,573,122	*MANBA*	NA	0	0	0.9803	30	88,322
rs700236	5	39,367,739	*DAB2*	0.9993	1	2,022	0.9901	33	109,426
rs3812036	5	176,813,404	*SLC34A1*	0.8917	49	119,217	0.9250	33	102,228
rs9348765	6	27,314,650	*ZNF204*	0.8132	48	101,835	0.8778	33	97,015
rs1317983	6	43,806,335	*VEGFA*	0.9441	3	12,357	0.9194	33	101,611
rs2279463	6	160,668,389	*SLC22A2*	0.9840	50	131,571	0.9851	33	108,870
rs62435145	7	1,286,567	*UNCX*	0.8994	1	1,819	0.5987	33	66,168
7:33113699 :INDEL	7	33,113,699	*AVL9*	NA	0	0	0.9900	31	91,611
rs112029703	7	77,238,678	*TMEM60*	0.9983	1	2,020	0.9824	33	108,567
rs10254101	7	151,415,536	*PRKAG2*	0.8337	3	10,913	0.9232	33	102,024
rs6971211	7	155,664,686	*AC005534.6*	0.9971	47	132,495	0.9993	29	109,345
rs36071802	8	23,715,871	*STC1*	0.8867	3	11,606	0.9473	33	104,688
rs10746942	9	71,434,465	*PIP5K1B*	0.9979	3	13,061	0.9990	33	110,403
rs80282103	10	899,071	*WDR37*	0.9851	1	1,993	0.9330	33	103,112
rs10994856	10	52,645,248	*A1CF*	0.9259	50	123,790	0.9655	33	106,705
rs84178	11	2,774,374	*KCNQ1*	0.9346	3	12,233	0.9797	33	108,276
rs3925584	11	30,760,335	*MPPED2*	0.9901	50	132,354	0.9938	33	109,833
rs11604462	11	65,551,648	*RNASEH2C*	0.9988	3	13,073	0.9904	33	109,453
rs11062167	12	364,739	*SLC6A13*	0.9982	1	2,019	0.9735	33	107,587
rs67551338	12	3,393,100	*TSPAN9*	0.8592	2	3,703	0.9062	33	100,149
rs12826808	12	15,323,380	*RERG*	0.9975	3	13,056	0.9971	33	110,193
rs3741414	12	57,844,049	*R3HDM2*	0.8621	50	115,255	0.9604	32	106,133
rs9529913	13	72,345,089	*DACH1*	0.9936	47	129,642	0.9815	33	108,467
rs6492982	15	41,399,951	*INO80*	0.9298	1	1,881	0.9170	33	101,344
rs2453533	15	45,641,225	*GATM*	0.9997	50	133,682	0.9999	33	110,511
15:53922280 :INDEL	15	53,922,280	*WDR72*	NA	0	0	0.9881	29	86,941
rs10851885	15	76,304,503	*UBE2Q2*	0.9972	47	133,425	1.0000	29	109,424
rs77924615	16	20,392,332	*UMOD*	0.9591	1	1,940	0.8480	33	93,718
rs428232	16	89,713,969	*DPEP1*	0.9557	1	1,933	0.9020	33	99,686
rs894680	17	19,440,538	*SLC47A1*	0.7776	49	103,165	0.8170	33	90,291
rs12451586	17	37,633,835	*CDK12*	0.9413	1	1,904	0.8270	33	91,396
rs9895661	17	59,456,589	*BCAS3*	0.8507	49	113,560	0.9162	33	101,259
rs71359461	18	77,156,103	*NFATC1*	0.8676	1	1,755	0.7875	32	85,365
rs7247977	19	33,358,355	*SLC7A9*	0.9944	3	13,016	0.9802	33	108,331
rs151087334	19	38,205,244	*SIPA1L3*	0.8540	1	1,728	0.5890	33	65,095

rs6058093	20	33,213,196	*TP53INP2*	0.9901	1	2,003	0.9066	33	100,191
rs6127099	20	52,731,402	*BCAS1*	0.8213	4	16,032	0.8749	33	96,686

Position is given on GRCh build 37. The gene closest to the variant is listed. The imputation quality is taken from each file in the meta-analysis. Typical metrics are the info quality (ImputeV2) and the RSQ (minimac).

*) The reported variant in known locus and the top variant in known locus are the same (reported just once)

**) rs7422339 has merged into rs1047891

7.2 Meta-analysis results from the HapMap and the 1000 Genomes meta-analyses

Figure 24 shows the Manhattan Plot of the CKDGen 1000 Genomes meta-analysis p-values, highlighting the kidney function disease loci detected by the CKDGen HapMap meta-analysis (in orange) and highlighting the kidney function disease loci identified by the CKDGen 1000 Genomes meta-analysis additional to the CKDGen HapMap meta-analysis (in dark red).

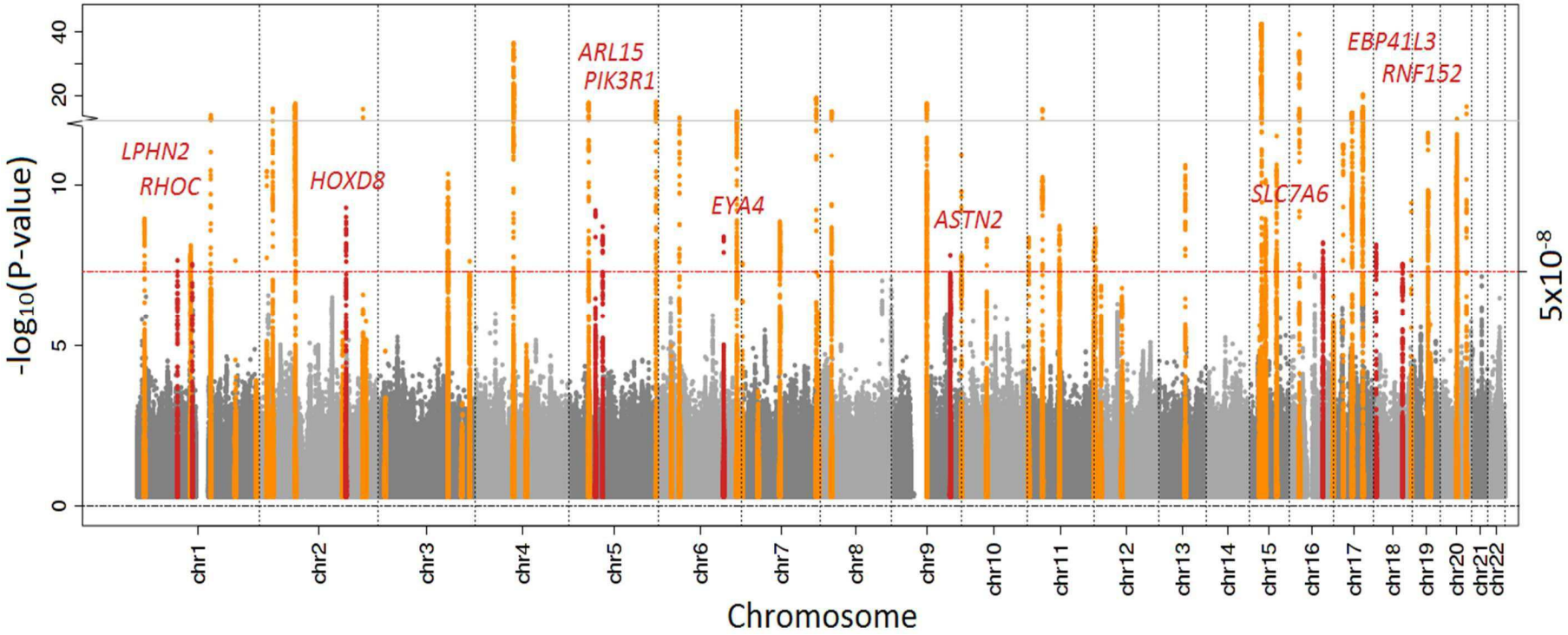

*Figure 24.**Manhattan Plot for eGFRcrea from the 1000 Genomes meta-analysis**. Shown are the p-values (on a -log10 scale) vs. genomic position (on GRCh build 37) in the CKDGen 1000 Genomes meta-analysis results. Genome wide loci uncovered additional to the HapMap meta-analysis are highlighted and labeled in dark red. Loci identified in the HapMap meta-analysis highlighted in bright orange. The human genome wide significance threshold at 5x10-8 is highlighted as horizontal, red dotted line.*

7.2.1 Comparison of the four lead variants identified by the CKDGen 1000 Genomes meta-analysis and not identified by the CKDGen Hapmap meta-analysis

Figure 18 and **Figure 19** illustrate the estimated effects and standard errors of the four lead variants identified by the CKDGen 1000 Genomes meta-analysis and not identified by the the CKDGen Hapmap meta-analysis in the studies analyzed in both HapMap and 1000 Genomes meta-analysis. The effect estimates, standard errors and number of subjects in the meta-analyses were illustrated as forestplots in **Figure 27** and **Figure 28**. Here the studies from the overlap are separated from the studies, which were analyzed in either HapMap or 1000 Genomes exclusively. Additionally, the meta-analysis results of the overlap, the exclusive studies and the overall meta-analysis are illustrated.

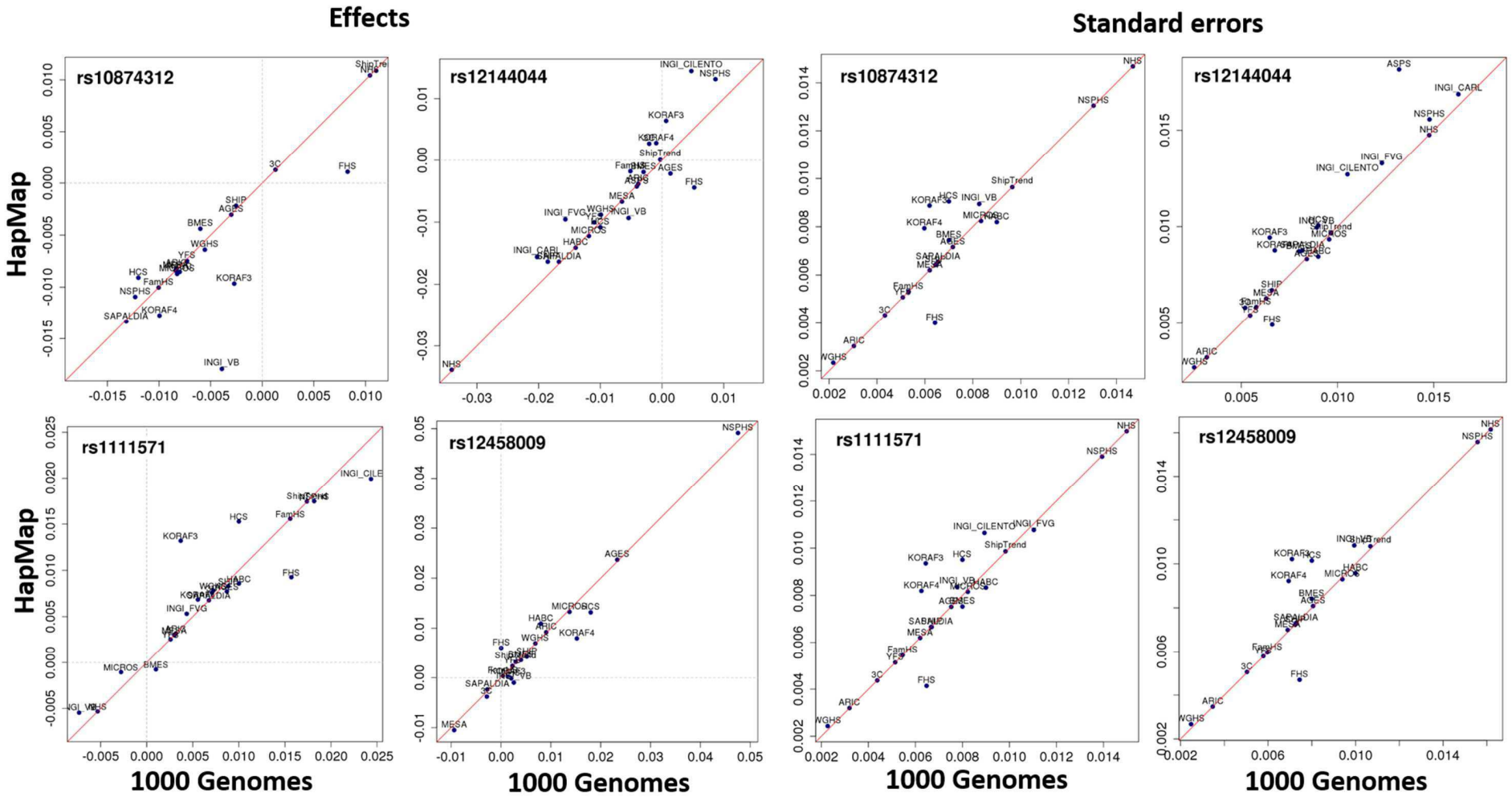

Figure 25. ***Comparison of the effects and standard errors in the HapMap and 1000 Genomes meta-analysis****.Shown are the effects (4 scatterpots on the left side) and the standard errors (4 scatterplot on the right side) in the CKDGen 1000 Genomes meta-analysis (x-axis) vs CKDGen HapMap meta-analysis (y-axis) of the four loci identified with genome-wide significance in the 1000 Genomes meta-analysis, but not in the HapMap meta-analysis.*

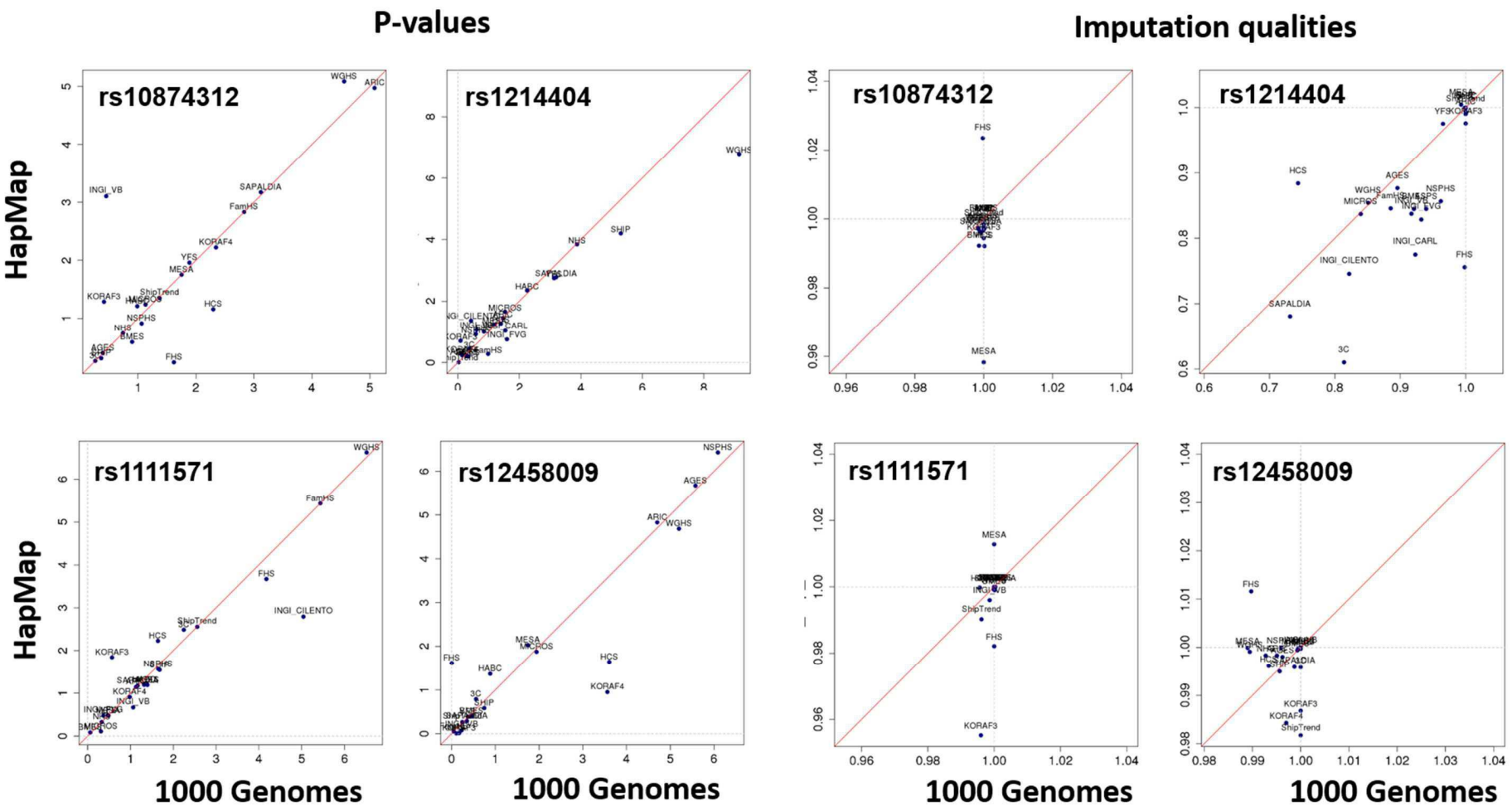

Figure 26. ***Comparison of the p-values and imputation qualities in the HapMap and 1000 Genomes meta-analysis****. Shown are the p-values (4 scatterpots on the left side) and the imputation qualities (four scatterplots on the right side) in the CKDGen 1000 Genomes meta-analysis (x-axis) vs CKDGen HapMap meta-analysis (y-axis) of the four loci identified with genome-wide significance in the 1000 Genomes meta-analysis, but not in the HapMap meta-analysis.*

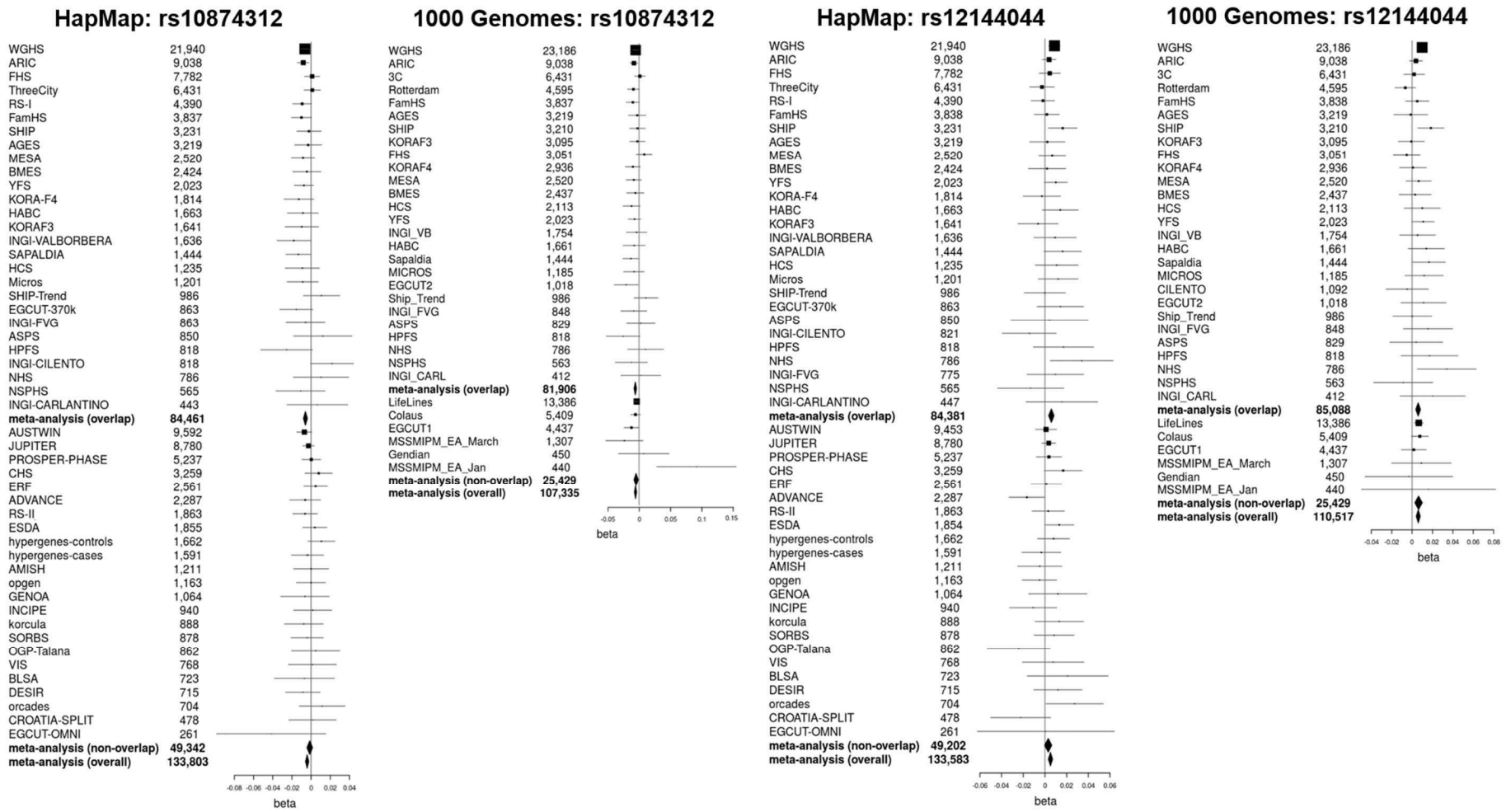

Figure 27. ***Forestplots of rs10874312 and rs12144044 on all studies and meta-analysis results from the CKDGen 1000 Genomes and the CKDGen HapMap meta-analysis****. Shown are the effect estimates (beta) with the 95% confidence intervals by study as forestplots of rs10874312 in the HapMap and 1000 Genomes meta-analysis and the forestplots of rs10874312 in the HapMap and 1000 Genomes meta-analysis (from left to right). The meta-analysis results of the studies in both analyses - meta-analysis (overlap), the meta-analysis results in the studies in either HapMap or 1000 Genomes meta-analysis - meta-analysis (non-overlap) and the meta-analysis of all studies meta-analysis (overall) are represented as diamonds.*

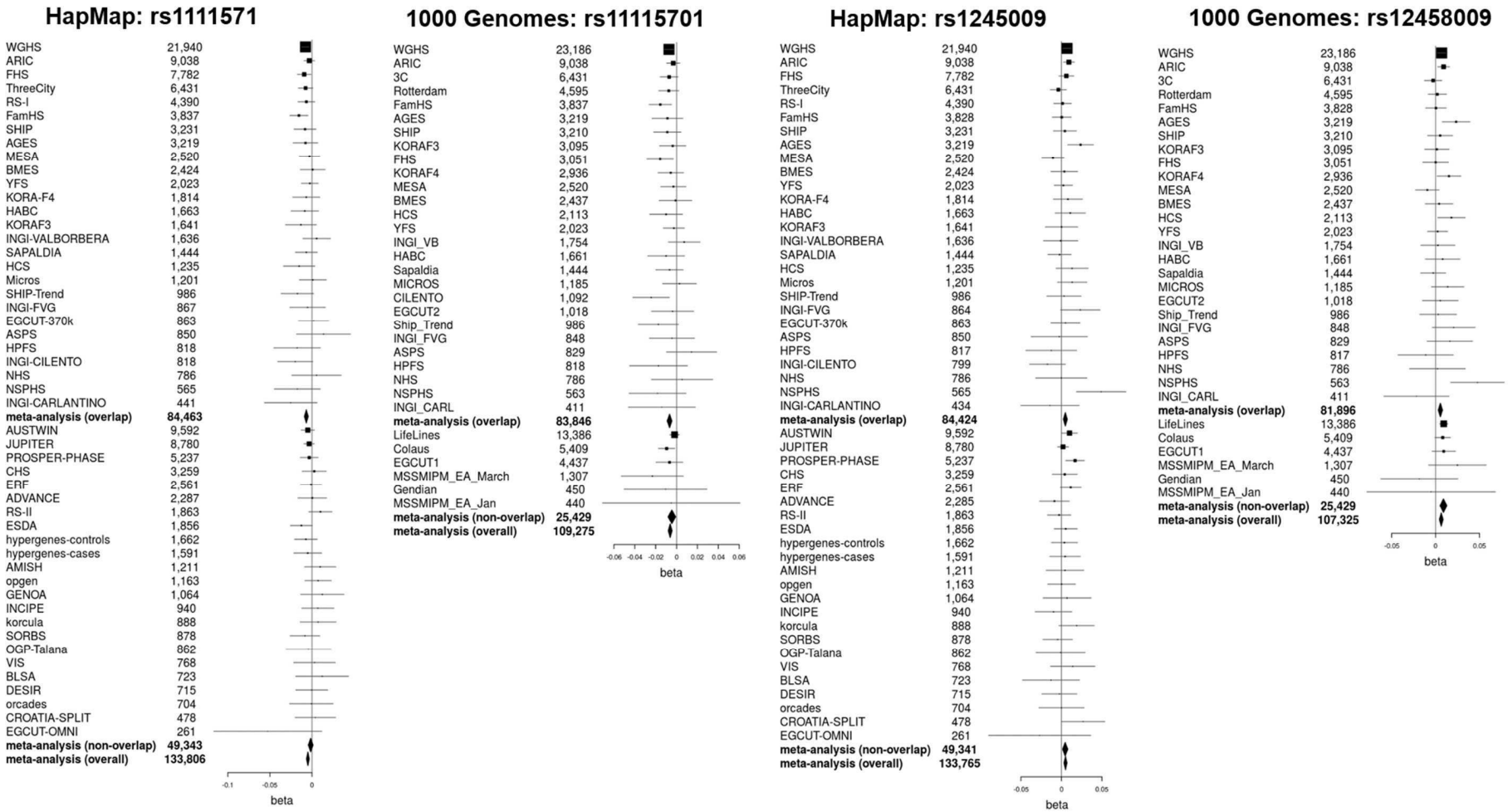

Figure 28. ***Forestplots of rs1111571 and rs12458009 of the studies in the CKDGen HapMap meta-analysis and in the CKDGen 1000 Genomes meta-analysis****. Shown are the effect estimates (beta) with the 95% confidence intervals by study as forestplots of rs1111571 in the HapMap and 1000 Genomes meta-analysis and as forestplots of rs12458009 in the HapMap and 1000 Genomes meta-analysis (from left to right). The meta-analysis results of the studies in both analyses - meta-analysis (overlap), the meta-analysis results in the studies in either HapMap or 1000 Genomes meta-analysis - meta-analysis (non-overlap) and the meta-analysis of all studies – meta-analysis (overall) - are represented as diamonds.*

7.2.2 Comparison of the one lead variant identified by the CKDGen 1000 Genomes meta-analysis and the CKDGen Hapmap meta-analysis

Figure 30 illustrates the estimated effects, standard errors, p-values and imputation qualities of the lead variant in the *DDX1* locus, identified by the CKDGen 1000 Genomes meta-analysis and the CKDGen Hapmap meta-analysis in the studies analyzed in both HapMap and 1000 Genomes meta-analysis. The effect estimates, standard errors and number of subjects in the meta-analyses were illustrated as forestplots in **Figure 29**. Here the studies from the overlap were separated from the studies, which were analyzed in either HapMap or 1000 Genomes exclusively. Additionally, the meta-analysis results of the overlap, the exclusive studies and the overall meta-analysis were illustrated.

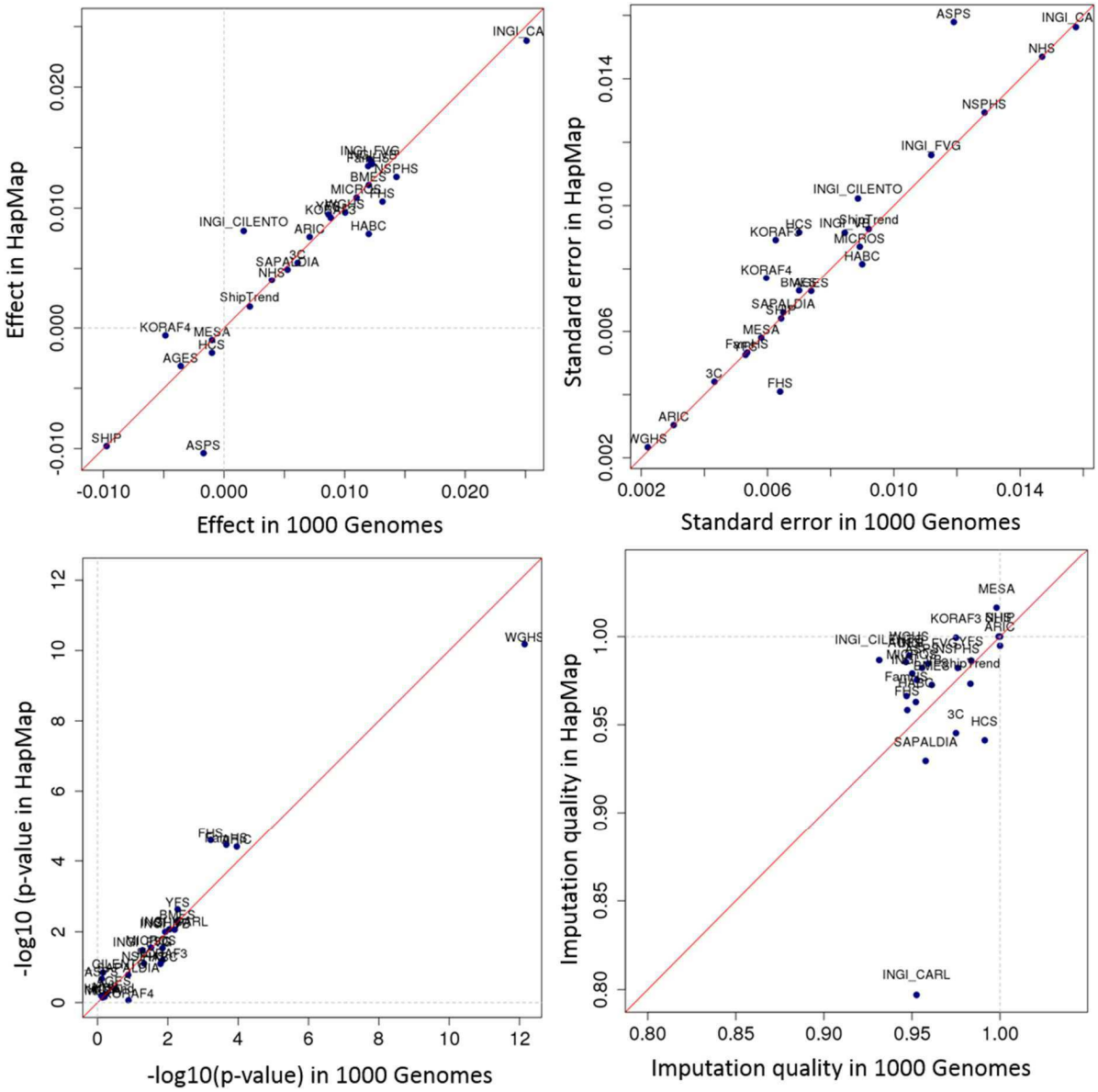

Figure 29. ***Scatterplots of effects, standard errors, p-values and imputation qualities of rs807601 of the studies overlapping the CKDGen 1000 Genomes meta-analysis and the CKDGen HapMap meta-analysis****.The p-values are given on a –log10 scale.*

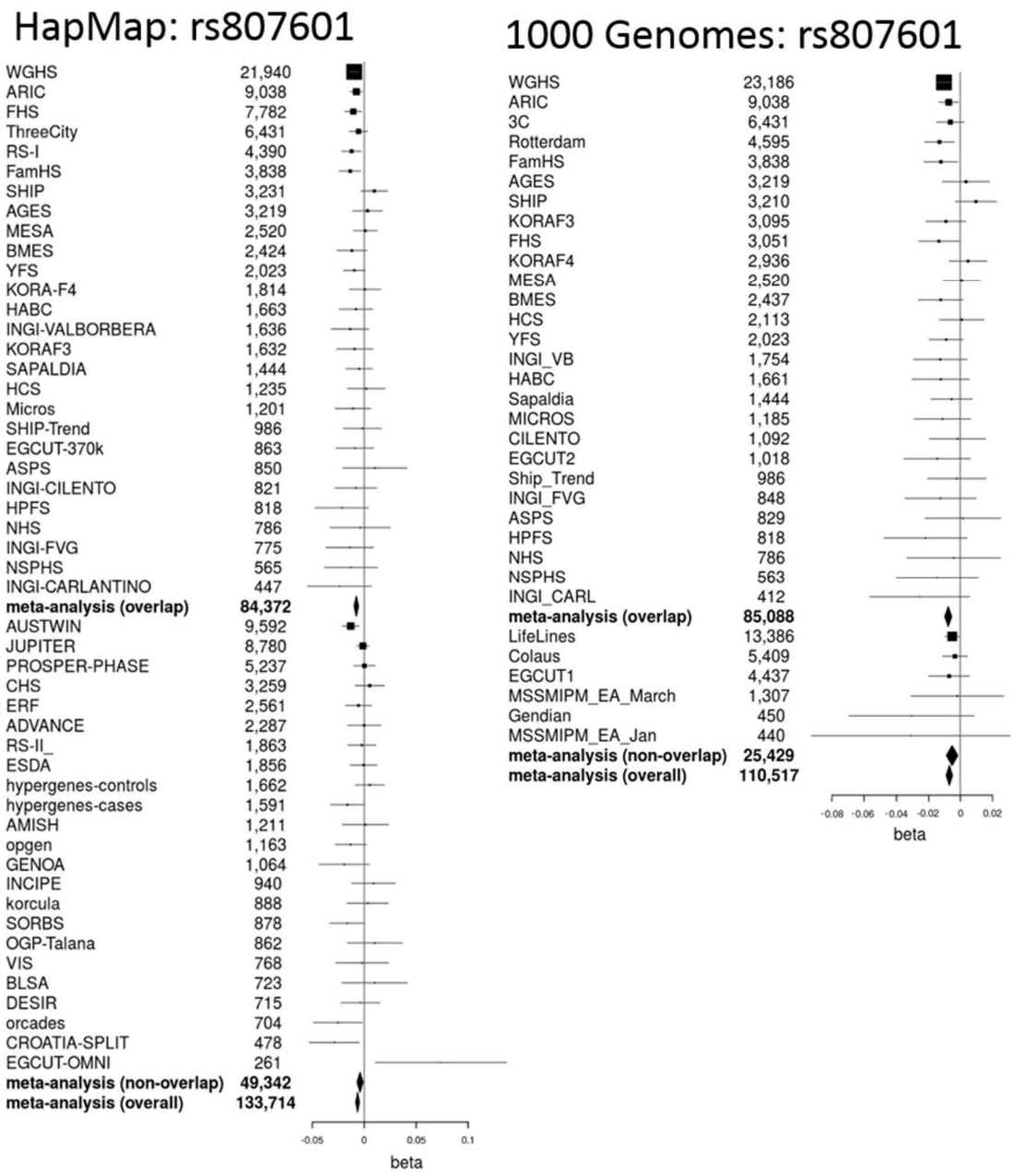

Figure 30. **Forestplots of rs807601 of the studies in the CKDGen HapMap meta-analysis and in the CKDGen 1000 Genomes meta-analysis**. Shown are the effect estimates (beta) with the 95% confidence intervals by study as forestplots of rs807601 in the HapMap and 1000 Genomes meta-analysis (from left to right). The meta-analysis results of the studies in both analyses - meta-analysis (overlap), the meta-analysis results in the studies in either HapMap or 1000 Genomes meta-analysis - meta-analysis (non-overlap) and the meta-analysis of all studies – meta-analysis (overall) - are represented as diamonds.

7.2.3 Comparison of the one lead variant not identified by the CKDGen 1000 Genomes meta-analysis and identified by the CKDGen Hapmap meta-analysis

The lead variant in the *LRP2* locus was not identified by the CKDGen 1000 Genomes meta-analysis and identified by the CKDGen Hapmap meta-analysis. The effect estimates, standard errors and number of studies in the CKDGen HapMap and CKDGen 1000 Genomes meta-analysis were illustrated as forestplots in **Figure 31.** Here the studies from the overlap were separated from the studies, which were analyzed in either HapMap or 1000 Genomes exclusively. Additionally, the meta-analysis results of the overlap, the exclusive studies and the overall meta-analysis were illustrated.

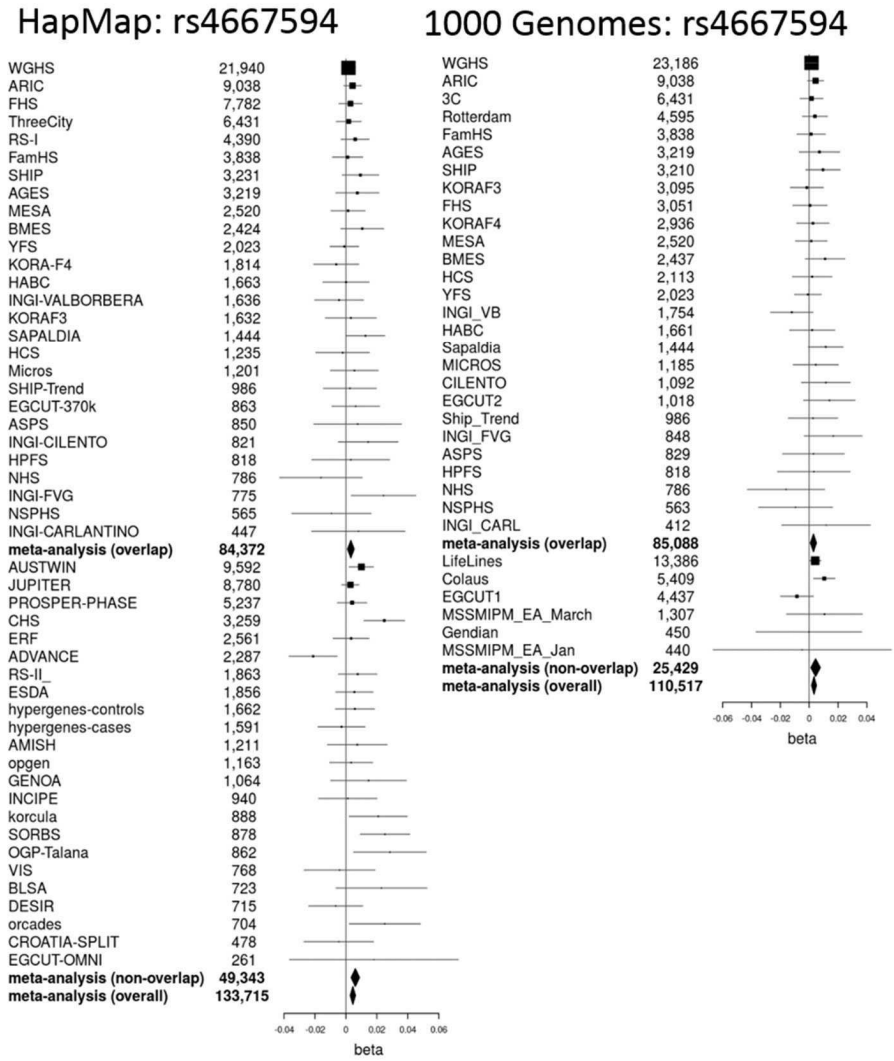

Figure 31. ***Forestplots of rs4667594 of the studies in the CKDGen HapMap meta-analysis and in the CKDGen 1000 Genomes meta-analysis****. Shown are the effect estimates (beta) with the 95% confidence intervals by study as forestplots of rs4667594 in the HapMap and 1000 Genomes meta-analysis (from left to right). The meta-analysis results of the studies in both analyses - meta-analysis (overlap), the meta-analysis results in the studies in either HapMap or 1000 Genomes meta-analysis - meta-analysis (non-overlap) and the meta-analysis of all studies – meta-analysis (overall) - are represented as diamonds.*

7.3 SNPs changing position

Figure 32 shows the change of positions from build 36.3 to build 37.1 on the complete chromosome 1 (a) and at the CFH region (b). Nearly all SNPs changed their positions. Illustrating these points for chromosome 1, shows the scatter by SNPs with relative position change, and highlights the offset of positions in a smaller region.

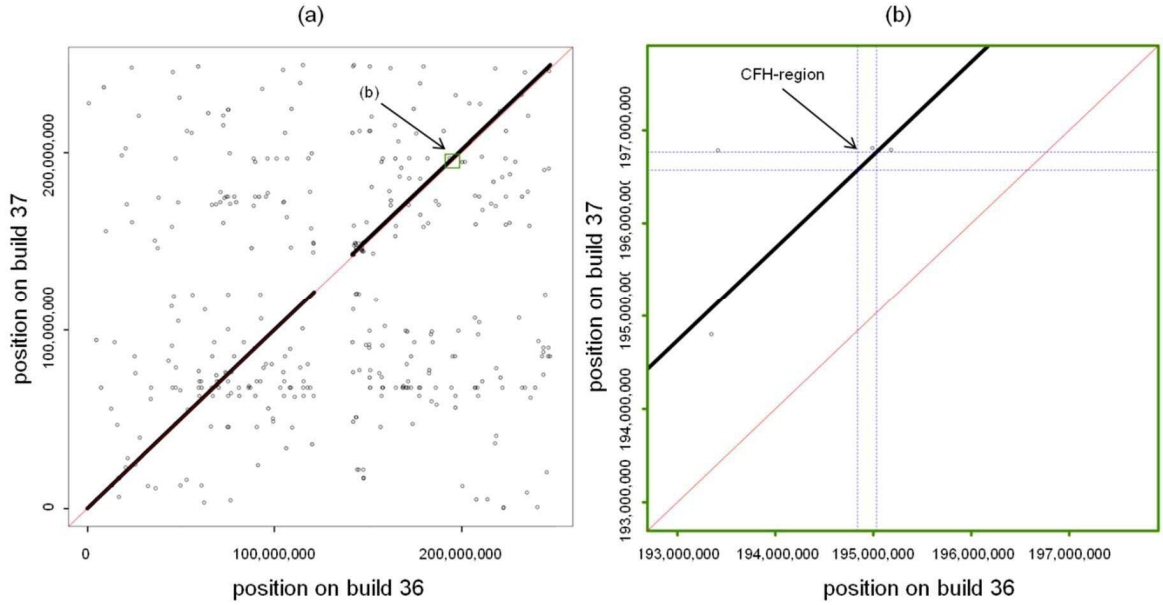

Figure 32. ***SNPs changing position from b36.3 to b37.1 on chromosome 1****. Shown are the positions of b36.3 compared to b37.1 (b130_SNPChrPosOnRef_36_3.bcp and b131_SNPChrPosOnRef_37_1.bcp, see Web Resources, 17,479,168 SNPs contained in both data sets) for (a) the chromosome 1 (1,347,160 SNPs) and (b) the CFH region (chr1:193 Mb–198 Mb, 14,532 SNPs).*

7.4 How similarities in dosages are reflected by similarities in imputation quality

I was interested whether similarity or difference in the imputation quality (RSQ) between the pre-phasing lift-over approach and the post-phasing lift-over approach also reflected similarity or difference in individuals' dosages. The figure below shows three SNPs from the CFH region example.

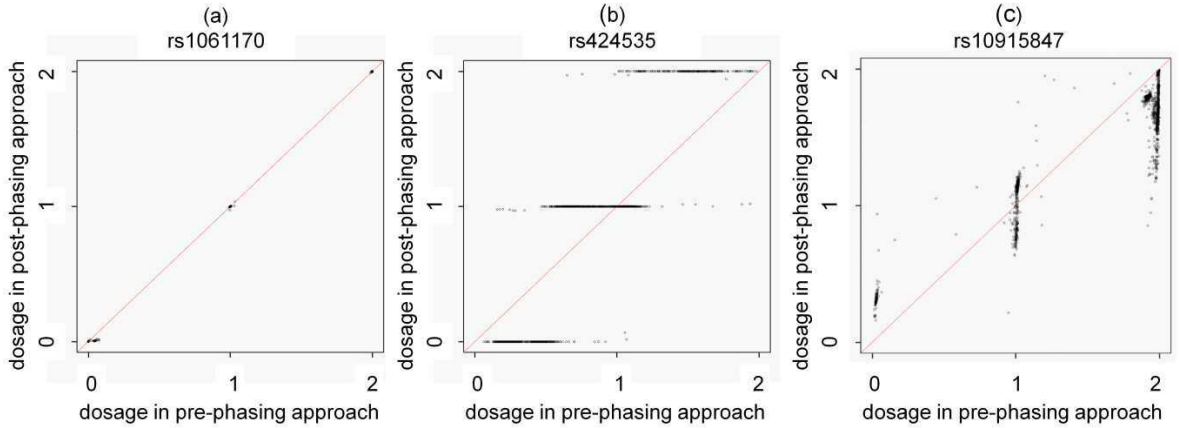

Figure 33. ***Similarities and differences in imputation quality (RSQ) reflect also in similarities and difference of the dosages.*** *Shown are dosages of three example SNPs in the CFH region imputed with pre- and post-phasing approach. (a) rs1061170, similar RSQ for both approaches (RSQpre=0.99, RSQpost=0.99), (b) rs424535, deteriorated RSQ for pre-phasing approach (RSQpre=0.47, RSQpost=0.99), (c) rs10915847, deteriorated RSQ for post-phasing approach (RSQpre=0.92, RSQpost=0.51). It should be noted that the dosages are nearly integer in the case of very high imputation quality near unity as expected.*

7.5 Differences between genomic builds, reference panels and pre-/ post phasing approach

7.5.1 Overview over the changes between builds

During the last five years a list of different builds have been released. **Table 28** shows a comprehensive overview over the different builds, released in the past five years:

*Table 28. **Overview on different builds released during the last five years**. Given is the Genome build, the number of SNPs and the date of release (dbSNP, see Web Resources).*

Genome Build	Number of SNPs (rs#'s)	Date available
37.5	62,676,337	Apr 25, 2013
37.4	53,567,890	Jun 26, 2012
37.3	52,327,221	Oct 13, 2011
37.2	41,365,915	Aug 15, 2011
37.1	30,442,771	Sep 26, 2010
37.1	23,653,737	Mar 25, 2010
36.3	17,804,034	May 03, 2009
36.2	14,380,528	Oct 25, 2007
36.2	11,811,594	Mar 08, 2007
36.1	11,961,761	May 18, 2006
35.1	10,430,753	Sep 29, 2005

7.5.2 Overview of reference panels

The constantly emerging reference panels from the 1000 Genomes project are shown in **Table 29**. The number of variations, subjects and the composition of populations have changed over time:

*Table 29. **Overview of reference panels for imputation with MaCH and minimac.** Stated are reference panels based on 1000G March 2010 and November 2010 data freeze. The 1000G 2010-03 and 1000G 2010-06 reference panels are on build 36.3, whereas the others are on build 37.1 or newer. The latest version (GIANT ALL 1000G Phase I v3) is currently recommended for imputation. It is a subset of 1000G Phase I v3 excluding singletons and monomorphic variants.*

Date of		Number of haplotypes in				Popu-	Number of	
data freeze	Name	CEU	ASN	AFR	AMR	lation	SNPs	X chromosome
Mar-10	1000G 2010-03	120	0	0	0	60	7,850,000	not included
Mar-10	1000G 2010-06	120	124	124	0	184	6,900,000	not included
Mar-10	1000G 2010-08	566	194	174	0	467	11,900,000	not included
Nov-10	1000G Phase I (α)	762	572	492	362	1094	37,400,000	included
Nov-10	1000G Phasel v2	758	572	492	362	1092	40,309,712	included
Nov-10	1000G Phasel v3	758	572	492	362	1092	~39,700,000	included
Nov-10	GIANT ALL 1000G Phasel v3	758	572	492	362	1092	~30,000,000	included

7.5.3 Position change between b36.3 and b37.1

Comparing the 17,479,168 SNPs contained in both data sets of b37.1 to b36.3 I notice 324,866 fewer SNPs and 238,400 SNPs re-located to other chromosomes; nearly all SNPs changed their positions:

Table 30. ***Quantifying the position change between b36.3 and b37.1.*** *Stated are the number of SNPs changing positions within a chromosome between NCBI build 36.3 and GRCh build 37.1 based on the data files 130_SNPChrPosOnRef_36_3.bcp and b131_SNPChrPosOnRef_37_1.bcp.*

chromosome number	number of SNPs increasing position		number of SNPS decreasing position		number of SNPs with equal position	
	number	%	number	%	number	%
1	1,322,443	99.85 %	2,047	0.15 %	0	0.00 %
2	1,320,125	99.68 %	4,229	0.32 %	0	0.00 %
3	398,103	36.16 %	702,905	63.84 %	0	0.00 %
4	366,593	33.50 %	727,723	66.50 %	0	0.00 %
5	237,164	24.56 %	728,565	75.44 %	0	0.00 %
6	155,769	14.04 %	953,640	85.96 %	0	0.00 %
7	895,544	98.92 %	9,770	1.08 %	0	0.00 %
8	88,136	10.60 %	743,143	89.40 %	0	0.00 %
9	665,401	99.19 %	5,462	0.81 %	0	0.00 %
10	575,618	72.72 %	215,933	27.28 %	0	0.00 %
11	770,364	99.61 %	3,011	0.39 %	0	0.00 %
12	756,388	98.71 %	9,853	1.29 %	0	0.00 %
13	585,891	98.73 %	7,524	1.27 %	0	0.00 %
14	497,351	98.94 %	5,310	1.06 %	0	0.00 %
15	454,527	96.50 %	16,508	3.50 %	0	0.00 %
16	507,964	99.64 %	1,819	0.36 %	0	0.00 %
17	459,234	99.14 %	1,462	0.32 %	2,536	0.55 %
18	450,987	99.92 %	345	0.08 %	0	0.00 %
19	159,136	43.30 %	208,408	56.70 %	0	0.00 %
20	433,011	99.72 %	1,205	0.28 %	0	0.00 %
21	240,074	98.05 %	4,779	1.95 %	0	0.00 %
22	259,259	99.99 %	15	0.01 %	0	0.00 %
X	467,383	88.29 %	61,989	11.71 %	0	0.00 %
Y	12,447	45.43 %	14,953	54.57 %	0	0.00 %
all	12,078,912	73.15 %	4,430,598	26.83 %	2,536	0.02 %

7.5.4 Differences in imputation quality between pre- and post-phasing approach on chromosome 1

The differences in imputation quality (RSQ) between the post-phasing approach and the pre-phasing approach of imputation by MAF categories is given in **Table 31**. The differences are symmetric across all 16 parts. Highlighted in bold is part12, which contains the CFH region.

*Table 31. **Differences in RSQ across chromosome 1.** Shown are the means and SD of the difference between RSQ_{pre} and RSQ_{post} (n=1,644, #SNPs=2,300,365) on chromosome1 in 16 regions (size: 12.5 Megabases) by MAF in pre-phasing approach. The region containing the CFH locus (chr1:196,594,399 to 196,718,099) is highlighted in bold letters.*

Region	Difference in imputation quality RSQ_{post} - RSQ_{pre} overall	MAF 0%-5%	MAF 5%-20%	MAF 20%-50%
1	0.005 (0.062)	0.004 (0.060)	0.007 (0.053)	0.009 (0.081)
2	0.003 (0.071)	0.002 (0.075)	0.003 (0.043)	0.009 (0.057)
3	0.004 (0.077)	0.003 (0.081)	0.010 (0.069)	0.007 (0.059)
4	0.007 (0.080)	0.006 (0.084)	0.009 (0.044)	0.014 (0.071)
5	0.015 (0.093)	0.014 (0.096)	0.017 (0.072)	0.021 (0.090)
6	0.005 (0.075)	0.005 (0.082)	0.006 (0.042)	0.006 (0.050)
7	0.005 (0.077)	0.005 (0.085)	0.005 (0.031)	0.005 (0.038)
8	0.006 (0.077)	0.005 (0.083)	0.008 (0.041)	0.008 (0.051)
9	0.011 (0.083)	0.010 (0.086)	0.012 (0.056)	0.015 (0.078)
10	0.003 (0.078)	0.002 (0.087)	0.006 (0.046)	0.005 (0.040)
11	0.006 (0.098)	0.006 (0.108)	0.005 (0.047)	0.010 (0.068)
12	**0.005 (0.074)**	**0.004 (0.080)**	**0.010 (0.054)**	**0.007 (0.045)**
13	0.008 (0.083)	0.007 (0.089)	0.008 (0.040)	0.014 (0.068)
14	0.006 (0.089)	0.005 (0.096)	0.005 (0.053)	0.008 (0.068)
15	0.006 (0.077)	0.005 (0.083)	0.008 (0.045)	0.009 (0.058)
16	0.008 (0.069)	0.007 (0.067)	0.012 (0.062)	0.011 (0.084)

7.6 Differences when repeating imputation and analysis with identical parameters

7.6.1 Differences between four mega-imputation and mega-analyses with identical parameters

The mega-imputation and the mega-analysis were repeated four times. Small differences in the results of phase estimation and genotype imputation are reflected in small differences between the Firth corrected logistic regression analyses of the four mega-imputed and mega-analyzed variants. Overall yielded all 34 variants in all four runs genome wide significance with two exceptions: The lead variant in the *CNN2* locus, which yielded p-values close to the human genome wide significance threshold, were in the third and fourth analysis not genome wide significant, reflecting the small differences between the four runs of mega-imputation and mega-analysis. But overall the differences in p-values are small. The odds ratios were very consistent across the four analyses. Whereas they were identical in most loci, the biggest differences were found in the *ACAD10* gene, with a negligible difference of 0.03 of the odds ratios between the second and third run.

In summary the four runs of the mega-analysis can be seen as equivalent and for the analyses and comparisons in my thesis the first results were used.

7.6.2 Differences between four meta-imputation and meta-analyses with identical parameters

The meta-imputation and the meta-analysis were both repeated twice. P-values were very similar, whereas variants were partially genome wide significant (see **chapter 3.2.5)**. The odds ratios were nearly identical, whereas the differences were not greater than 0.01. The differences of the observed heterogeneity in the meta-analysis were smaller but overall equivalent across all 34 loci.

In summary the four runs of the meta-imputation and meta-analysis were equivalent and for the analyses and comparisons in my thesis the results from the first round of phasing and the first round of imputation were used.

Table 32. Firth corrected logistic regression results of the mega-imputed and mega-analyzed lead SNPs in the 34 loci associated with AMD in the mega-analysis

				OR				P-value			
Variant ID	Chr	Position (bp)	Closest Gene	1	2	3	4	1	2	3	4
rs10922109	1	196,704,632	*CFH*	0.38	0.38	0.38	0.38	$1.7x10^{-615}$	$1.2x10^{-618}$	$1.2x10^{-615}$	$1.2x10^{-618}$
rs11884770	2	228,086,920	*COL4A3*	0.90	0.90	0.90	0.90	$2.5x10^{-8}$	$2.3x10^{-8}$	$2.1x10^{-8}$	$2.5x10^{-8}$
rs62247658	3	64,715,155	*ADAMTS9-AS2*	1.14	1.14	1.14	1.14	$1.3x10^{-14}$	$1.3x10^{-14}$	$1.4x10^{-14}$	$1.5x10^{-14}$
rs140647181	3	99,180,668	*COL8A1*	1.60	1.61	1.59	1.59	$5.9x10^{-12}$	$5.9x10^{-12}$	$7.1x10^{-12}$	$8.1x10^{-12}$
rs10033900	4	110,659,067	*CFI*	1.15	1.15	1.15	1.15	$5.7x10^{-17}$	$5.6x10^{-17}$	$5.7x10^{-17}$	$6.2x10^{-17}$
rs114092250	5	35,494,448	*PRLR/SPEF2*	0.70	0.70	0.70	0.70	$2.9x10^{-8}$	$2.6x10^{-8}$	$3.0x10^{-8}$	$2.9x10^{-8}$
rs62358361	5	39,327,888	*C9*	1.79	1.79	1.79	1.79	$1.6x10^{-14}$	$1.6x10^{-14}$	$2.2x10^{-14}$	$2.0x10^{-14}$
rs116503776	6	31,930,462	*C2/CFB/SKIV2L*	0.57	0.57	0.57	0.57	$8.2x10^{-104}$	$8.0x10^{-104}$	$1.7x10^{-103}$	$1.7x10^{-103}$
rs943080	6	43,826,627	*VEGFA*	0.88	0.88	0.88	0.88	$1.1x10^{-14}$	$1.1x10^{-14}$	$1.1x10^{-14}$	1.1x10-14
rs7803454	7	99,991,548	*PILRB/PILRA*	1.13	1.13	1.13	1.13	$5.0x10^{-9}$	$5.2x10^{-9}$	$4.7x10^{-9}$	$4.8x10^{-9}$
rs1142	7	104,756,326	*KMT2E/SRPK2*	1.11	1.11	1.11	1.11	$1.3x10^{-9}$	$1.4x10^{-9}$	$1.5x10^{-9}$	$1.5x10^{-9}$
rs79037040	8	23,082,971	*TNFRSF10A*	0.90	0.90	0.90	0.90	$2.5x10^{-11}$	$2.6x10^{-11}$	$2.4x10^{-11}$	$2.5x10^{-11}$
rs71507014	9	73,438,605	*TRPM3*	1.10	1.10	1.10	1.10	$3.2x10^{-8}$	$2.9x10^{-8}$	$2.7x10^{-8}$	$2.7x10^{-8}$
rs10781182	9	76,617,720	*MIR6130/RORB*	1.11	1.11	1.11	1.11	$2.5x10^{-9}$	$2.5x10^{-9}$	$2.7x10^{-9}$	$2.7x10^{-9}$
rs1626340	9	101,923,372	*TGFBR1*	0.88	0.88	0.88	0.88	$3.9x10^{-10}$	$3.8x10^{-10}$	$4.1x10^{-10}$	$3.8x10^{-10}$
rs2740488	9	107,661,742	*ABCA1*	0.90	0.90	0.90	0.90	$1.7x10^{-8}$	$1.9x10^{-8}$	$1.4x10^{-8}$	$1.7x10^{-8}$
rs12357257	10	24,999,593	*ARHGAP21*	1.11	1.11	1.11	1.11	$3.9x10^{-8}$	$3.9x10^{-8}$	$3.7x10^{-8}$	$3.8x10^{-8}$
rs3750846	10	124,215,565	*ARMS2/HTRA1*	2.81	2.81	2.81	2.81	$1.5x10^{-735}$	$1.5x10^{-735}$	$1.5x10^{-735}$	$1.5x10^{-735}$
rs3138141	12	56,115,778	*RDH5/CD63*	1.17	1.17	1.17	1.17	$1.6x10^{-9}$	$1.4x10^{-9}$	$2.5x10^{-9}$	$2.3x10^{-9}$
rs61941274	12	112,132,610	*ACAD10*	1.50	1.52	1.49	1.49	$1.7x10^{-9}$	$1.4x10^{-9}$	$1.7x10^{-9}$	$1.7x10^{-9}$
rs9564692	13	31,821,240	*B3GALTL*	0.89	0.89	0.89	0.89	$3.1x10^{-10}$	$3.1x10^{-10}$	$3.2x10^{-10}$	$3.3x10^{-10}$
rs61985136	14	68,769,199	*RAD51B*	0.90	0.90	0.90	0.90	$1.5x10^{-10}$	$1.6x10^{-10}$	$1.6x10^{-10}$	$1.4x10^{-10}$
rs2043085	15	58,680,954	*LIPC*	0.88	0.88	0.88	0.88	$6.7x10^{-14}$	$6.9x10^{-14}$	$6.7x10^{-14}$	$6.7x10^{-14}$
rs5817082	16	56,997,349	*CETP*	0.84	0.84	0.84	0.84	$5.1x10^{-19}$	$5.4x10^{-19}$	$4.3x10^{-19}$	$4.3x10^{-19}$
rs72802342	16	75,234,872	*CTRB2/CTRB1*	0.80	0.79	0.79	0.79	$9.7x10^{-12}$	$8.7x10^{-12}$	$7.1x10^{-12}$	$6.3x10^{-12}$
rs11080055	17	26,649,724	*TMEM97/VTN*	0.91	0.91	0.91	0.91	$1.1x10^{-8}$	$9.5x10^{-9}$	$1.2x10^{-8}$	$1.1x10^{-8}$
rs6565597	17	79,526,821	*NPLOC4/TSPAN10*	1.13	1.13	1.13	1.12	$2.4x10^{-12}$	$2.4x10^{-12}$	$3.7x10^{-11}$	$5.3x10^{-11}$
rs67538026	19	1,031,438	*CNN2*	0.90	0.90	0.91	0.91	$4.4x10^{-8}$	$4.9x10^{-8}$	$1.9x10^{-7}$	$1.8x10^{-7}$
rs2230199	19	6,718,387	*C3*	1.43	1.43	1.43	1.43	$1.6x10^{-69}$	$3.8x10^{-68}$	$6.9x10^{-69}$	$1.4x10^{-68}$
rs429358	19	45,411,941	*APOE*	0.70	0.70	0.70	0.70	$2.2x10^{-42}$	$3.4x10^{-42}$	$4.5x10^{-42}$	$6.6x10^{-42}$
rs142450006	20	44,614,991	*MMP9*	0.85	0.85	0.85	0.85	$3.1x10^{-10}$	$3.3x10^{-10}$	$3.1x10^{-10}$	$2.8x10^{-10}$
rs201459901	20	56,653,724	*C20orf85*	0.76	0.76	0.76	0.76	$3.3x10^{-16}$	$2.6x10^{-16}$	$3.5x10^{-16}$	$3.8x10^{-16}$
rs5754227	22	33,105,817	*SYN3/TIMP3*	0.77	0.78	0.77	0.77	$9.8x10^{-25}$	$3.5x10^{-24}$	$1.6x10^{-24}$	$7.1x10^{-25}$
rs8135665	22	38,476,276	*SLC16A8*	1.14	1.14	1.14	1.14	$9.3x10^{-11}$	$9.2x10^{-11}$	$6.1x10^{-11}$	$6.0x10^{-11}$

Chr = Chromosome; Position (bp) is the chromosomal position given based on NCBI build 37, Closest gene is gene(s) nearest to the variant, OR = odds ratio, the p-values are given in scientific notation.

Table 33. ***Meta-analysis results of the meta-imputed and meta-analyzed lead SNPs in the 34 loci associated with AMD in the mega-analysis.***

				OR				P-value				Heterogenity			
Variant ID	Chr	Position (bp)	Closest Gene	1	2	3	4	1	2	3	4	1	2	3	4
rs10922109	1	196,704,632	*CFH*	0.40	0.40	0.40	0.40	0	0	0	0	58.7	57.7	58.5	58.5
rs11884770	2	228,086,920	*COL4A3*	0.90	0.90	0.90	0.90	$1.1x10^{-6}$	$9.7x10^{-7}$	$1.0x10^{-6}$	$1.1x10^{-6}$	13.3	12.8	14.4	13.9
rs62247658	3	64,715,155	*ADAMTS9-AS2*	1.12	1.12	1.12	1.12	$3.4x10^{-10}$	$3.5x10^{-10}$	$3.3x10^{-10}$	$3.6x10^{-10}$	47.1	47.1	47.2	47.3
rs140647181	3	99,180,668	*COL8A1*	1.49	1.49	1.48	1.48	$2.9x10^{-7}$	$2.9x10^{-7}$	$5.8x10^{-7}$	$5.4x10^{-7}$	0	0	0	0
rs10033900	4	110,659,067	*CFI*	1.13	1.13	1.13	1.13	$6.8x10^{-12}$	$6.0x10^{-12}$	$5.7x10^{-12}$	$5.7x10^{-12}$	0	0	0	0
rs114092250	5	35,494,448	*PRLR/SPEF2*	0.75	0.75	0.76	0.76	$1.1x10^{-04}$	$1.0x10^{-4}$	$1.6x10^{-04}$	$1.4x10^{-4}$	0	0	0	0
rs62358361	5	39,327,888	*C9*	1.64	1.64	1.65	1.65	$1.8x10^{-8}$	$1.8x10^{-8}$	$1.3x10^{-8}$	$1.3x10^{-8}$	0	0	0	0
rs116503776	6	31,930,462	*C2/CFB/SKIV2L*	0.57	0.57	0.57	0.57	$3.1x10^{-85}$	$3.1x10^{-85}$	$3.0x10^{-85}$	$3.1x10^{-85}$	46.2	46.1	45.7	45.7
rs943080	6	43,826,627	*VEGFA*	0.88	0.88	0.88	0.88	$5.3x10^{-12}$	$5.3x10^{-12}$	$5.3x10^{-12}$	$5.3x10^{-12}$	0	0	0	0
rs7803454	7	99,991,548	*PILRB/PILRA*	1.14	1.14	1.14	1.14	$1.9x10^{-8}$	$2.0x10^{-8}$	$2.3x10^{-8}$	$2.5x10^{-8}$	11	11	11.6	11.6
rs1142	7	104,756,326	*KMT2E/SRPK2*	1.10	1.10	1.10	1.10	$1.5x10^{-6}$	$1.6x10^{-6}$	$1.5x10^{-6}$	$1.5x10^{-6}$	19.3	19.5	20.4	20.6
rs79037040	8	23,082,971	*TNFRSF10A*	0.90	0.90	0.90	0.90	$1.7x10^{-9}$	$1.7x10^{-9}$	$2.1x10^{-9}$	$2.4x10^{-9}$	0	0	0	0
rs71507014	9	73,438,605	*TRPM3*	1.09	1.09	1.09	1.09	$7.3x10^{-6}$	$6.0x10^{-6}$	$6.0x10^{-6}$	$4.7x10^{-6}$	1.9	0.9	0	0.7
rs10781182	9	76,617,720	*MIR6130/RORB*	1.09	1.09	1.09	1.10	$5.2x10^{-6}$	$5.1x10^{-6}$	$4.6x10^{-6}$	$4.3x10^{-6}$	0	0	0	0
rs1626340	9	101,923,372	*TGFBR1*	0.87	0.87	0.87	0.87	$9.6x10^{-10}$	$9.9x10^{-10}$	$1.1x10^{-9}$	$1.1x10^{-9}$	0	0	0	0
rs2740488	9	107,661,742	*ABCA1*	0.88	0.88	0.88	0.88	$5.5x10^{-9}$	$4.4x10^{-9}$	$5.9x10^{-9}$	$6.2x10^{-9}$	51.3	51.1	50.4	50.9
rs12357257	10	24,999,593	*ARHGAP21*	1.11	1.11	1.11	1.11	$3.2x10^{-6}$	$3.5x10^{-6}$	$4.1x10^{-6}$	$4.2x10^{-6}$	10.6	9.9	12.1	11.5
rs3750846	10	124,215,565	*ARMS2/HTRA1*	2.71	2.71	2.71	2.71	0	0	0	0	63.7	63.7	64	63.8
rs3138141	12	56,115,778	*RDH5/CD63*	1.16	1.16	1.16	1.16	$3.1x10^{-7}$	$2.2x10^{-7}$	$2.5x10^{-7}$	$2.1x10^{-7}$	0	0	0	0
rs61941274	12	112,132,610	*ACAD10*	1.44	1.44	1.43	1.43	$2.1x10^{-6}$	$1.8x10^{-6}$	$3.4x10^{-6}$	$2.6x10^{-6}$	28.1	28	24	23.1
rs9564692	13	31,821,240	*B3GALTL*	0.90	0.90	0.90	0.90	$1.1x10^{-7}$	$1.1x10^{-7}$	$1.1x10^{-7}$	$1.1x10^{-7}$	23.1	23	23	22.8
rs61985136	14	68,769,199	*RAD51B*	0.91	0.91	0.91	0.91	$6.4x10^{-7}$	$6.5x10^{-7}$	$6.2x10^{-7}$	$6.2x10^{-7}$	21.9	22.1	22.1	23
rs2043085	15	58,680,954	*LIPC*	0.88	0.88	0.88	0.88	$1.2x10^{-10}$	$1.2x10^{-10}$	$1.2x10^{-10}$	$1.2x10^{-10}$	0	0	0	0
rs5817082	16	56,997,349	*CETP*	0.84	0.84	0.84	0.84	$3.2x10^{-16}$	$5.4x10^{-16}$	$3.0x10^{-16}$	$3.5x10^{-16}$	0	0	0	0
rs72802342	16	75,234,872	*CTRB2/CTRB1*	0.83	0.83	0.83	0.83	$4.5x10^{-7}$	$4.3x10^{-7}$	$3.9x10^{-7}$	$4.3x10^{-7}$	0	0	0	0
rs11080055	17	26,649,724	*TMEM97/VTN*	0.91	0.91	0.91	0.91	$2.0x10^{-7}$	$1.7x10^{-7}$	$1.6x10^{-7}$	$1.7x10^{-7}$	0	0	0	0
rs6565597	17	79,526,821	*NPLOC4/TSPAN10*	1.12	1.12	1.12	1.12	$1.7x10^{-8}$	$2.9x10^{-8}$	$9.2x10^{-8}$	$4.9x10^{-8}$	8.8	6.7	13.2	11.2
rs67538026	19	1,031,438	*CNN2*	0.91	0.91	0.91	0.91	$1.3x10^{-5}$	$1.3x10^{-5}$	$9.9x10^{-6}$	$9.2x10^{-6}$	0	0	0	0
rs2230199	19	6,718,387	*C3*	1.37	1.37	1.38	1.38	$1.2x10^{-44}$	$9.6x10^{-45}$	$9.5x10^{-46}$	$2.2x10^{-45}$	0	0	0	0
rs429358	19	45,411,941	*APOE*	0.72	0.72	0.72	0.72	$2.4x10^{-28}$	$3.1x10^{-28}$	$1.9x10^{-28}$	$2.2x10^{-28}$	22.5	23.1	25.5	24
rs142450006	20	44,614,991	*MMP9*	0.84	0.84	0.84	0.84	$1.5x10^{-9}$	$1.2x10^{-9}$	$1.4x10^{-9}$	$1.4x10^{-9}$	0	0	0	0
rs201459901	20	56,653,724	*C20orf85*	0.76	0.76	0.76	0.76	$8.5x10^{-13}$	$9.2x10^{-13}$	$1.0x10^{-12}$	9.0x10+	0	0	0	0
rs5754227	22	33,105,817	*SYN3/TIMP3*	0.77	0.77	0.77	0.77	$3.1x10^{-20}$	$3.1x10^{-20}$	$4.3x10^{-20}$	$2.7x10^{-20}$	0	0	0	0
rs8135665	22	38,476,276	*SLC16A8*	1.14	1.14	1.14	1.14	$6.5x10^{-9}$	$6.5x10^{-9}$	$4.5x10^{-9}$	$4.5x10^{-9}$	30.2	30.2	28.9	28.9

Chr = Chromosome; Position (bp) is the chromosomal position given based on NCBI build 37, Closest gene is gene(s) nearest to the variant, OR = odds ratio, the p-values are given in scientific notation, it is zero (in genes *CFH* and *ARMS2/HTRA1*) if the model did not work with the meta-analysis software. Heterogeneity is the between-study-heterogeinity.

7.7 Power to identify the lead variants in the 34 loci associated with AMD

7.7.1 Power to identify variants with genome-wide significance

The power to detect variants with varying odds ratios is shown in **Figure 34**. It shows that a variant with an odds ratio of 1.11 and MAF ≥ 0.3 yield at least 80% chance to be identified with genome wide significance (dark blue curve). Less frequent variants with a MAF of 10% need an odds ratio of at least 1.17 for 80% power (orange curve). This can be observed in gene *KMT2E/SRPK2*, which yields ~86% power with a MAF = 0.35. A complete overview of all lead variants with their odds ratios and MAF in controls in the 34 susceptibility loci for AMD can be found in **Power to** detect the lead variants in the 34 AMD disease loci

Shown are the odds ratios, MAF in controls and the power to detect the lead variants from the 34 AMD disease loci in **Table 34.**

Table 34.

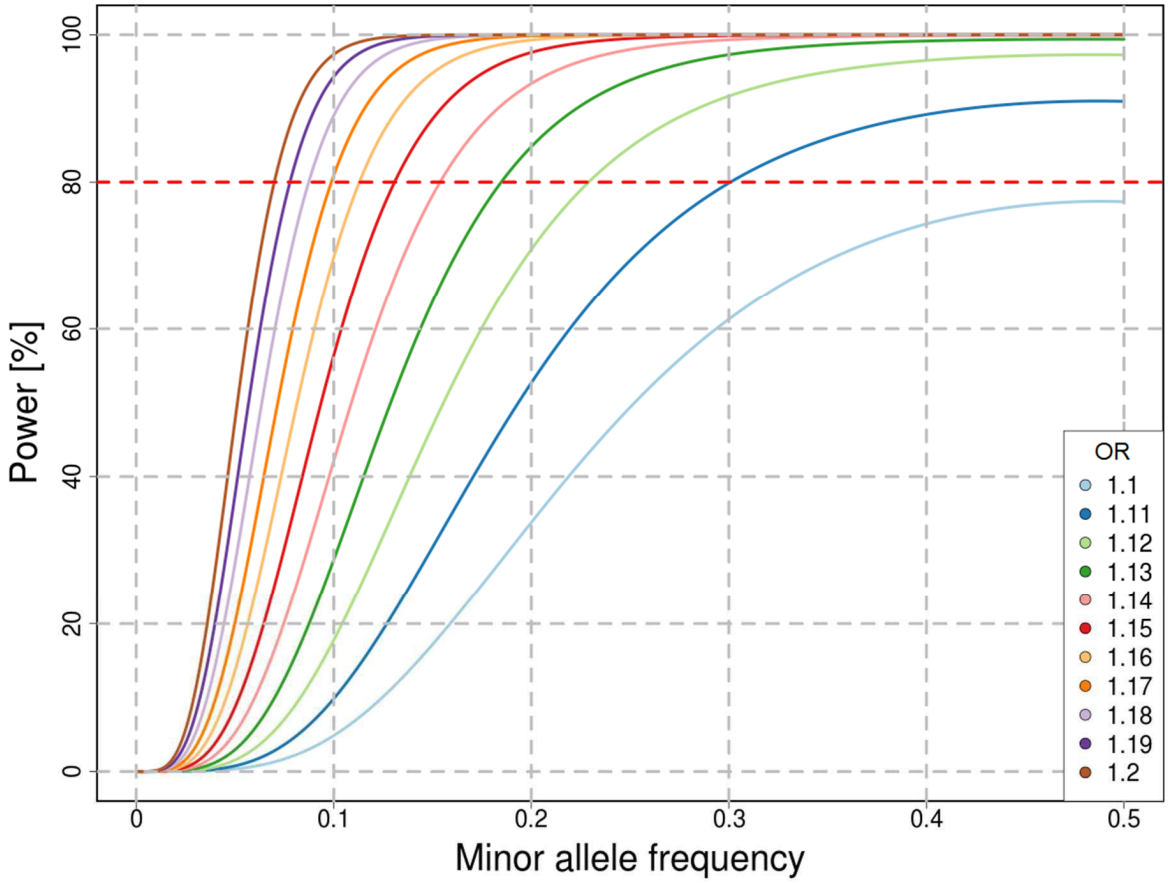

Figure 34. ***Power of variants from the IAMDGC for rare, less frequent and common variants****. Shown is the power (Y) for perfectly imputed variants by MAF (X) for a binary trait with varying odds ratios for 16,144 cases, 17,832 controls.*

7.7.2 Power to detect rare variants with genome-wide significance

Next, I was interested in the power the IAMDGC data yields to detect rare variants. The change of power is visualized in **Figure 35.** For this evaluation, the number of subjects in the analysis in the IAMDGC data, the case-control-ratio of 16,144/17,832 ~= 0.905 and perfect imputation quality was assumed. Panel A varies the estimated odds ratio of the variants. In **Panel B-D** an odds ratio of 1.8 was assumed, which is the odds ratio of the rare lead variant in the C9 locus in the IAMDGC SFS analysis. In **Panels A-D**, the odds ratio, imputation quality, case-control-ratio and the number of subjects in the evaluation were varied.

First, assuming perfect imputation, 80% power yield variants with MAF ≥ 0.196%, 0.253%, 0.344%, **0.507%,** 0.845%, 1.786% and 6.992% for OR of 2.4, 2.2, 2.0, **1.8**, 1.6, 1.4 and 1.2, respectively. A realistic example is the SNP rs622358361 from the C9 locus with $MAF_{controls}$ =0.889% with OR=1.794. A comparable variant with $MAF_{controls}$=0.009 and OR=1.8 yields a power of 99.82% assuming perfect imputation (**Panel A**).

Second, effective sample size (imputation quality * sample size) is taken into account: To demonstrate the loss of power, again a variant with $MAF_{controls}$=0.09% and OR=1.8 is assumed. To yield 80% power for imputation qualities of 1, **0.8**, 0.6, 0.4 and 0.2 a MAF of at least 0.51%, **0.63%,** 0.85%, 1.28% and 2.62% are needed, respectively. For the median imputation quality in rare variants from **Table 20** (median imputation qualities in all 1,206,753 rare variants, $MedianRSQ_{mega\text{-}imputed}$=0.478 and $MedianRSQ_{meta\text{-}imputed}$=0.084) a MAF ≥ 1.07% and ≥ 6.7% is needed, respectively (**Panel B**).

Third, the case-control-ratio in the IAMDGC study data is 16,144/17,832=0.905. Assuming the variant with $MAF_{controls}$=0.09 and OR=1.8 and a constant number of subjects analyzed of 33,976, but varying the case-control ratio between 0.4 and 2, the minimal minor allele frequency needed for 80% power for detecting a truly associated variant with genome wide significance varies only little (0.057%, 0.052%, **0.051%**, 0.051%, 0.053%, 0.055%, 0.057% for a case-control-ratio of 0.5, 0.75, **1.0**, 1.25, 1.5, 1,75 and 2.0, respectively, **Panel C**).

Fourth, assuming the variant with $MAF_{controls}$=0.09, OR=1.8 and a case-control-ratio as in the IAMDGC data (0.905), the 80% power is reached for 1.76%, 0.87%, **0.57%,** 0.43%,0.34%, 0.29%,0.25%,0.22%, 0.19%, and 0.17% when analyzing 10,000, 20,000, **30,000**, 40,000, 50,000, 60,000, 70,000, 80,000 90,000 and 100,000 subjects, respectively (**Panel D**).

As a summary, assuming 16,144 cases and 17,832 controls: even if perfectly imputed, we need a MAF ≥ 0.507% for 80% power for a variant with OR=1.8 (the *C9* variant). To have 80% power to detect the variant with genome wide significance in the analysis of well imputed variants (with imputation quality ≥ 0.8) the variant must have a MAF ≥ 0.63%. The optimal case-control-ratio is 1 and deviating from this optimal setting changes power only marginally. The number of subjects analyzed has strong influence on the power: At least 30k subjects are needed to have at least 80% power for a rare variant

with MAF≥0.57%. Greater number of subjects reduce the minimally needed MAF drastically. Overall the data set from the IAMDGC yields sufficient power to detect rare variants with moderate effects.

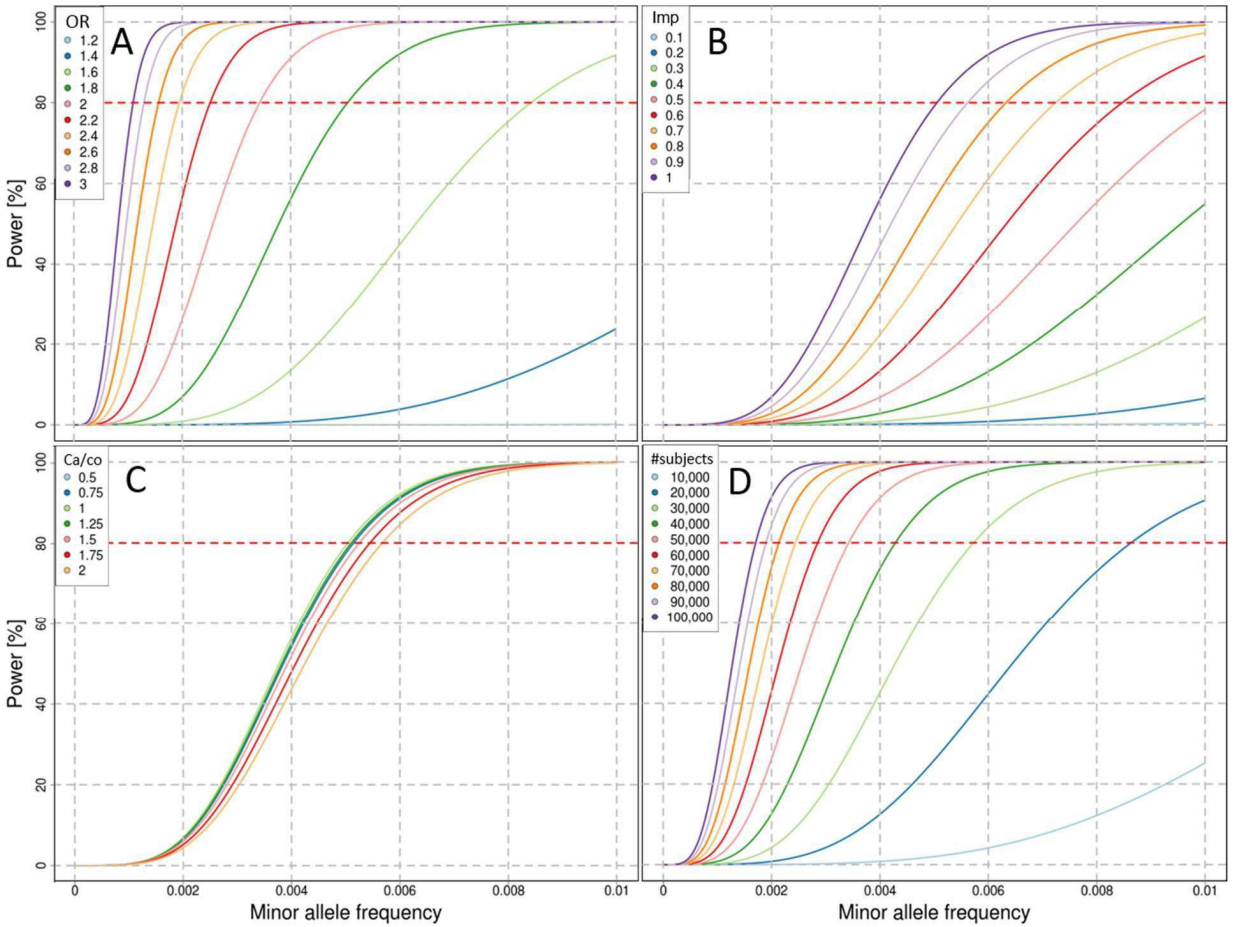

Figure 35. ***Power to detect rare variants depending on odds ratio, imputation quality, case-control-ration and the number of subjects in the analysis****. Shown are scatterplots of the MAF (X-axis) vs. Power [%] (Y-axis). The odds ratios, imputation qualities, case-control-ratios and the numbers of subjects in the analysis were varied in panels A, B, C and D, respectively. Panels A, B and D assumed a case-control-ratio of 0.905 as in the IAMDGC data. 16,144 cases and 17,832 controls were assumed in panels A, B. Panel C assumes an analysis of 33,976 subjects. Perfect imputation quality was assumed in all panels except in panel B. All panels except panel A assume an odds ratio of 1.8 (as observed in the C9 lead variant).*

7.7.3 Power to detect the lead variants in the 34 AMD disease loci

Shown are the odds ratios, MAF in controls and the power to detect the lead variants from the 34 AMD disease loci in **Table 34**.

Table 34. ***Power of the 34 lead variants in the disease loci associated with AMD****. Data is taken from the regression analysis of the genotypes phased and imputed in the mega-analysis.*

Variant ID	Chr	Position (bp)	Closest gene	OR	$MAF_{controls}$	Power
rs10922109	1	196,704,632	*CFH*	2.63	0.43	100.00%
rs11884770	2	228,086,920	*COL4A3*	1.11	0.28	71.00%
rs62247658	3	64,715,155	*ADAMTS9-AS2*	1.14	0.43	99.75%
rs140647181	3	99,180,668	*COL8A1*	1.60	0.02	99.89%
rs10033900	4	110,659,067	*CFI*	1.15	0.48	99.97%
rs114092250	5	35,494,448	*PRLR/SPEF2*	1.43	0.02	79.76%
rs62358361	5	39,327,888	*C9*	1.79	0.01	99.76%
rs116503776	6	31,930,462	*C2/CFB/SKIV2L*	1.75	0.15	100.00%
rs943080	6	43,826,627	*VEGFA*	1.14	0.50	99.75%
rs7803454	7	99,991,548	*PILRB/PILRA*	1.13	0.19	78.31%
rs1142	7	104,756,326	*KMT2E/SRPK2*	1.11	0.35	85.98%
rs79037040	8	23,082,971	*TNFRSF10A*	1.12	0.48	95.49%
rs71507014	9	73,438,605	*TRPM3*	1.10	0.40	69.98%
rs10781182	9	76,617,720	*MIR6130/RORB*	1.11	0.31	83.08%
rs1626340	9	101,923,372	*TGFBR1*	1.14	0.21	89.63%
rs2740488	9	107,661,742	*ABCA1*	1.11	0.28	75.41%
rs12357257	10	24,999,593	*ARHGAP21*	1.11	0.22	68.68%
rs3750846	10	124,215,565	*ARMS2/HTRA1*	2.81	0.21	100.00%
rs3138141	12	56,115,778	*RDH5/CD63*	1.17	0.21	99.87%
rs61941274	12	112,132,610	*ACAD10*	1.50	0.02	98.63%
rs9564692	13	31,821,240	*B3GALTL*	1.12	0.30	89.97%
rs61985136	14	68,769,199	*RAD51B*	1.12	0.48	93.08%
rs2043085	15	58,680,954	*LIPC*	1.14	0.38	99.49%
rs5817082	16	56,997,349	*CETP*	1.19	0.26	100.00%
rs72802342	16	75,234,872	*CTRB2/CTRB1*	1.26	0.08	98.87%
rs11080055	17	26,649,724	*TMEM97/VTN*	1.10	0.49	74.80%
rs6565597	17	79,526,821	*NPLOC4/TSPAN10*	1.13	0.38	99.52%
rs67538026	19	1,031,438	*CNN2*	1.11	0.50	87.99%
rs2230199	19	6,718,387	*C3*	1.43	0.21	100.00%
rs429358	19	45,411,941	*APOE*	1.43	0.14	100.00%
rs142450006	20	44,614,991	*MMP9*	1.17	0.14	93.66%
rs201459901	20	56,653,724	*C20orf85*	1.32	0.07	99.96%
rs5754227	22	33,105,817	*SYN3/TIMP3*	1.29	0.14	100.00%
rs8135665	22	38,476,276	*SLC16A8*	1.14	0.20	91.99%

OR= odds ratio, as visualized in **Figure 34**, which are transformed to show values greater than 1 and the $MAF_{controls}$= minor allele frequency of controls, for the given odds ratio. Power reflects the 80% chance to detect the variant with genome wide significance in the IAMDGC data.

7.8 Imputation qualities of the 34 lead variants in the loci associated with AMD

When quantifying the power to detect rare variants and the gain in power of jointly imputed data compared to meta-imputation, the imputation qualities of untyped genotypes were evaluated in all studies. **Table 36** and **Table 37** show the imputation qualities in the 34 genome wide significant lead variants from **Table 18** per study. The imputation qualities were consistently high among all studies, except for those studies imputed with amplified DNA (*Columbia*, *NHS/HPF* and *UTAH* in genes *NPLOC4/TSPAN* and *CNN2*, also compare **Table 17**). When the meta-imputation was repeated in those studies without subjects with amplified DNA, the imputation qualities were at comparable levels to the other studies, which contained no subjects with amplified DNA (data not shown).

Table 35. ***Comparison of imputation qualities between meta-analysis and mega-analysis****. Shown are the imputation qualities of the 34 lead variants in the loci associated with AMD in the mega-analysis. The imputation quality in the meta-analysis is quantified with the descriptive statistics of the imputation qualities across all 25 studies.*

		Imputation quality in						
		Mega-	Meta-imputation					
		impu-		Lower		Upper		Imp Qual
Variant ID	Closest gene	tation	Min	25	Median	75	Max	(Mega – meta)
rs10922109	CFH	**0.9976**	0.9920	0.9959	**0.9977**	0.9986	0.9996	-0.0001
rs11884770	COL4A3	**0.9803**	0.9565	0.9772	**0.9808**	0.9851	0.9870	-0.0005
rs62247658	ADAMTS9-AS2	**0.9986**	0.9974	0.9983	**0.9986**	0.9992	1.0006	0.0000
rs140647181	COL8A1	**0.7711**	0.6874	0.7424	**0.7605**	0.7711	0.8429	0.0106
rs10033900	CFI	**0.9984**	0.9864	0.9983	**0.9986**	0.9990	1.0005	-0.0003
rs114092250	PRLR/SPEF2	**0.8750**	0.7534	0.8421	**0.8621**	0.8835	0.9293	0.0128
rs62358361	C9	**0.9807**	0.9446	0.9700	**0.9807**	0.9841	0.9905	0.0000
rs116503776	C2/CFB/SKIV2L	**0.9998**	0.9989	0.9998	**1.0001**	1.0004	1.0009	-0.0003
rs943080	VEGFA	**1.0000**	1.0000	1.0002	**1.0003**	1.0004	1.0009	-0.0003
rs7803454	PILRB/PILRA	**0.9972**	0.9661	0.9968	**0.9976**	0.9981	0.9989	-0.0004
rs1142	KMT2E/SRPK2	**0.9866**	0.9800	0.9841	**0.9860**	0.9874	0.9889	0.0006
rs79037040	TNFRSF10A	**0.9975**	0.9726	0.9978	**0.9985**	0.9990	1.0000	-0.0011
rs71507014	TRPM3	**0.9793**	0.9443	0.9824	**0.9867**	0.9894	0.9923	-0.0074
rs10781182	MIR6130/RORB	**0.9951**	0.9854	0.9945	**0.9951**	0.9968	0.9980	0.0000
rs1626340	TGFBR1	**0.9991**	0.9972	0.9992	**0.9994**	0.9997	1.0007	-0.0003
rs2740488	ABCA1	**0.9567**	0.9265	0.9529	**0.9602**	0.9625	0.9676	-0.0035
rs12357257	ARHGAP21	**0.9809**	0.9692	0.9778	**0.9808**	0.9829	0.9858	0.0001
rs3750846	ARMS2/HTRA1	**0.9997**	0.9957	0.9994	**0.9998**	1.0000	1.0009	-0.0001
rs3138141	RDH5/CD63	**0.5827**	0.4232	0.5625	**0.5747**	0.5840	0.6143	0.0079
rs61941274	ACAD10	**0.7353**	0.2241	0.6990	**0.7212**	0.7353	0.7851	0.0140
rs9564692	B3GALTL	**0.9972**	0.9926	0.9966	**0.9974**	0.9980	0.9994	-0.0001
rs61985136	RAD51B	**0.9712**	0.9591	0.9679	**0.9710**	0.9726	0.9865	0.0002
rs2043085	LIPC	**0.9996**	0.9983	0.9998	**1.0000**	1.0002	1.0008	-0.0004
rs5817082	CETP	**0.9856**	0.9683	0.9831	**0.9855**	0.9869	0.9906	0.0001
rs72802342	CTRB2/CTRB1	**0.8494**	0.4671	0.8494	**0.8605**	0.8650	0.8936	-0.0111
rs11080055	TMEM97/VTN NPLOC4/	**0.9797**	0.9430	0.9712	**0.9741**	0.9783	0.9824	0.0056
rs6565597	TSPAN10	**0.8389**	0.2839	0.8619	**0.8766**	0.8805	0.8953	-0.0377
rs67538026	CNN2	**0.8254**	0.0681	0.8057	**0.8197**	0.8240	0.8397	0.0057
rs2230199	C3	**0.9152**	0.7684	0.8940	**0.9138**	0.9212	0.9374	0.0014
rs429358	APOE	**0.9933**	0.9720	0.9906	**0.9939**	0.9950	0.9990	-0.0006
rs142450006	MMP9	**0.9242**	0.8195	0.9173	**0.9251**	0.9331	0.9435	-0.0009
rs201459901	C20orf85	**0.9814**	0.9680	0.9814	**0.9853**	0.9869	0.9921	-0.0039
rs5754227	SYN3/TIMP3	**0.9942**	0.9895	0.9923	**0.9927**	0.9950	0.9968	0.0015
rs8135665	SLC16A8	**0.9994**	0.9936	0.9999	**1.0001**	1.0004	1.0009	-0.0007

Closest gene= gene closest to variant, Chr = chromosome, Pos (bp) = position of variant on GRCh build 37, Mega = imputation quality of variant from mega-imputation, Meta= imputation quality of variant from meta-imputation, Min = minimum, Lower25= 25th percentile, Upper75 = 75th percentile, Max 0 maximum, Imp Qual (Mega-meta) = Difference between imputation quality of mega-imputation and meta-imputation.

Table 36. ***Imputation quality from the mega analysis and the studies from meta-analysis part 1****. Imputation quality is color coded (continuous from red = low to green = high imputation quality).*

Variant ID	Locus name	Chr	Position (bp)	Mega-analysis	AREDS	BDES	Cam-bridge	Cologne	Columbia	CWRU	Edinburgh	EU_JHU	Jerusalem	Mars-hfield	Mel-bourne	Miami
rs10922109	CFH	1	196,704,632	0.9976	0.9955	0.9977	0.9978	0.9958	0.9986	0.9959	0.9993	0.9960	0.9996	0.9966	0.9989	0.9977
rs11884770	COL4A3	2	228,086,920	0.9803	0.9782	0.9861	0.9851	0.9860	0.9585	0.9769	0.9860	0.9808	0.9634	0.9858	0.9836	0.9772
rs62247658	ADAMTS9-AS2	3	64,715,155	0.9986	0.9986	0.9997	0.9994	0.9993	0.9984	0.9982	0.9993	0.9980	1.0006	0.9985	0.9998	0.9989
rs140647181	COL8A1	3	99,180,668	0.7711	0.7560	0.7062	0.7679	0.7648	0.7544	0.7500	0.8049	0.7925	0.7605	0.6874	0.8230	0.7674
rs10033900	CFI	4	110,659,067	0.9984	0.9986	0.9989	0.9995	0.9994	0.9864	0.9988	0.9986	0.9988	1.0005	0.9985	0.9988	0.9990
rs114092250	PRLR/SPEF2	5	35,494,448	0.8750	0.8723	0.8892	0.8835	0.9031	0.8290	0.8621	0.8531	0.8561	0.7576	0.8894	0.8461	0.7534
rs62358361	C9	5	39,327,888	0.9807	0.9810	0.9867	0.9761	0.9771	0.9608	0.9836	0.9905	0.9828	0.9446	0.9819	0.9745	0.9666
rs116503776	C2/CFB/SKIV2L	6	31,930,462	0.9998	0.9990	0.9999	1.0004	1.0004	0.9996	1.0003	1.0009	1.0002	1.0008	1.0000	1.0005	1.0000
rs943080	VEGFA	6	43,826,627	1.0000	1.0001	1.0003	1.0004	1.0004	1.0004	1.0003	1.0009	1.0003	1.0009	1.0000	1.0005	1.0002
rs7803454	PILRB/PILRA	7	99,991,548	0.9972	0.9973	0.9981	0.9983	0.9968	0.9918	0.9974	0.9989	0.9977	0.9933	0.9973	0.9981	0.9980
rs1142	KMT2E/SRPK2	7	104,756,326	0.9866	0.9860	0.9849	0.9877	0.9860	0.9832	0.9841	0.9889	0.9841	0.9800	0.9855	0.9877	0.9857
rs79037040	TNFRSF10A	8	23,082,971	0.9975	0.9983	0.9985	0.9985	1.0000	0.9879	0.9992	0.9994	0.9985	0.9996	0.9974	0.9990	0.9986
rs71507014	TRPM3	9	73,438,605	0.9793	0.9866	0.9881	0.9904	0.9903	0.9708	0.9862	0.9923	0.9867	0.9443	0.9884	0.9894	0.9791
rs10781182	MIR6130/RORB	9	76,617,720	0.9951	0.9959	0.9969	0.9968	0.9970	0.9891	0.9929	0.9979	0.9945	0.9980	0.9952	0.9944	0.9951
rs1626340	TGFBR1	9	101,923,372	0.9991	0.9995	1.0000	0.9997	0.9994	0.9995	0.9992	1.0007	0.9994	0.9972	0.9989	0.9985	0.9999
rs2740488	ABCA1	9	107,661,742	0.9567	0.9610	0.9582	0.9626	0.9612	0.9265	0.9484	0.9666	0.9579	0.9357	0.9602	0.9625	0.9610
rs12357257	ARHGAP21	10	24,999,593	0.9809	0.9792	0.9827	0.9825	0.9755	0.9744	0.9796	0.9858	0.9748	0.9709	0.9808	0.9778	0.9800
rs3750846	ARMS2/HTRA1	10	124,215,565	0.9997	0.9997	0.9993	1.0001	1.0004	0.9957	0.9997	1.0007	0.9998	1.0009	0.9999	0.9998	0.9996
rs3138141	RDH5/CD63	12	56,115,778	0.5827	0.5830	0.6143	0.6064	0.5840	0.5236	0.5535	0.5806	0.5495	0.4232	0.6017	0.5756	0.5625
rs61941274	ACAD10	12	112,132,610	0.7353	0.7268	0.7098	0.7289	0.6853	0.6934	0.6801	0.7851	0.7731	0.2241	0.7059	0.7438	0.7413
rs9564692	B3GALTL	13	31,821,240	0.9972	0.9969	0.9983	0.9974	0.9974	0.9964	0.9960	0.9980	0.9943	0.9980	0.9979	0.9965	0.9980
rs61985136	RAD51B	14	68,769,199	0.9712	0.9676	0.9740	0.9669	0.9773	0.9679	0.9726	0.9591	0.9661	0.9735	0.9710	0.9634	0.9694
rs2043085	LIPC	15	58,680,954	0.9996	0.9999	0.9994	1.0002	1.0004	1.0000	1.0002	1.0008	1.0001	1.0001	0.9992	1.0000	1.0003
rs5817082	CETP	16	56,997,349	0.9856	0.9854	0.9855	0.9906	0.9887	0.9683	0.9821	0.9891	0.9863	0.9722	0.9859	0.9848	0.9831
rs72802342	CTRB2/CTRB1	16	75,234,872	0.8494	0.8620	0.8702	0.8616	0.8619	0.7497	0.8507	0.8858	0.8666	0.7283	0.8936	0.8503	0.8494
rs11080055	TMEM97/VTN	17	26,649,724	0.9797	0.9798	0.9713	0.9712	0.9804	0.9703	0.9763	0.9737	0.9717	0.9430	0.9798	0.9713	0.9783
rs6565597	NPLOC4/TSPAN10	17	79,526,821	0.8389	0.8773	0.8539	0.8876	0.8808	0.3965	0.8766	0.8805	0.8619	0.8590	0.8805	0.8792	0.8759
rs67538026	CNN2	19	1,031,438	0.8254	0.8198	0.7932	0.8227	0.8019	0.1325	0.8057	0.8392	0.8240	0.7500	0.8244	0.8232	0.8118
rs2230199	C3	19	6,718,387	0.9152	0.9235	0.8884	0.9285	0.9153	0.7684	0.9138	0.9374	0.8843	0.8722	0.9212	0.9189	0.9087
rs429358	APOE	19	45,411,941	0.9933	0.9928	0.9946	0.9990	0.9949	0.9720	0.9897	0.9960	0.9939	0.9898	0.9959	0.9910	0.9933
rs142450006	MMP9	20	44,614,991	0.9242	0.9216	0.9280	0.9409	0.9422	0.8440	0.9258	0.9414	0.9236	0.8929	0.9253	0.9317	0.9345
rs201459901	C20orf85	20	56,653,724	0.9814	0.9855	0.9843	0.9873	0.9898	0.9757	0.9779	0.9921	0.9829	0.9799	0.9854	0.9849	0.9853
rs5754227	SYN3/TIMP3	22	33,105,817	0.9942	0.9918	0.9923	0.9966	0.9927	0.9927	0.9925	0.9926	0.9895	0.9927	0.9932	0.9909	0.9950
rs8135665	SLC16A8	22	38,476,276	0.9994	1.0001	1.0002	1.0004	1.0004	0.9936	1.0001	1.0005	1.0000	1.0009	1.0000	1.0005	1.0003

*Table 37.**Imputation quality from the mega analysis and the studies from meta-analysis part 2**. Imputation quality is color coded (continuous from red = low to green = high imputation quality).*

Variant ID	Locus name	Chr	Position (bp)	Mega-analysis	Michigan	NHS HPF	Oregon	Penn	Pitt	Regens-burg	South-ampton	UCSD	UMCN	Utah	UWA LEI Flinders	Vander-bilt	Westmead Sydney
rs10922109	CFH	1	196,704,632	0.9976	0.9933	0.9979	0.9978	0.9987	0.9953	0.9962	0.9991	0.9985	0.9989	0.9920	0.9978	0.9930	0.9990
rs11884770	COL4A3	2	228,086,920	0.9803	0.9803	0.9792	0.9858	0.9655	0.9775	0.9848	0.9843	0.9767	0.9870	0.9565	0.9831	0.9812	0.9809
rs62247658	ADAMTS9-AS2	3	64,715,155	0.9986	0.9979	0.9989	0.9983	0.9980	0.9986	0.9988	0.9982	0.9990	0.9992	0.9974	0.9983	0.9990	0.9992
rs140647181	COL8A1	3	99,180,668	0.7711	0.7283	0.7513	0.7761	0.7623	0.8429	0.7419	0.8280	0.7378	0.7641	0.6880	0.7540	0.7424	0.7519
rs10033900	CFI	4	110,659,067	0.9984	0.9996	0.9948	0.9984	0.9989	0.9983	0.9974	0.9997	0.9986	0.9996	0.9964	0.9976	0.9983	0.9977
rs114092250	PRLR/SPEF2	5	35,494,448	0.8750	0.8750	0.8374	0.8929	0.8415	0.8628	0.8535	0.8933	0.8527	0.9293	0.8367	0.8633	0.8421	0.8711
rs62358361	C9	5	39,327,888	0.9807	0.9851	0.9700	0.9750	0.9640	0.9843	0.9829	0.9875	0.9841	0.9892	0.9568	0.9788	0.9637	0.9674
rs116503776	C2/CFB/SKIV2L	6	31,930,462	0.9998	1.0001	1.0003	0.9990	0.9989	0.9998	0.9997	1.0004	1.0001	1.0004	1.0001	1.0001	1.0003	1.0004
rs943080	VEGFA	6	43,826,627	1.0000	1.0002	1.0003	1.0005	1.0002	1.0002	1.0002	1.0004	1.0000	1.0005	1.0002	1.0001	1.0004	1.0004
rs7803454	PILRB/PILRA	7	99,991,548	0.9972	0.9964	0.9661	0.9981	0.9962	0.9976	0.9980	0.9981	0.9969	0.9989	0.9960	0.9976	0.9982	0.9979
rs1142	KMT2E/SRPK2	7	104,756,326	0.9866	0.9859	0.9837	0.9871	0.9806	0.9870	0.9837	0.9868	0.9879	0.9825	0.9879	0.9874	0.9885	0.9892
rs79037040	TNFRSF10A	8	23,082,971	0.9975	0.9986	0.9726	0.9987	0.9989	0.9978	0.9982	0.9990	0.9985	0.9999	0.9866	0.9986	0.9978	0.9986
rs71507014	TRPM3	9	73,438,605	0.9793	0.9869	0.9815	0.9859	0.9824	0.9894	0.9894	0.9867	0.9842	0.9910	0.9668	0.9850	0.9875	0.9901
rs10781182	MIR6130/RORB	9	76,617,720	0.9951	0.9954	0.9949	0.9973	0.9880	0.9939	0.9968	0.9969	0.9949	0.9964	0.9854	0.9951	0.9948	0.9964
rs1626340	TGFBR1	9	101,923,372	0.9991	0.9998	0.9992	0.9996	0.9996	1.0001	0.9993	0.9992	0.9994	0.9995	0.9992	0.9992	1.0001	0.9992
rs2740488	ABCA1	9	107,661,742	0.9567	0.9625	0.9479	0.9594	0.9313	0.9529	0.9653	0.9676	0.9601	0.9667	0.9468	0.9639	0.9613	0.9547
rs12357257	ARHGAP21	10	24,999,593	0.9809	0.9832	0.9745	0.9832	0.9785	0.9816	0.9807	0.9829	0.9839	0.9826	0.9692	0.9837	0.9852	0.9798
rs3750846	ARMS2/HTRA1	10	124,215,565	0.9997	1.0000	0.9993	1.0000	0.9994	0.9994	0.9989	1.0001	0.9996	1.0003	0.9990	0.9998	0.9998	0.9997
rs3138141	RDH5/CD63	12	56,115,778	0.5827	0.5661	0.5485	0.5736	0.5747	0.5685	0.5673	0.5749	0.5997	0.5897	0.4527	0.5935	0.5747	0.5637
rs61941274	ACAD10	12	112,132,610	0.7353	0.7200	0.5987	0.7287	0.6916	0.7212	0.7345	0.7323	0.7459	0.7203	0.7070	0.6990	0.7446	0.6799
rs9564692	B3GALTL	13	31,821,240	0.9972	0.9959	0.9972	0.9978	0.9926	0.9981	0.9980	0.9966	0.9974	0.9978	0.9994	0.9972	0.9981	0.9977
rs61985136	RAD51B	14	68,769,199	0.9712	0.9693	0.9714	0.9705	0.9775	0.9721	0.9714	0.9656	0.9719	0.9752	0.9865	0.9692	0.9691	0.9702
rs2043085	LIPC	15	58,680,954	0.9996	0.9996	1.0002	0.9998	0.9995	1.0000	0.9998	1.0004	0.9999	1.0004	0.9983	0.9999	0.9998	1.0003
rs5817082	CETP	16	56,997,349	0.9856	0.9863	0.9805	0.9900	0.9816	0.9849	0.9867	0.9870	0.9855	0.9893	0.9739	0.9869	0.9847	0.9833
rs72802342	CTRB2/CTRB1	16	75,234,872	0.8494	0.8605	0.8591	0.8346	0.8288	0.8494	0.8650	0.8796	0.8532	0.8649	0.4671	0.8729	0.8646	0.8351
rs11080055	TMEM97/VTN	17	26,649,724	0.9797	0.9766	0.9771	0.9673	0.9760	0.9708	0.9682	0.9712	0.9762	0.9741	0.9656	0.9824	0.9810	0.9704
rs6565597	NPLOC4/TSPAN10	17	79,526,821	0.8389	0.8807	0.2839	0.8759	0.8707	0.8745	0.8757	0.8852	0.8804	0.8953	0.6594	0.8863	0.8804	0.8810
rs67538026	CNN2	19	1,031,438	0.8254	0.8232	0.0681	0.8197	0.8237	0.8109	0.8154	0.8397	0.8189	0.8169	0.2987	0.8251	0.8316	0.8176
rs2230199	C3	19	6,718,387	0.9152	0.9082	0.8029	0.9143	0.8944	0.9004	0.8940	0.9136	0.9315	0.9329	0.8456	0.9348	0.9157	0.8986
rs429358	APOE	19	45,411,941	0.9933	0.9948	0.9799	0.9943	0.9819	0.9950	0.9966	0.9962	0.9906	0.9961	0.9756	0.9945	0.9914	0.9906
rs142450006	MMP9	20	44,614,991	0.9242	0.9251	0.8195	0.9331	0.9070	0.9124	0.9249	0.9435	0.9209	0.9285	0.8449	0.9410	0.9173	0.9262
rs201459901	C20orf85	20	56,653,724	0.9814	0.9887	0.9844	0.9857	0.9680	0.9920	0.9869	0.9873	0.9868	0.9827	0.9780	0.9856	0.9804	0.9817
rs5754227	SYN3/TIMP3	22	33,105,817	0.9942	0.9937	0.9934	0.9913	0.9917	0.9966	0.9957	0.9968	0.9936	0.9959	0.9926	0.9958	0.9916	0.9948
rs8135665	SLC16A8	22	38,476,276	0.9994	0.9998	0.9991	1.0003	0.9978	1.0005	1.0000	1.0003	1.0001	1.0005	0.9984	0.9999	0.9999	1.0000

7.9 Forestplots of the lead variants in the 34 AMD loci from the mega-imputed and mega-analyzed IAMDGC data

The odds ratios and standard errors in the lead variants of the 34 susceptibility loci from the mega-imputed and mega-analyzed were shown as forestplots in **Figure 36**, **Figure 37** and **Figure 38** . These forestplots illustrated the odds ratios as boxes and the 95% confidence intervals as whiskers of the 25 studies used for the meta-analysis. The studies were sorted by the number of subjects in the analysis in decreasing order. The box size represented the number of subjects in the study. The header gave the gene names of the genome wide associated lead variants from **Table 18**. Then chromosome, position and rs-identifier of the variant were given.

Most studies were direction consistent and standard errors were consistent with the number of subjects analyzed in the study. No or little heterogeneity was furthermore confirmed by the reported heterogeneity in **Table 19**. Studies with whole genome amplified DNA (for example studies *Columbia*, *NHS_HPF* and *Utah* in genes *NPLOC4/TSPAN* and ***CNN2*** in **Figure 38** showed an excess of confidence intervals. When conducting genotype imputation without subjects with amplified DNA, this excess vanished. The lead variants in gene *NPLOC4/TSPAN10* increased imputation qualities from 0.3965 to 0.61, from 0.2839 to 0.57 and from 0.6594 to 0.78 in studies Columbia, NHS and Utah, respectively). The lead variants in gene *CNN2* increased imputation qualities from 0.1325 to 0.87, from 0.0681 to 0.85 and from 0.2987 to 0.82 (in studies Columbia, NHS and Utah, respectively; no further consequences for heterogeneity in meta-analysis shown).

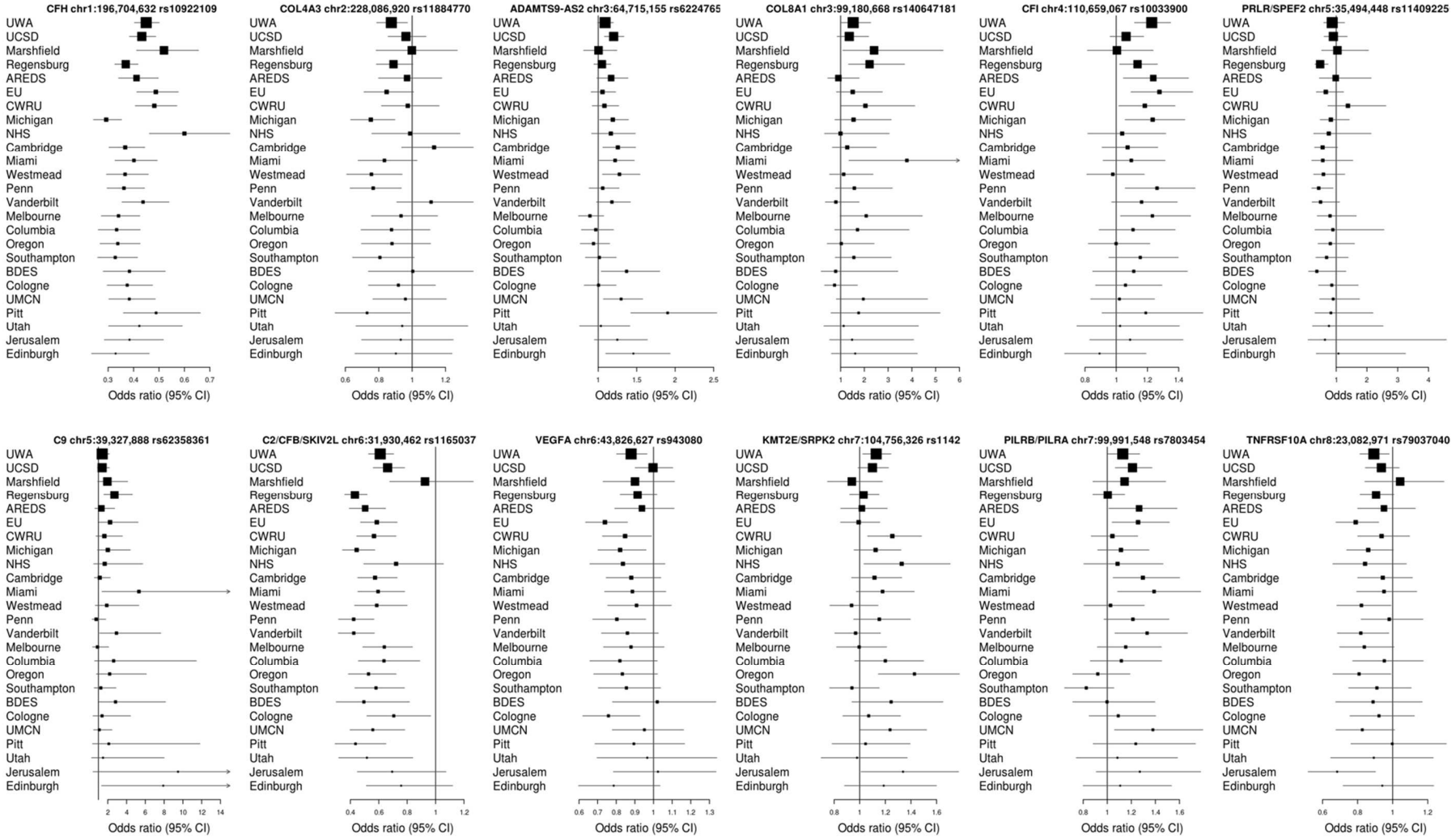

Figure 36. ***Forestplots of 12 loci associated with AMD on chromosomes 1 to 8.***

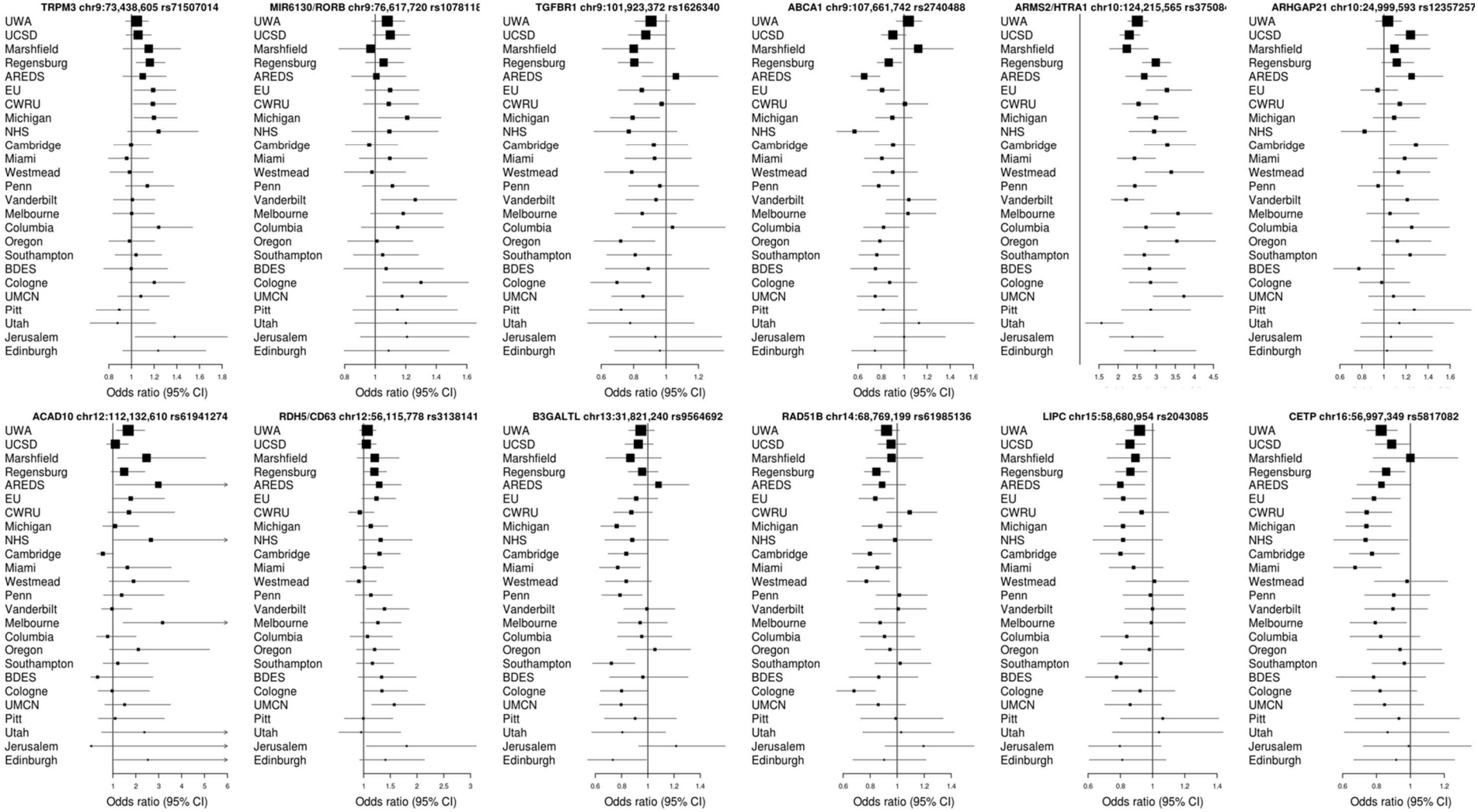

Figure 37. ***Forestplots of 12 loci associated with AMD on chromosomes 8 to 16.***

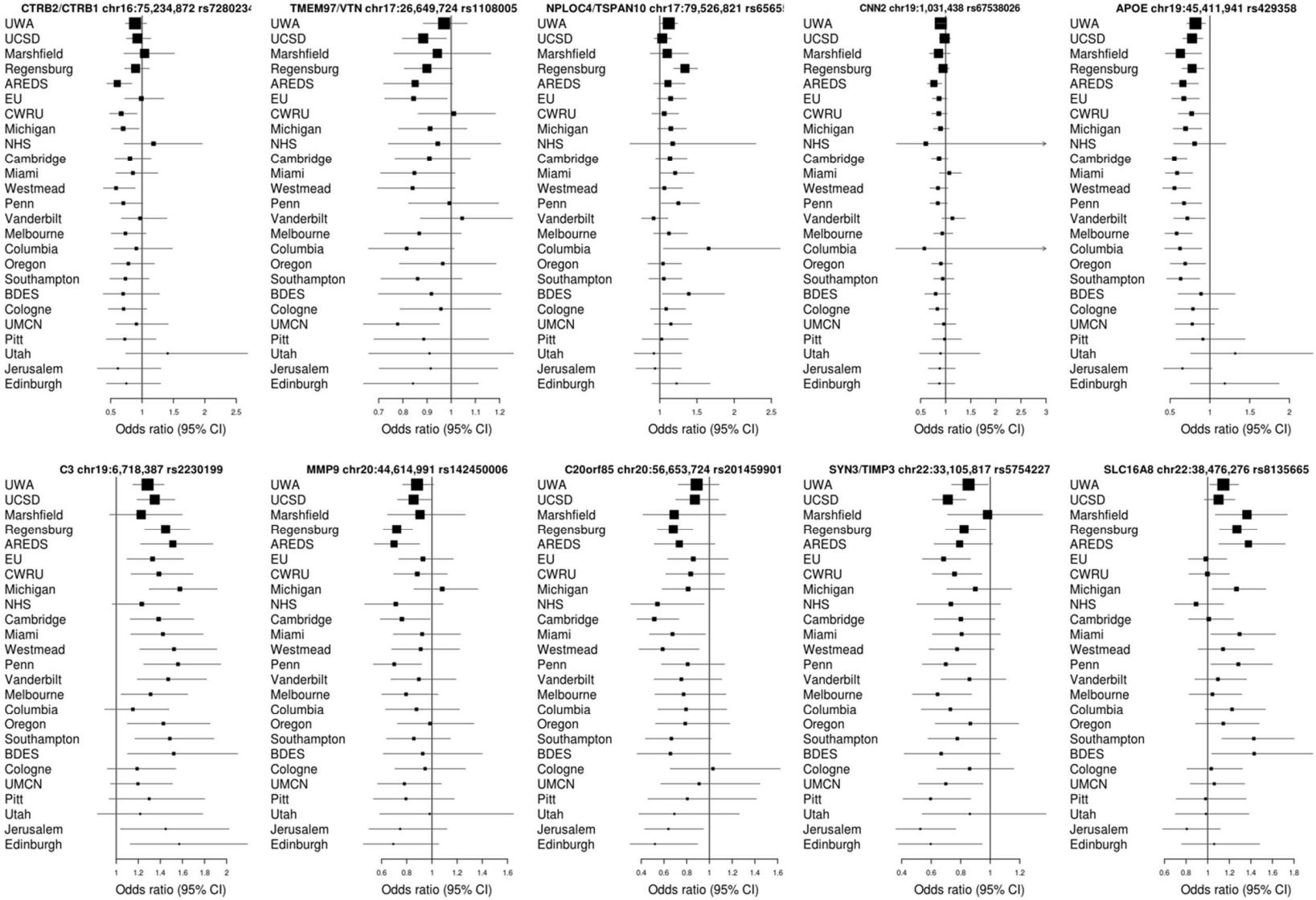

Figure 38. ***Forestplots of 12 loci associated with AMD on chromosomes 16 to 22.***

8 References

1. Collins, F.S., M. Morgan, and A. Patrinos, *The Human Genome Project: lessons from large-scale biology.* Science, 2003. **300**(5617): p. 286-90.
2. Seddon, J.M., et al., *The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences.* Arch Ophthalmol, 2005. **123**(3): p. 321-7.
3. Langefeld, C.D., et al., *Heritability of GFR and albuminuria in Caucasians with type 2 diabetes mellitus.* Am J Kidney Dis, 2004. **43**(5): p. 796-800.
4. Sud, M., et al., *Progression to Stage 4 chronic kidney disease and death, acute kidney injury and hospitalization risk: a retrospective cohort study.* Nephrol Dial Transplant, 2015.
5. Jordan, C.L., et al., *Incidence, risk factors, and outcomes of opportunistic infections in pediatric renal transplant recipients.* Pediatr Transplant, 2015.
6. Holland, P.M., et al., *Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase.* Proc Natl Acad Sci U S A, 1991. **88**(16): p. 7276-80.
7. Distefano, J.K. and D.M. Taverna, *Technological issues and experimental design of gene association studies.* Methods Mol Biol, 2011. **700**: p. 3-16.
8. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
9. Hirschhorn, J.N. and M.J. Daly, *Genome-wide association studies for common diseases and complex traits.* Nat Rev Genet, 2005. **6**(2): p. 95-108.
10. Schuster, S.C., *Next-generation sequencing transforms today's biology.* Nat Methods, 2008. **5**(1): p. 16-8.
11. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
12. Mardis, E.R., *A decade's perspective on DNA sequencing technology.* Nature, 2011. **470**(7333): p. 198-203.
13. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.
14. Albert, A.A., J. A., *On the Existence of Maximum Likelihood Estimates in Logistic Regression Models.* Biometrika, 1984. **71**(1): p. 1-10.
15. Johnson, R.C., et al., *Accounting for multiple comparisons in a genome-wide association study (GWAS).* BMC Genomics, 2010. **11**: p. 724.
16. Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.
17. Fritsche, L.G., et al., *Seven new loci associated with age-related macular degeneration.* Nat Genet, 2013. **45**(4): p. 433-9, 439e1-2.
18. Pattaro, C., et al., *Genome-wide association and functional follow-up reveals new loci for kidney function.* PLoS Genet, 2012. **8**(3): p. e1002584.
19. Winkler, T.W., et al., *The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study.* PLoS Genet, 2015. **11**(10): p. e1005378.
20. Kottgen, A., et al., *New loci associated with kidney function and chronic kidney disease.* Nat Genet, 2010. **42**(5): p. 376-84.
21. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.
22. Howie, B., et al., *Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.* Nat Genet, 2012. **44**(8): p. 955-9.
23. Browning, S.R. and B.L. Browning, *Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering.* Am J Hum Genet, 2007. **81**(5): p. 1084-97.

24. Williams, A.L., et al., *Phasing of many thousands of genotyped samples.* Am J Hum Genet, 2012. **91**(2): p. 238-51.
25. Niu, T., *Algorithms for inferring haplotypes.* Genet Epidemiol, 2004. **27**(4): p. 334-47.
26. Stephens, M. and P. Scheet, *Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation.* Am J Hum Genet, 2005. **76**(3): p. 449-62.
27. Scheet, P. and M. Stephens, *A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.* Am J Hum Genet, 2006. **78**(4): p. 629-44.
28. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2012. **9**(2): p. 179-81.
29. Browning, S.R., *Missing data imputation and haplotype phase inference for genome-wide association studies.* Hum Genet, 2008. **124**(5): p. 439-50.
30. Hawley, M.E. and K.K. Kidd, *HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes.* J Hered, 1995. **86**(5): p. 409-11.
31. Porcu, E., et al., *Genotype imputation in genome-wide association studies.* Curr Protoc Hum Genet, 2013. **Chapter 1**: p. Unit 1 25.
32. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.
33. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
34. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
35. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
36. Anderson, C.A., et al., *Data quality control in genetic case-control association studies.* Nat Protoc, 2010. **5**(9): p. 1564-73.
37. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
38. Geer, L.Y., et al., *The NCBI BioSystems database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D492-6.
39. Placha, G., et al., *A genome-wide linkage scan for genes controlling variation in renal function estimated by serum cystatin C levels in extended families with type 2 diabetes.* Diabetes, 2006. **55**(12): p. 3358-65.
40. Bochud, M., et al., *Heritability of renal function in hypertensive families of African descent in the Seychelles (Indian Ocean).* Kidney Int, 2005. **67**(1): p. 61-9.
41. Fox, C.S., et al., *Genomewide linkage analysis to serum creatinine, GFR, and creatinine clearance in a community-based population: the Framingham Heart Study.* J Am Soc Nephrol, 2004. **15**(9): p. 2457-61.
42. Chasman, D.I., et al., *Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function.* Hum Mol Genet, 2012. **21**(24): p. 5329-43.
43. Kottgen, A., et al., *Multiple loci associated with indices of renal function and chronic kidney disease.* Nat Genet, 2009. **41**(6): p. 712-7.
44. Chambers, J.C., et al., *Genetic loci influencing kidney function and chronic kidney disease.* Nat Genet, 2010. **42**(5): p. 373-5.
45. Pattaro, C., *Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function.* Accepted for publication at Nature Communications doi: 10.1038/ncomms10023.
46. Raychaudhuri, S., et al., *A rare penetrant mutation in CFH confers high risk of age-related macular degeneration.* Nat Genet, 2011. **43**(12): p. 1232-6.
47. Helgason, H., et al., *A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration.* Nat Genet, 2013. **45**(11): p. 1371-4.

48. Seddon, J.M., et al., *Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration.* Nat Genet, 2013. **45**(11): p. 1366-70.
49. Zhan, X., et al., *Identification of a rare coding variant in complement 3 associated with age-related macular degeneration.* Nat Genet, 2013. **45**(11): p. 1375-9.
50. van de Ven, J.P., et al., *A functional variant in the CFI gene confers a high risk of age-related macular degeneration.* Nat Genet, 2013. **45**(7): p. 813-7.
51. Arakawa, S., et al., *Genome-wide association study identifies two susceptibility loci for exudative age-related macular degeneration in the Japanese population.* Nat Genet, 2011. **43**(10): p. 1001-4.
52. Fritsche, L.G., *A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants.* Accepted for publication at Nature Genetics.
53. Levey, A.S., et al., *A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group.* Ann Intern Med, 1999. **130**(6): p. 461-70.
54. Inker, L.A., et al., *Estimating glomerular filtration rate from serum creatinine and cystatin C.* N Engl J Med, 2012. **367**(1): p. 20-9.
55. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
56. Rosner, B., *Fundamentals of Biostatistcs.* 2006.
57. Trudu, M., et al., *Common noncoding UMOD gene variants induce salt-sensitive hypertension and kidney damage by increasing uromodulin expression.* Nat Med, 2013. **19**(12): p. 1655-60.
58. Visscher, P.M., et al., *Five years of GWAS discovery.* Am J Hum Genet, 2012. **90**(1): p. 7-24.
59. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
60. Delaneau, O., J.F. Zagury, and J. Marchini, *Improved whole-chromosome phasing for disease and population genetic studies.* Nat Methods, 2013. **10**(1): p. 5-6.
61. Wichmann, H.E., et al., *KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes.* Gesundheitswesen, 2005. **67 Suppl 1**: p. S26-30.
62. Voight, B.F., et al., *The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits.* PLoS Genet, 2012. **8**(8): p. e1002793.
63. Yu, Y., et al., *Common variants near FRK/COL10A1 and VEGFA are associated with advanced age-related macular degeneration.* Hum Mol Genet, 2011. **20**(18): p. 3699-709.
64. Zheng, J., et al., *A comparison of approaches to account for uncertainty in analysis of imputed genotypes.* Genet Epidemiol, 2011. **35**(2): p. 102-10.
65. Chen, J., et al., *On combining reference data to improve imputation accuracy.* PLoS One, 2013. **8**(1): p. e55600.
66. Feng, J., et al., *Compilation of a comprehensive gene panel for systematic assessment of genes that govern an individual's drug responses.* Pharmacogenomics, 2010. **11**(10): p. 1403-25.
67. Belmonte Mahon, P., et al., *Genome-wide association analysis of age at onset and psychotic symptoms in bipolar disorder.* Am J Med Genet B Neuropsychiatr Genet, 2011. **156B**(3): p. 370-8.
68. Major Depressive Disorder Working Group of the Psychiatric, G.C., et al., *A mega-analysis of genome-wide association studies for major depressive disorder.* Mol Psychiatry, 2013. **18**(4): p. 497-511.
69. Swaroop, A., et al., *Unraveling a multifactorial late-onset disease: from genetic susceptibility to disease mechanisms for age-related macular degeneration.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 19-43.
70. Joachim, N., et al., *The Incidence and Progression of Age-Related Macular Degeneration over 15 Years: The Blue Mountains Eye Study.* Ophthalmology, 2015.

71. Cohen, J., *Statistical power analysis for the behavioral sciences.* Hissldale, NJ: Lawrence Erlabaum Associates, 1988. **2nd edition**.
72. Wood, A.R., et al., *Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation.* PLoS One, 2013. **8**(5): p. e64343.
73. Consortium, C.A.D., *A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease.* Nat Genet, 2015. **47**(10): p. 1121-30.
74. Horikoshi, M., et al., *Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation.* PLoS Genet, 2015. **11**(7): p. e1005230.
75. Kinnersley, B., et al., *Genome-wide association study identifies multiple susceptibility loci for glioma.* Nat Commun, 2015. **6**: p. 8559.
76. Buchanan, C.C., et al., *A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data.* J Am Med Inform Assoc, 2012. **19**(2): p. 289-94.
77. Zheng, H.F., et al., *Performance of genotype imputation for low frequency and rare variants from the 1000 genomes.* PLoS One, 2015. **10**(1): p. e0116487.
78. Huang, L., et al., *Genotype-imputation accuracy across worldwide human populations.* Am J Hum Genet, 2009. **84**(2): p. 235-50.
79. Huang, G.H. and Y.C. Tseng, *Genotype imputation accuracy with different reference panels in admixed populations.* BMC Proc, 2014. **8**(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo): p. S64.
80. Hancock, D.B., et al., *Assessment of genotype imputation performance using 1000 Genomes in African American studies.* PLoS One, 2012. **7**(11): p. e50610.
81. Gao, X., et al., *Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels.* Front Genet, 2012. **3**: p. 117.
82. Scott, L.J., et al., *Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry.* Proc Natl Acad Sci U S A, 2009. **106**(18): p. 7501-6.
83. Ferreira, M.A., et al., *Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder.* Nat Genet, 2008. **40**(9): p. 1056-8.
84. Olkin, I. and A. Sampson, *Comparison of meta-analysis versus analysis of variance of individual patient data.* Biometrics, 1998. **54**(1): p. 317-22.
85. Mathew, T. and K. Nordstrom, *On the equivalence of meta-analysis using literature and using individual patient data.* Biometrics, 1999. **55**(4): p. 1221-3.
86. Lin, D.Y. and D. Zeng, *On the relative efficiency of using summary statistics versus individual-level data in meta-analysis.* Biometrika, 2010. **97**(2): p. 321-332.
87. Lin, D.Y. and D. Zeng, *Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data.* Genet Epidemiol, 2010. **34**(1): p. 60-6.
88. Sung, Y.J., et al., *An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions.* Genet Epidemiol, 2014. **38**(4): p. 369-78.
89. Yamazaki, M., et al., *Effect of obesity on the effectiveness of hormonal contraceptives: an individual participant data meta-analysis.* Contraception, 2015.
90. Kozuki, N., et al., *Short Maternal Stature Increases the Risk of Small-for-Gestational-Age and Preterm Births in Low- and Middle-Income Countries: Individual Participant Data Meta-Analysis and Population Attributable Fraction.* J Nutr, 2015.
91. Saccone, G., et al., *Celiac disease and obstetric complications: a systematic review and meta-analysis.* Am J Obstet Gynecol, 2015.
92. Musso, G., et al., *Association of non-alcoholic fatty liver disease with chronic kidney disease: a systematic review and meta-analysis.* PLoS Med, 2014. **11**(7): p. e1001680.
93. Steinberg, K.K., et al., *Comparison of effect estimates from a meta-analysis of summary data from published studies and from a meta-analysis using individual patient data for ovarian cancer studies.* Am J Epidemiol, 1997. **145**(10): p. 917-25.

94. de Bakker, P.I., et al., *Practical aspects of imputation-driven meta-analysis of genome-wide association studies.* Hum Mol Genet, 2008. **17**(R2): p. R122-8.
95. Evangelou, E. and J.P. Ioannidis, *Meta-analysis methods for genome-wide association studies and beyond.* Nat Rev Genet, 2013. **14**(6): p. 379-89.
96. van Leeuwen, E.M., et al., *Population-specific genotype imputations using minimac or IMPUTE2.* Nat Protoc, 2015. **10**(9): p. 1285-96.
97. Roshyara, N.R. and M. Scholz, *Impact of genetic similarity on imputation accuracy.* BMC Genet, 2015. **16**: p. 90.
98. Kanterakis, A., et al., *Molgenis-impute: imputation pipeline in a box.* BMC Res Notes, 2015. **8**: p. 359.
99. Nothnagel, M., et al., *A comprehensive evaluation of SNP genotype imputation.* Hum Genet, 2009. **125**(2): p. 163-71.
100. Fuchsberger, C., G.R. Abecasis, and D.A. Hinds, *minimac2: faster genotype imputation.* Bioinformatics, 2015. **31**(5): p. 782-4.
101. Reed, E., et al., *A guide to genome-wide association analysis and post-analytic interrogation.* Stat Med, 2015.
102. Coleman, J.R., et al., *Quality control, imputation and analysis of genome-wide genotyping data from the Illumina HumanCoreExome microarray.* Brief Funct Genomics, 2015.

9 Web Resources

1000 Genomes Project homepage

http://browser.1000genomes.org/index.html

1000 Genomes reference panels

http://csg.sph.umich.edu/abecasis/MaCH/download/1000G.2013-09.html

http://csg.sph.umich.edu/abecasis/MaCH/download/1000G.Phase3.v5.html

ChunkChromosome

http://www.sph.umich.edu/csg/cfuchsb/generic-ChunkChromosome-2012-08-28.tar.gz

Haplotype Reference consortium

http://www.haplotype-reference-consortium.org/home

Cookbook ImputeV2

http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

Cookbook MaCH/ minimac

http://genome.sph.umich.edu/wiki/Minimac:_GIANT_1000_Genomes_Imputation_Cookbook

dbSNP

http://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi

Galaxy – an open, web-based platform for data intensive biomedical research

https://usegalaxy.org

GIANT ALL 1000G PhaseI v3 reference panel

http://www.sph.umich.edu/csg/abecasis/MaCH/download/1000G.2012-03-14.html

HapMap Phase II release 22 reference panel

http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml

LD of subjects with European ancestry

http://www.ensembl.org/info/genome/variation/data_description.html

LiftOver online tool and wiki page from the University of California

https://genome.ucsc.edu/cgi-bin/hgLiftOver

http://genome.sph.umich.edu/wiki/LiftOver

NCBI remapping service

http://www.ncbi.nlm.nih.gov/genome/tools/remap

Power analysis: R package 'pwr'

https://cran.r-project.org/web/packages/pwr/index.html

Pubmed

http://www.ncbi.nlm.nih.gov/pubmed

UCSC Lift-Over Chain File

http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver

RSMergeArch

ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/

UK 10K – rare genetic variants in health and disease

http://www.uk10k.org

SNPChrPosOnRef

ftp://ftp.ncbi.nih.gov/snp/organisms/human_9606/database/organism_data/

The International AMD Genomics Consortium

http://eaglep.case.edu/iamdgc_web

The International HapMap project

ftp://ftp.ncbi.nlm.nih.gov/hapmap

The 1000 Genomes Project Phase I version 3 reference data set

http://www.1000genomes.org/category/frequently-asked-questions/population

List of abbreviations

1000G	1000 Genomes
AMD	Age-related Macular Degeneration
CIDR	Center for Inherited Diseases Research
cM	Centi Morgan
eGFR	Estimated Glomerular Filtration Rate
eGFRcrea	Estimated Glomerular Filtration Rate based on creatinine measurement
GB	Gigabyte
GWAS	Genome Wide Association Study
GWAMA	Genome Wide Association Meta-Analysis
H_0	Null hypothesis
H_A	Alternative hypothesis
IAMDGC	International Age-related Macular Degeneration Genomics Consortium
IPD	Individual Participant Data
LD	Linkage Disequilibrium
MAC	Minor allele count
MAF	Minor Allele Frequency
MB	Megabytes
QQ-Plot	Quantile-Quantile-Plot
RSQ	R-Squared or coefficient of determination for quality of an imputed variant
SFS	Sequential forward selection
SNP	Single Nucleotide Polymorphism
WGA	Whole Genome Amplified

List of publications

Publications directly related to this work and description of own contribution

Part of this work has been published or accepted for publication:

Gorski, M., Winkler, T.W., Stark, K., Muller-Nurasyid, M., Ried, J.S., Grallert, H., Weber, B.H., and Heid, I.M. (2014). Harmonization of Study and Reference Data by PhaseLift: Saving Time When Imputing Study Data. Genet Epidemiol 38, 381-388.

- ➔ **Own contribution:** Methods and software development for all analysis; data preparation and quality control; interpretation of results, paper writing.

Pattaro C.*, Teumer A.*, **Gorski M.***, Chu, A.*, Li, M.*, Mijatovic, V., Garnaas M., et al. Genetic Associations at 53 Loci Highlight Cell Types and Biologic Pathways for Kidney Function.
* joint contribution

- ➔ **Own contribution:** Statistical analysis, data preparation, quality control, meta-analysis, interpretation of results.
- ➔ **Status:** Accepted for publication at Nature Communications.

Fritsche L., Igl W.,Cooke Bailey J., Grassmann, F., …, **Gorski M.**, …, Heid I. et al. Insights into Rare and Common Genetic Variation From a Large Study of Age-Related Macular Degeneration.

- ➔ **Own contribution:** Quality control and imputation. Contribution to statistical analysis and interpretation of results.
- ➔ **Status:** Accepted for publication in Nature Genetics.

Further work is currently in the stage of manuscript writing:

Gorski M., Fuchsberger C., Fox C. et al. Meta-analysis of high density 1000G imputed data reveal 10 novel Genetic Associations for Kidney Function.

- ➔ **Own contribution:** Statistical analysis, data preparation, quality control and meta-analysis, interpretation of results, paper writing.
- ➔ **Status:** In writing.

Mathias Gorski et al. How much can we gain from Individual Participant Study Data as compared to meta-analysis of study-specific statistics in the light of imputed data?

- ➔ **Own contribution:** Statistical analysis, data preparation, quality control and meta-analysis, interpretation of results, paper writing.
- ➔ **Status:** In writing.

Other first authored work related to kidney function genetics:

Gorski, M., , Tin A, Garnaas M, McMahon GM, Chu AY, Tayo BO, Pattaro C, Teumer A, Chasman DI, Chalmers J, et al., Genome-wide association study of kidney function decline in individuals of European descent. Kidney Int, 2015. **87**(5): p. 1017-29.

- ➔ **Own contribution:** Statistical analysis, Data preparation, quality control and meta-analysis, Interpretation of results; Paper writing.

Other co-authored work on genome-wide association analysis or methods for genome-wide association analysis

Chasman, D.I Fuchsberger C, Pattaro C, Teumer A, Boger CA, Endlich K, Olden M, Chen MH, Tin A, Taliun D, Li M, Gao X, **Gorski M**, Yang Q, Hundertmark C, et al., *Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function.* Hum Mol Genet, 2012. **21**(24): p. 5329-43.

Parsa A, … , **Gorski M**, … , Fox CS, Kao WL, Böger CA, Common variants in Mendelian kidney disease genes and their association with renal function, J Am Soc Nephrol. 2013 Dec;24(12):2105-17. doi: 10.1681/ASN.2012100983. Epub 2013 Sep 12.

Further published work related to genome-wide association analysis and description of susceptibility loci for complex diseases

Winkler, T. W., Justice, A. E., Graff, M., Barata, L., Feitosa, M. F., Chu, S., ..., **Gorski M**, ..., Loos, R. J. (2015). The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. PLoS Genet, 11(10), e1005378. doi: 10.1371/journal.pgen.1005378.

The Interleukin Genetics Consortium (2015). Cardiometabolic effects of genetic upregulation of the interleukin 1 receptor antagonist: a Mendelian randomisation analysis. Lancet Diabetes Endocrinol 3, 243-253.

Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, Powell C, Vedantam S, Buchkovich ML, Yang J, ..., **Gorski M**, ..., et al. 2015. Genetic studies of body mass index yield new insights for obesity biology. Nature 518(7538):197-206.

Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Justice, A.E., ..., **Gorski M**, ..., at al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187-196.

Winkler, T.W., Kutalik, Z., **Gorski, M.**, Lottaz, C., Kronenberg, F., and Heid, I.M. (2014). EasyStrata: Evaluation and Visualization of stratified genome-wide association meta-analysis data. Bioinformatics.

Yoneyama S, … , **Gorski M**, …, Yusuf S; the GIANT Consortium; the CARe IBC Consortium, Hakonarson H, Lange LA, Demerath EW, Fox CS, North KE, Reiner AP, Keating B, Taylor KC, Gene-centric meta-analyses for central adiposity traits in up to 57 412 individuals of European descent confirm known loci and reveal several novel associations, Hum Mol Genet. 2014 Jan 6.

Dorhofer, L., Lammert, A., Krane, V., **Gorski, M.**, Banas, B., Wanner, C., Kramer, B.K., Heid, I.M., and Boger, C.A. (2013). Study design of DIACORE (DIAbetes COhoRtE) - a cohort study of patients with diabetes mellitus type 2. BMC Med Genet 14, 25.

Burghardt T, Kastner J, Suleiman H, Rivera-Milla E, Stepanova N, Lottaz C, Kubitza M, Böger CA, Schmidt S, **Gorski M**, de Vries U, Schmidt H, Hertting I, Kopp J, Rascle A, Moser M, Heid IM, Warth R, Spang R, Wegener J, Mierke CT, Englert C, Witzgall R., LMX1B is essential for the maintenance of differentiated podocytes in adult kidneys, J Am Soc Nephrol. 2013 Nov;24(11):1830-48. doi: 10.1681/ASN.2012080788. Epub 2013 Aug 29.

Pinto, L.A., Michel, S., Klopp, N., Vogelberg, C., von Berg, A., Bufe, A., Heinzmann, A., Laub, O., Simma, B., Frischer, T., Genuneit J, **Gorski M**, Illig T, Kabesch M. (2013). Polymorphisms in the IRF-4 gene, asthma and recurrent bronchitis in children. Clin Exp Allergy 43, 1152-1159. Oct 2013.

Pattaro, C., Kottgen A, Teumer A, Garnaas M, Boger CA, Fuchsberger C, Olden M, Chen MH, Tin A, Taliun D, Li M, Gao X, **Gorski M**, Yang Q, Hundertmark C, (2012) Genome-wide association and functional follow-up reveals new loci for kidney function. PLoS Genet 8: e1002584.

Guo Y, … , **Gorski M**, … , Keating Dagger BJ.(2012) Gene-centric meta-analysesof 108 912 individuals confirm known body mass index loci and reveal three novel signals. Hum Mol Genet.

Boger CA, **Gorski M**, Li M, Hoffmann MM, Huang C et al; CKDGen Consortium. (2011) Association of eGFR-Related Loci Identified by GWAS with Incident CKD and ESRD. PLoS Genet 7: e1002292.

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

Selbstständigkeitserklärung

Ich, Gorski, Mathias geboren am 19. Mai 1981 in Neumarkt i. d. OPf, 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. November 2015 Mathias Gorski

Acknowledgements

First of all, I want to thank Prof. Dr. Iris Heid, Head of the Department of Genetic Epidemiology at the University of Regensburg, for her constant support and for introducing me to the great world of Genetic Epidemiology and the many possibilities it offers me as a computer scientist. I am grateful for the inspiring work environment with the manifold chances to present my work at national and international conferences and for the chance to work in the CKDGen consortium and the IAMDGC.

I would like to thank Prof. Dr. Carsten Böger, Department of Nephrology at the University Hospital Regensburg for his constant advice and encouragement and for giving me the chance to be part of the genetic kidney working group of Regensburg. Many thanks for our collaborative work in the CKDGen consortium as well as in the Diacore and Gendian studies and for the support at all times.

Many thanks to Prof. Dr. Bernhard Weber, Head of the Department of Human Genetics at the University of Regensburg for mentoring this thesis and for enriching my work with many fruitful ideas and advices.

Next I would like to thank Prof. Dr. Rainer Spang, Head of the Institute of Functional Genomics at the University of Regensburg for mentoring this thesis, for his advice and for our great discussions.

I would like to thank all my collaborators from the CKDGen consortium. Among many others my special thanks to Prof. Dr. Anna Köttgen, Prof. Dr. Caroline Fox and Dr. Cristian Pattaro for many discussions and their advice on meetings and on our weekly telephone conferences.

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 Dr. Thomas Winkler for his statistical advice, Dr. Matthias Olden for our great team work on any topic and PD Dr. Klaus Stark for our discussions on human biology and on anything else.

Finally, I am grateful to my wife Stefanie who never failed with encouragement, advice and her endless patience and understanding.