Statistical Methods and Applications for Biomarker Discovery Using Large Scale Omics Data Set

Inauguraldissertation

zur

Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Universität Greifswald

vorgelegt von

Sahar Ghasemi

Greifswald, April 2023

Dekan: Prof. Dr. Gerald Kerth

Erstgutachter: Prof. Dr. Volkmar Liebscher

Zweitgutachter: Prof. Dr. Inke R. König

Tag der Promotion: 07.09.2023

Table of Contents

1 **INTRODUCTION** 1

1.1 BACKGROUND AND AIM 1
1.2 IDENTIFYING GENETIC SUSCEPTIBILITY LOCI 1
1.2.1 Genetic markers 1
1.2.2 Haplotype and linkage disequilibrium (LD) 2
1.2.3 Genome-wide association study (GWAS) 3
1.2.4 Meta-analysis of GWAS results 7
1.2.5 Significance level corrections for multiple comparisons 8
1.2.6 Conditional analysis of GWAS summary statistics 9
1.3 ADDITIONAL OMICS DATA AS A BASIS FOR LOCUS DISCOVERY 10
1.4 ASSOCIATION WITH GENE EXPRESSION AND COLOCALIZATION 10
1.4.1 Expression quantitative trait loci (eQTL) study 10
1.4.2 Colocalization- integration of GWAS and eQTLs 12
1.5 QUASI-ADAPTIVE METHOD 13
1.5.1 Improved quasi-adaptive method 15
1.5.2 Conditional analysis in quasi-adaptive method 15

2 **SUMMARY** 16

2.1 SUMMARY OF THE THESIS 16
2.2 SUMMARY OF THE PUBLICATIONS 17
2.2.1 Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria (Article I) 17
2.2.2 Assessment of significance of conditionally independent GWAS signals (Article II) 19
2.2.3 Discovery of novel eGFR-associated multiple independent signals using quasi-adaptive method (Article III) 20
2.2.4 Association studies 22

3 **REFERENCES** 23

AUTHOR CONTRIBUTIONS 26

GENOME-WIDE ASSOCIATION META-ANALYSES AND FINE-MAPPING ELUCIDATE PATHWAYS INFLUENCING ALBUMINURIA (ARTICLE I) 26
ASSESSMENT OF SIGNIFICANCE OF CONDITIONALLY INDEPENDENT GWAS SIGNALS (ARTICLE II) 26
DISCOVERY OF NOVEL EGFR-ASSOCIATED MULTIPLE INDEPENDENT SIGNALS USING A QUASI-ADAPTIVE METHOD (ARTICLE III) 26

ARTICLES 27

GENOME-WIDE ASSOCIATION META-ANALYSES AND FINE-MAPPING ELUCIDATE PATHWAYS INFLUENCING ALBUMINURIA (ARTICLE I) 28
ASSESSMENT OF SIGNIFICANCE OF CONDITIONALLY INDEPENDENT GWAS SIGNALS (ARTICLE II) 47
DISCOVERY OF NOVEL EGFR-ASSOCIATED MULTIPLE INDEPENDENT SIGNALS USING A QUASI-ADAPTIVE METHOD (ARTICLE III) 56

APPENDIX 69

EIGENSTÄNDIGKEITSERKLÄRUNG 69
CURRICULUM VITAE 70
LIST OF PUBLICATIONS 71
FUNDING 72
ACKNOWLEDGMENT 73

1 Introduction

1.1 Background and aim

Understanding the causes of common diseases and estimating their risk factors are biomedical research's fundamental goals to develop new treatments and medications. Many of the most common diseases, known as complex diseases, are caused by multiple environmental, lifestyle, and genetic risk factors. Lifestyle modifications and medications have been quite successful in treating various ailments. Nevertheless, a better understanding of the molecular mechanisms underlying complex diseases and their potential interactions with other factors is necessary to identify possible causes of diseases and develop more effective treatments. This thesis aims to contribute to our understanding of the genetic basis of complex human diseases by utilizing multi-omics data, including DNA polymorphism variation, and integrating various levels of molecular data to uncover the biological mechanisms underlying the findings. The primary objective is identifying novel genetic markers related to kidney function by conducting trans-ethnic GWAS meta-analyses of multiple studies with large sample sizes. To identify additional independent genetic markers, a statistical method called the quasi-adaptive method has been developed which assesses the significance level of secondary signals in GWAS conditional analysis. Furthermore, the method has been improved and applied to a prior report on kidney function to reveal more trait-associated genetic variants. This research aims to gain insights into the underlying mechanisms of complex diseases, paving the way for more effective treatments and medications.

1.2 Identifying genetic susceptibility loci

1.2.1 Genetic markers

A genetic marker is a sequence of DNA with a known physical location on a chromosome used to identify individuals, populations, species, or genes involved in inherited disease. The genome of an individual may differ from others in numerous ways, including base differences known as single nucleotide polymorphisms (SNPs), insertions or deletions (INDELs), or differences in the number of copies of a sequence or gene (copy number variations (CNV)). The most common type of genetic marker are SNPs (Figure 1) occur when a single nucleotide adenine (A), thymine (T), cytosine (C), or guanine (G) in the genome differs between individuals.

Genetic markers can refer to genes associated with various complex traits or common diseases. When SNPs occur in coding or non-coding regions (a regulatory region), they may significantly affect the function of the gene(s) and influence diseases. In this context, identifying associated genetic loci is the first step toward deciphering disease-related biological pathways and understanding the etiology of a specific illness.

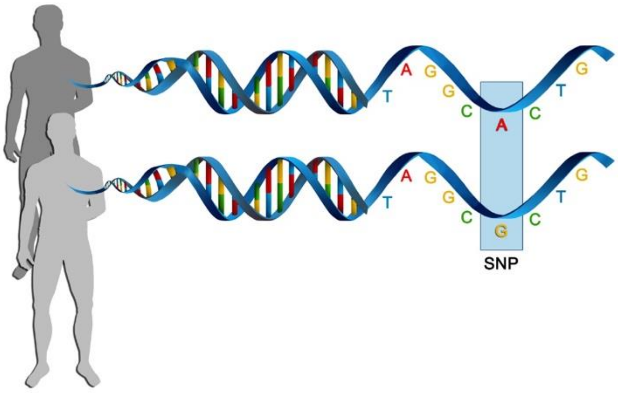


Figure 1. Depicts two sequenced DNA fragments from different individuals containing a difference in a single nucleotide A and G. This variation in the DNA sequence is referred to as an SNP with two alleles, A and G. Reprinted from "Single nucleotide polymorphisms as genomic markers for high-throughput pharmacogenomic studies" by Lonetti A. et al. (2016), Methods Mol Biol, 1368, 143-59. Copyright © 2016, Springer Science Business Media New York.

1.2.2 Haplotype and linkage disequilibrium (LD)

A haplotype is a collection of specific SNPs alleles at adjacent loci belonging to the same chromosome that are statistically associated and tend to be inherited together. Identification of these statistical associations and fewer alleles of a particular haplotype sequence can facilitate the identification of all other nearby polymorphic sites on a chromosome. Such information is critical for research into the genetics of common diseases. For example, if there are SNP_A with two alleles, T and G, at locus one and SNP_B with two alleles, G and A, at locus two, the corresponding plausible haplotypes are TG, TA, GG, and GA. Theoretically, if the association between the allele T and G is supposed to be random and the allele frequency $P_A(T)$ and $P_B(G)$ are known. In this case, the frequency for each haplotype, for instance, $P_{AB}(TG)$ can be calculated as $P_{AB}(TG) = \; P_A(T) \times P_B(G)$. Sometimes there is a difference between the theoretical haplotype frequency and empirical haplotype frequency in a population. This difference $D = P_{AB}(TG) - P_A(T)\, P_B(G)$ is a measure of linkage disequilibrium (LD), which refers to a non-random association between tightly linked SNPs at different loci in a given population[1].

Since the range of D depends on the frequencies of the alleles to which it refers, this is not straightforward to compare the extent of linkage disequilibrium between different pairs of alleles. Lewontin, R. C. (1964)[2] suggested normalizing D as follows:

$$D^{'} = \frac{D}{D_{max}} \quad \text{where,} \quad D_{max} = \begin{cases} \max\{-P_A(T)\, P_B(G), -\big(1 - P_A(T)\big)\big(1 - P_B(G)\big)\}, & D < 0 \\ \min\{P_A(T)\big(1 - P_B(G)\big),\; P_B(G)\big(1 - P_A(T)\big)\}, & D > 0 \end{cases}$$

An alternative to $D^{'}$ is the pairwise correlation coefficient (r^2) between two SNPs as follows:

$$r^2 = \frac{D^2}{P_A(T)(1 - P_A(T))P_B(G)(1 - P_B(G))}$$

Two SNPs are in complete LD if $D^{'} = 1$ and $r^2 < 1$ and are in perfect LD if $D^{'} = 1$ and $r^2 = 1$.

1.2.3 Genome-wide association study (GWAS)

Genome-wide association study (GWAS) is an observational approach widely used to identify genetic variants, mainly SNPs, statistically associated with a particular complex trait or a common disease[3,4]. GWAS examines SNPs across the entire set of DNA (the whole genome) from a large group of participants to identify variations that account for disease risk. These participants may be people with a disease (cases) and people without the disease (controls), or people with different phenotypes for a particular trait.

To perform GWAS, first, each individual must be genotyped using available genotyping arrays or sequencing technologies. Second, the disease outcome or non-disease-related trait is measured for each individual. Third, a statistical association test is performed at the genome-wide level to look for genetically associated SNPs. The associated SNPs are then considered to highlight the genomic regions that may influence disease risk.

1.2.3.1 Sample description

I conducted the analyses based on individual-level data using the Study of Health in Pomerania (SHIP)[5] and UK-Biobank (UKBB)[6].

SHIP cohort

SHIP is a population-based epidemiological study in Western Pomerania, the northeast area of Germany, comprising two prospective independent cohorts, SHIP-Start (The initial cohort SHIP was renamed to SHIP-Start to avoid confusion) and SHIP-Trend. Both cohorts were collected from the general adult population aged 20–79. First Baseline examinations of SHIP-Start (SHIP-Start-0) were performed from 1997-2001, with the final sample of 4,308 participants (response 68.8%). The mean age of the SHIP-Start-0 baseline sample was 50.3 years (±16.4 years) and 50.9% of participants were women. First Baseline examinations of SHIP-Trend (SHIP-Trend-0) were performed from 2008-2012, with the final sample of 4,420 participants (response 50.1%). The mean age of the SHIP-Trend-0 baseline sample was 52.0 years (±15.5 years), with 51.4% women included. The main objective of SHIP is to assess the prevalence and incidence of common diseases and their risk factors. The study design has been previously described in detail in the respective publication[5].

UKBB cohort

The UKKB is a large prospective cohort study with deep genetic data and a wide variety of phenotypic information collected on approximately 500,000 individuals aged between 40 and 69 across the United Kingdom. The first baseline information was collected between 2006 and 2010. The study continues to collect detailed information about participants' demographic, lifestyles, health-related factors, and physical measures. The mean age of the UKBB baseline sample was 56.53 years (±15.87 years) and 54.4% of participants were women. The UKBB aims to offer samples for different studies undertaking vital research into the most common and life-threatening diseases. The study design has been previously described comprehensively in the respective publications[6,7].

The CKDGen consortium

The CKDGen consortium[8] [https://ckdgen.imbi.uni-freiburg.de/], focusing on the genetic basis of kidney function, was conducted in international consortia to increase the power of the statistical analyses by increasing the sample size. Samples from multiple cohorts (including SHIP) were analyzed individually according to a centrally generated analysis plan and then subjected to meta-analysis. The projects carried out within this consortium include more than 100 individual studies with more than one million samples.

1.2.3.2 Array-based genotyping analyses

SHIP cohort

In 2008, 4252 samples from the SHIP-Start were genotyped at 909,622 SNPs and 946,000 additional non-polymorphic copy number probes by the Affymetrix Genome-Wide Human SNP Array 6.0 at Affymetrix, Inc. (Santa Clara, CA, USA) using the Birdseed2 clustering algorithm. Following this, at the end of 2010, a subset of SHIP-Trend (SHIP-Trend multi-OMICS) including 1000 samples, was selected for genotyping at 2.45 million genetic variants using the Illumina Human Omni 2.5 array. Genotyping was performed at the Helmholtz Zentrum München, Munich, Germany.

The genotyping quality was checked for all arrays, and the corresponding SNPs were called and quality controlled (QC) using an established workflow. The QC steps included filters for both low array genotyping efficiency (< 94% or < 92% depending on array type), SNPs call rates (< 95%), Hardy-Weinberg equilibrium ($pHWE \leq 10^{-4}$), monomorphic SNPs, and excessive heterozygosity as indicators of genotyping errors (more than ±4 standard deviations of the mean). Additional filters included mismatches between reported and genetically estimated gender, individuals call rates (< 94%), and duplicate samples. Furthermore, sample outlier detection was applied using principal component analyses (more than ±8 standard deviations of the mean for the first 10 principal components and five iterations).

UKBB cohort

Two similar genotyping arrays were used to assay 488,377 participants involved in the UKBB. A subset of 49,950 participants in the UKBB Lung Exome Variant Evaluation (UK BiLEVE) study was genotyped (807,411 markers) using the Applied Biosystems™ UK BiLEVE Axiom™ Array by Affymetrix. The rest of the 438,427 participants were genotyped (825,927 markers) using the closely-related Applied Biosystems™ UK Biobank Axiom™ Array[9].

The QC steps included filters for variants that showed batch effects, plate effects, departures from HWE, sex effects, array effects, discordance across control replicates, SNPs call rates (< 95%), and MAF < 10^{-4}. In addition, samples with ancestry outliers, outliers for heterozygosity, and missingness (high heterozygosity or > 5% missing rate) were removed as described comprehensively in the respective publication[9].

1.2.3.3 Imputation of genotype SNPs

An essential step in most GWAS is the imputation of genotype SNPs. This process dramatically increases the number of SNPs that can be tested for association, increases the power of the study, and facilitates meta-analysis of GWAS across distinct cohorts[10]. Different datasets may have used different genotyping platforms and may have different genotyped variants containing a non-overlapping set of SNPs. Each study imputed the genotype data before running GWAS by imputation methods. These methods integrate the LD structure obtained from an SNP reference data set such as the HapMap[11,12], the 1000 Genomes[13], or the Haplotype Reference Consortium (HRC)[14] to infer the alleles of missing SNPs in the study.

SHIP cohort

Genotyped SHIP data were imputed based on the HapMap II (nSNPs=2.5 million), the 1000 Genomes (nSNPs=16 million), and the HRC (nSNPs=40 million) reference panels. The imputation to the HRC panel was performed using the Michigan[15] [https://imputationserver.sph.umich.edu/] and Sanger[16] [https://imputation.sanger.ac.uk/] imputation servers. The imputation to the HRC was performed by myself using the Michigan imputation server. The SNPs were annotated according to the GRCh37 (hg19) reference build.

UKBB cohort

Genotyped UKBB data were imputed using the HRC (as the main imputation reference panel). The UK10K and 1000 Genomes phase 3 reference panels were merged by the IMPUTE4 program (https://jmarchini.org/software/) to approximately 90 million autosomal SNPs, short indels, and large structural variants in 487,442 individuals. The genotype imputation process used in the UKBB study has been thoroughly discussed in the corresponding publications[6,7,9].

1.2.3.4 GWAS workflow

GWAS was performed using a linear regression model for a quantitative trait (continuous) and a binary logistic regression for a dichotomous trait (cases and controls). In regression models, the trait was used as a response variable and the amount of coded allele of SNP as an exposure variable (SNP`s allele dosages). For example, for SNP with two alleles, A and G (G as a coded allele), allele dosages were coded as 0 for AA, 1 for AG, and 2 for GG genotypes. The regression models typically adjusted for sex, age, family structure, and population stratification (genetic principle component) to correct for the possible influences of other parameters during association testing. GWAS tested the association of coded allele with desire trait, the effect estimate of the coded allele against deviation from the null hypothesis (no association).

In the SHIP studies (SHIP-Start and SHIP-Trend), no further adjustment to the family structure was made in the association model because participants in the studies were predominantly unrelated. Other adjustments (study-specific covariates) in the association models were described in the respective publications[3,4]. GWAS was performed using the software Efficient and Parallelizable Association Container Toolbox (EPACTS) (https://genome.sph.umich.edu/wiki/EPACTS), adding the SNP's allele dosage to a linear regression model via "q.linear" test for quantitative and a logistic regression model via "b.wald" test for binary phenotypes. EPACTS was able to work with the imputed genotypes output, Variant Call Format (VCF) file, from the Michigan imputation server.

In the UKBB study, genetic association analysis was performed by the BOLT-LMM mixed model algorithm[17], an efficient mixed model for identifying genetic associations and avoiding confounding. Compared to the standard infinitesimal mixed model, BOLT-LMM requires only a small number of time iterations and accordingly increases the power to detect associations. The BOLT-LMM adapts the mixed model via modeling non-infinitesimal genetic architectures with a Bayesian mixture prior to SNP effect sizes that better accommodate both small and large effect loci.

1.2.3.5 Quality control and visualization of GWAS results

Population structure, including population stratification and cryptic relatedness, can lead to spurious associations in GWAS. Consequently, GWAS results have to be quality controlled. Typically, a Quantile-Quantile (QQ) plot (plot of the observed p-value of the meta-analysis association test versus the expected distribution under the null hypothesis of no association) is generated to discover the undetected problems in GWAS results. In addition, genomic control (GC) can then be applied to adjust test statistics at individual loci when the genomic inflation factor λ_{GC} reflects the evidence of inflation of the GWAS p-values. λ_{GC} is defined as the median of the observed chi-squared test statistic divided by the median of the corresponding chi-squared distribution with one degree of freedom. The QQ plot and λ_{GC} can help to detect and correct the possible inflation of the results in terms of the unexpectedly high number of low p-values from genotype associations with the outcome.

A standard method for visualizing the GWAS results is generating a genome-wide Manhattan plot or a detailed association plot of the specific region (regional association plot-locus zoom plot). In both cases, SNPs are displayed on the x-axis according to their position on each chromosome versus the $-\log_{10}(\mathrm{p})$ of the association on the y-axis. More significant associations show higher peaks on the y-axis.

Before meta-analysis, study-specific GWAS files were quality-controlled (GWAS-QC) based on effective sample size, imputation quality score (INFO), genotyping callrate, MAF, effect size (Beta), standard error (SE), and p-value. GC correction was applied when the λ_{GC} within the study was greater than one. GWAS-QC has been discussed in the corresponding publications[3,4].

1.2.4 Meta-analysis of GWAS results

In complex diseases, the causative common genetic variants have relatively small effect sizes, and single studies are underpowered to detect true positive associations. Meta-analysis of GWAS is a statistical technique that increases the sample size and examines more variants throughout the genome by combining the results of multiple smaller independent GWAS studies (on the same research question). Meta-analysis improves the power to detect genetic variants with small to moderate effect sizes and investigates the consistency or inconsistency (heterogeneity) of detected associations across diverse datasets and study populations.

A meta-analysis combines directly genotyped or imputed genotyped variants across studies up to several millions of common variants[18].

The fixed and random effects models are two basic approaches to meta-analysis. In the fixed effects meta-analysis model, all studies in the meta-analysis are assumed to have a common true effect size, and the combined effect is the estimate of this value. The differences in the observed effect sizes are due to the random error inherent in each study. By contrast, the random effects model assumes a distribution of true effect sizes (not one true effect size), and the combined estimate is the average distribution of effects. The differences in the observed effect sizes are due to a combination of true difference and random error. The random effects meta-analysis model needs a larger number of studies and is less used in GWAS.

I ran a trans-ethnic meta-analysis of GWAS across diverse populations using fixed effects inverse-variance weighted meta-analysis implemented by METAL[19] for the respective publication[3]. Heterogeneity between studies was assessed using I^2 statistic[20], indicating the percentage of total variation between studies due to heterogeneity rather than chance. Study-specific variant filtering and QC, followed by fixed-effects inverse-variance weighted meta-analysis, were described in the respective publication[3].

Trans-ethnic meta-analysis may increase the power to detect complex trait loci when causal variants are shared between ancestry groups. However, at these loci, heterogeneity in allelic effects between GWAS correlated with ancestry may occur for several reasons. This may occur due to differences in LD structures of the causal variant(s) between ethnic groups. Or it may be due to the interaction between causal variant(s) with different environmental risk factors that affect exposure differently across populations or with SNPs that differ in allele frequency in different ethnic groups. Finally, the quality of imputation may vary between populations depending on the reference panel used. This introduces a downward bias in allelic effect estimates within ethnic groups where genotypes are less well predicted. Trans-ethnic meta regression[21] was developed to assess the contribution of ancestry to heterogeneity in effects between GWAS. In this approach, a matrix of mean pairwise allele frequency differences between GWAS (genome-wide metrics of diversity among populations) is used to derive axes of genetic variation by multi-dimensional scaling (MDS). Allele effects of a variant across GWAS, weighted by their corresponding standard errors, are modeled in a linear regression framework with the axes of genetic variation included as covariates. To evaluate heterogeneity correlated with ancestry, I implemented the meta-regression model, including the three axes explaining the largest genetic variation, using Meta-Regression of Multi-Ethnic Genetic Association (MR-MEGA v0.1.2.25)[21] software for the respective publication[3].

1.2.5 Significance level corrections for multiple comparisons

In genome research, e.g., GWAS and gene expression data analysis, simultaneous association tests of a large number of genetic variants increase the risk of false discovery rate[22]. For multiple hypothesis testing, the family-wise error rate ($FWER$) is the probability of making at least one type I error for the family of N independent tests corresponding to N (null) hypotheses $\boldsymbol{H} = (H_1, H_2, \dots, H_N)$[22,23]. To avoid many false positives, the significance threshold must be lowered to control the $FWER$ at a significant level of α ($FWER \leq \alpha$). The $FWER$ can be written as equation (1), where $\alpha[PT]$ is the probability of making a false discovery rate for a single test.

$$FWER = 1 - (1 - \alpha[PT])^N \qquad (1)$$

Equation (1) can be rewritten by equation (2) to find $\alpha[PT]$ when $FWER$ is kept at the fixed α level.

$$\alpha[PT] = 1 - (1 - \alpha)^{\frac{1}{N}} \qquad (2)$$

Equation (2), called the Šidàk correction[24], illustrates that the value of $\alpha[PT]$ must be adjusted to control the $FWER$ at level α. Let p_j donates the p-value associated with the hypothesis H_j $(1 \leq j \leq N)$. In the Šidàk correction if $p_j \leq 1 - (1 - \alpha)^{\frac{1}{N}}$, then reject H_j.

Using the first term of a Taylor expansion of the Šidàk equation, a simpler approximation known as Bonferroni[25] was derived, calculated by $\alpha[PT] \approx \frac{\alpha}{N}$. In the Bonferroni correction if $p_j \leq \frac{\alpha}{N}$, then reject H_j. The Šidàk and Bonferroni corrections are widely used to control $FWER$ for multiple hypothesis testing. For instance, in GWAS, a fixed established genome-wide significance level of $\alpha = 5 \times 10^{-8}$ is frequently applied to determine the association between a common genetic variant and a trait of interest. However, they have limited statistical power and become very conservative when the number of tests increases or when the tests are not independent[26,27]. The power of multiple testing corrections can be increased by using weighted p-values[28]. To this end, Kang et al. (2009)[27] proposed a weighted Šidák correction by incorporating a set of nonnegative weights $\boldsymbol{w} = (w_1, w_2, \dots, w_N)$ specified for N independent tests associated with $\boldsymbol{H} = (H_1, H_2, \dots, H_N)$, respectively into the Šidák correction. The weighted Šidák correction assigns specific $\alpha[PT_j]$ to every single test (j) by equation (3) while controlling the $FWER$ at level α.

$$\alpha[PT_j] = 1 - (1 - \alpha)^{\frac{w_j}{N}}, \quad \text{where } \frac{1}{N}\sum_{j=1}^{N} w_j = 1, \qquad j = 1,2,\dots,N \qquad (3)$$

In the weighted Šidák correction if $p_j \leq 1 - (1 - \alpha)^{\frac{w_j}{N}}$, then reject H_j.

The weights can be determined by prior available information. For example, in GWAS, LD structure can be used as prior information to estimate the optimal weights in multiple testing frameworks. However, how to use prior information to estimate the optimal weight is an open problem[28].

1.2.6 Conditional analysis of GWAS summary statistics

Despite the success of GWAS in identifying thousands of genetic variants associated with various diseases and traits, interpreting the discovered variants remains a challenging task. Only a handful of the GWAS-associated variants are true causal due to extensive LD structure. In association studies, causal variants are responsible for association signals and have a biological effect on a disease or trait. The LD structure creates both opportunities and difficulties in gene mapping. On the one hand, GWAS has been greatly facilitated by using knowledge of LD structure to predict variation in the genome by genotyping only a small fraction of polymorphic sites. On the other hand, it is difficult to identify true causal variants in a set of sites in strong LD by using only association data. Detection of causal variants is complicated because the index SNP (the SNP with the smallest p-value) at the given locus (the ±500kb-region around index SNP) may not be casual but instead be in high LD with an unknown functional variant. In addition to the LD structure, the presence of multiple genetic variants at the same locus (allelic heterogeneity) is a common characteristic of complex traits. Consequently, the total phenotypic variance explained by genetic variation might be underestimated under the simplifying assumption that each GWAS-associated locus harbors exactly one causal variant or if only index SNPs were considered causal variants. To address these issues, conditional analysis was developed to detect multiple conditionally independent association signals at GWAS loci. Conditionally independent association signals define as signals that remain or become significantly associated after conditioning on other nearby signals, which are more significant. The conditional analysis is an interactive process starting with an index SNP at a locus. It is performed by including the allele dosages of index SNP as a covariate in an association model. The process is followed by the stepwise procedure of selecting additional significant SNPs, one by one, according to the conditional p-values. After the first iteration of conditional analysis (conditioning on the index SNP at a locus), the marginal statistics of all remaining variants are re-computed. A locus is considered to have conditionally independent signals when the conditional p-value for at least one of the variants is less than a predefined threshold. The predefined threshold is referred to as the stopping threshold. Usually, the established genome-wide significance level of $\alpha = 5 \times 10^{-8}$ is applied as a significance threshold. The conditional analyses are performed by conditioning on index SNP and all SNPs selected in previous steps until no additional multiple independent signals are found at a locus. This standard method can be applied to either summary statistics or individual-level data sets.

Approximate conditional and joint genome-wide association analysis[29] implemented in GCTA[30] software (GCTA COJO Slct algorithm) performs the conditional analysis while utilizing the GWAS summary statistics. The method does not depend on genotype and phenotype data at the individual level, except for an LD reference sample with individual-level genotype data. The method can use summary-level statistics from a meta-analysis of GWAS and estimate the LD from a reference sample. To estimate the unbiased LD correlation, the reference sample should be from either one of the participating studies of the meta-analysis or an ancestry-matched population with a large sample size > 2,000[29].

1.3 Additional OMICS data as a basis for locus discovery

Epigenetic mechanisms involve modifications to genomic DNA, affecting transcript abundance and influencing common diseases and complex traits. They may not be accounted for in SNP-based association studies, hence alternative approaches are needed to unravel the biology underlying diseases. The technique includes epigenetic changes like DNA methylation, which can be applied by the epigenome-wide association study (EWAS) to increase our understanding of the role of methylation in many diseases.

EWAS investigate the association between a phenotype of interest and genome-wide epigenetic variants, most commonly DNA methylation at CpGs. DNA methylation can occur at the DNA sites, where a C nucleotide is followed by a G nucleotide, cytosine-phosphate-guanine (CpG). They regulate gene expression through the presence or absence of a methyl group on CpG dinucleotides. The methylation level at each site was measured and modeled as the dependent variables with phenotype being either continuous or binary variable.

In SHIP-Trend study, DNA methylation level assessed from blood samples of 256 participants via the Illumina Human Methylation Bead Chips, which covers 850,000 DNA methylation sites per array. The methylation level at each site was calculated as β-value. β-value is the estimate of methylation level using the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities and 100 array probe intensity values).

I performed study-specific EWAS for SHIP-Trend for the respective populations[48-50].

1.4 Association with gene expression and colocalization

1.4.1 Expression quantitative trait loci (eQTL) study

In recent years, much effort has been devoted to the analysis of genome function, particularly in the context of genome variation. One of the most important directions is the expression quantitative trait loci (eQTL) study[31], which uses gene expression measurements derived from microarray[32] or RNA sequencing[33] studies as an outcome trait for the GWAS design and identifies variants influence the expression level of genes in different tissues and cell types (Figure 2).

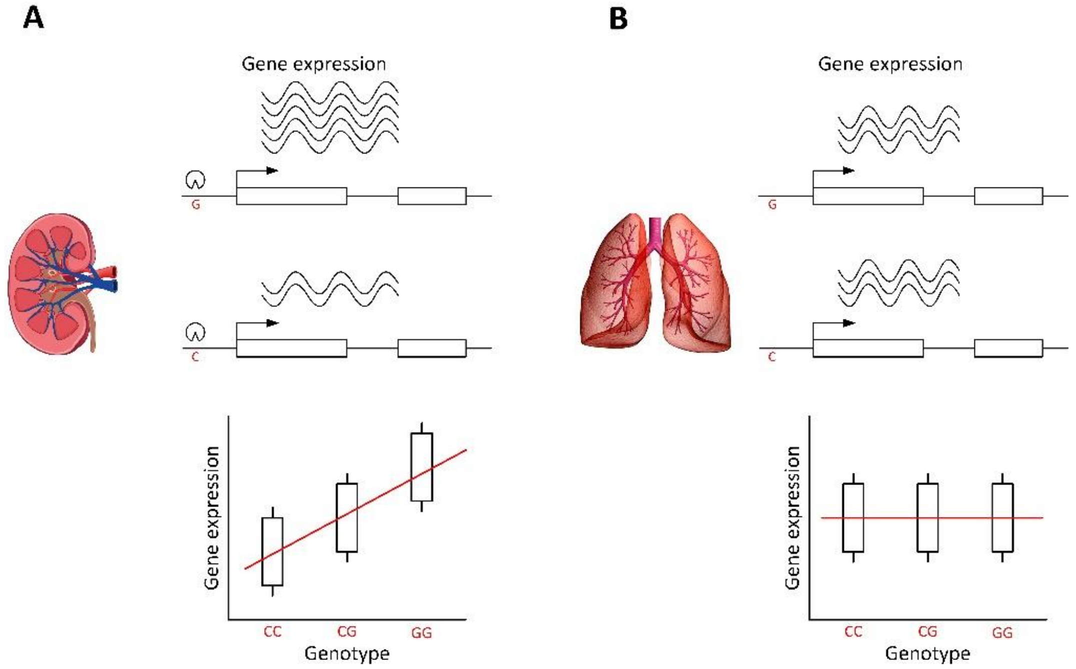


Figure 2. Illustrates the statistical associations between genotype and gene expression levels in different tissues. **A** Shows a significant association between genotype and gene expression level when mRNA levels are significantly higher in one allele than the other allele, while **B** shows no difference in mRNA levels between genotype groups.

The eQTLs are typically classified according to their relative locations of the eQTLs and the gene(s) they affect, and the type of mechanism by which they affect gene expression. eQTLs can be divided into *cis* and *trans*-eQTL. *Cis*-eQTL has local effects, which commonly refers to the genetic variant within 1-Mb on either side of a gene's transcriptional start site (TSS) directly affects the expression of its associated local gene. While *trans*-eQTL refers to the genetic variant with distant effects located at least 5-Mb downstream or upstream of the TSS or on a different chromosome. In *Trans*-eQTL, gene expression is affected through possibly complex effects involving the genetic variant (Figure 3).

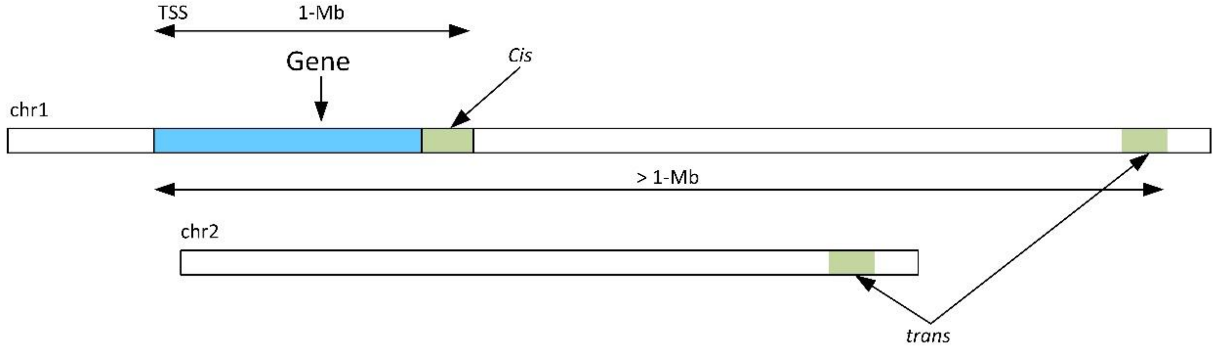


Figure 3. Illustrates *cis* and *trans*-eQTL testing. *Cis*-eQTL tests all SNPs within 1-Mb on either side of a gene's transcriptional start site (TSS) and *trans*-eQTL tests all other SNPs located in higher distance commonly at least 5-Mb downstream or upstream of the TSS or on a different chromosome.

The Genotype-Tissue Expression (GTEx) project (https://gtexportal.org) is a comprehensive public resource to study tissue-specific gene expression and regulation. GTEx has generated gene expression levels of different tissues and the eQTL dataset, which is publicly available and has been used extensively to help interpret GWAS signals from complex traits. Samples for GTEx v8 were collected from 54 non-diseased tissue sites across nearly 1000 individuals. However, the human kidney tissues have been poorly covered by the GTEx study, and only the kidney cortex with a small sample size is included in this dataset. To overcome this limitation, kidney tissue can be investigated using a *cis*-eQTL dataset from microdissected human glomerular and tubulointerstitial kidney portions from 187 individuals from the NEPTUNE study[34].

1.4.2 Colocalization- integration of GWAS and eQTLs

GWAS has identified many genetic variants in non-coding regions of the genome. These variants may alter the individual's disease risk through their effect on gene expression in different tissues. One approach to understanding the biological basis of these GWAS risk loci can be achieved by integrating GWAS and eQTLs to assess whether two association signals are consistent with a shared causal variant. The abundance of eQTLs in the human genome and across different tissues makes an accidental overlap between eQTLs and GWAS signals very likely. Therefore, formal statistical tests must be used to make inferences about causality. Colocalization analysis (frequentist[35] and Bayesian[36] approaches) has emerged as a powerful tool to combine GWAS and eQTLs to estimate the relation between gene expression of nearby genes and GWAS association signals.

The Bayesian approach of the colocalization method makes the "one causal variant in a locus" (OCV) assumption for each trait (GWAS and eQTL). This specific assumption outlines five different possible hypotheses within each region. One: there are no causal variants for either trait (H0). Two: there is only one causal eQTL variant but no causal GWAS variant (H1). Three: there is only one causal GWAS variant but no causal eQTL variant (H2). Four: there are different causal SNPs for both eQTL and GWAS (H3). Five: there is a colocalized signal (H4). Four hypotheses are shown in Figure 4.

The corresponding posterior probability is calculated for each hypothesis by considering all latent association states from GWAS and eQTL data using Bayesian model averaging (BMA). Colocalization within each region is quantified by the posterior probability of H4 (PP). A variant was defined as a co-localized signal (same causal variant underlying both the GWAS and eQTL association) if a variant's posterior probability (PP) was greater than 80%. The colocalization method has been described in detail by Giambartolomei, C. et al. (2014)[36].

In addition to functional characterization of GWAS risk loci, colocalization provides a systematic approach for correlating gene expression levels (not measured directly in the sample of interest) with a trait or disease where the data can be obtained from extrinsic individuals. This means individuals can have either measured expression levels or assessed disease status. The link between these samples is made via genetics. Information on genetic variation must be available from all individuals included in the analysis.

I ran colocalization by coloc.fast function from the R package "gtx" version 2.1.6 (https://github.com/tobyjohnson/gtx) which provides an adaptation of Giambartolomei's colocalization method[36]. Details were described in the respective publication.

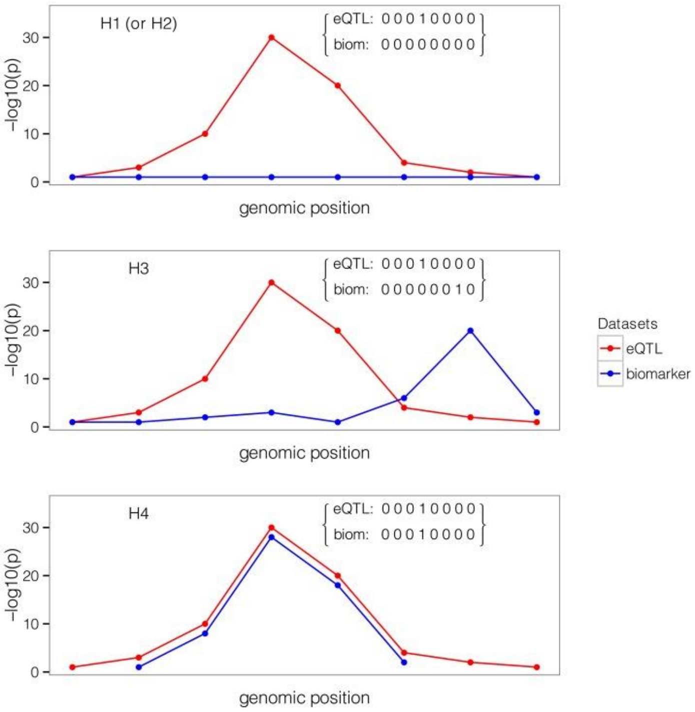


Figure 4. Illustrates four different hypotheses within a region. Variant causality was represented with a binary vector (0,1) of length n = 8 (the number of common variants in a region) for each trait (biom (GWAS) and eQTL). A value of 1 means that the variant is causally involved in the disease, and 0 means that it is not. The first plot shows the case where only one trait (eQTL) shows an association. The second plot shows that both traits show an association, but the causal SNP in GWAS dataset is different compared to the eQTL dataset. The third plot shows that the fourth SNP is the underlying causal variant for both GWAS and eQTL. Reprinted from "Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics" Giambartolomei, C. et al. (2014), PLoS Genet. 10, e1004383. Copyright: 2014 Giambartolomei et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

1.5 Quasi-adaptive method

The conditional analysis has successfully identified multiple conditionally independent SNPs at loci with allelic heterogeneity in GWAS studies. Generally, the established genome-wide significance level (e.g., $\alpha = 5 \times 10^{-8}$) is used as the significance threshold in conditional analysis, which is also the significance level for the primary GWAS. Unlike the tests for the primary association, conditional tests are not applied genome-wide but are restricted to the specific genomic regions surrounding the GWAS index signals. Consequently, applying the genome-wide significance level in the conditional analysis is too strict and implies an unnecessary loss of power. To address this issue, we developed a quasi-adaptive method to establish significance thresholds and evaluate the conditional independence of secondary signals in conditional analyses. The method prioritizes SNPs and applies less restrictive significance levels to those with higher priority, while maintaining an $FWER$ level at $\alpha = 0.05$ to control type I error rate.

Our method is based on independent genome-wide significant loci from primary GWAS. The number of candidate SNPs from N_2 independent loci is represented by m_2, which was referred to as simply "m" and "N" in the respective publication[24]. Within each locus, an SNP with the smallest p-value is designated as the index SNP, while the remaining SNPs are considered candidate SNPs. The quasi-adaptive method, utilizes the weighted Šidák correction (3) to take into account the LD structure (pairwise correlation (r^2) and chromosomal distance between the index SNP and candidate SNPs (d)) as the prior information. The method estimates optimal weights to prioritize SNPs and assigns an SNP-specific α-thresholds to candidate SNPs in conditional analysis.

The pre-weight ($w_{r^2{}_i}$) based on r^2 with optimal $r^2 = 0.3$ and a pre-weight (w_{d_i}) based on d which down-weighted SNPs at higher distance step-wise-strong are assigned to a candidate SNP(i), ($1 \le i \le m_2$) as:

$$w_{r^2{}_i} = \frac{1 - |r_i{}^2 - 0.3| - 0.3}{1 - 0.3},$$

$$w_{d_i} = \begin{cases} 1 & if\ 0 < d \le 1\text{Kb} \\ 0.5 & if\ 1\text{Kb} < d \le 10\text{Kb} \\ 0.25 & if\ 10\text{Kb} < d \le 50\text{Kb} \\ 0.125 & if\ 50\text{Kb} < d \le 100\text{Kb} \\ 0.0625 & if\ 100\text{Kb} < d \le 500\text{Kb} \end{cases}$$

The pre-weights $w_{r^2{}_i}$ and w_{d_i} are combined (with more emphasis on d than on r^2) by the geometric mean $w_i = (w_{d_i}^k \times w_{r^2{}_i})^{\frac{1}{k+1}}$, with $k = 5$, to assign an optimal weight $W_i = \frac{w_i \times m_2}{\sum_{i=1}^{m_2} w_i}$ to SNP(i).

The quasi-adaptive method is applied on N_2 loci, by incorporating W_i into the weighted Šidák correction. The method distributes type I error rate (α) among m_2 candidate SNPs, and assigns the SNP-specific α-thresholds to SNP(i) by $G_i(\alpha, r^2, d)$ as follows:

$$G_i(\alpha, r^2, d) = 1 - (1 - \alpha)^{\frac{W_i}{m_2}}, \qquad i = 1,2,\dots,m_2. \qquad (4)$$

SNP(i) is considered a secondary signal if the conditional p-value is smaller than $G_i(\alpha, r^2, d)$. The regional association plot in Figure 5 illustrates the secondary signal exclusively identified by the quasi-adaptive method, with conditional p-value smaller than $G_i(\alpha, r^2, d)$ and greater than 5×10^{-8}.

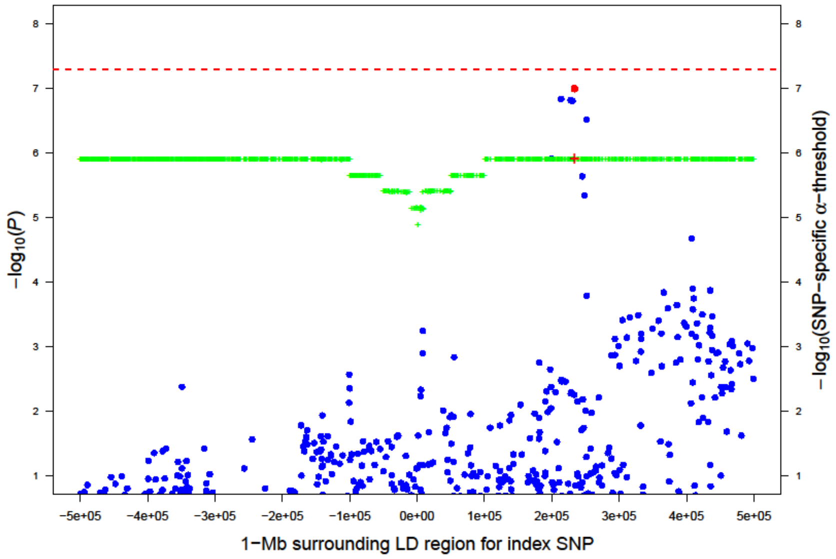


Figure 5. Regional association plot with secondary signal (red dot) detected exclusively by the quasi-adaptive method. Y-axis is $-log_{10}$ of conditional p-value (blue dot) for candidate SNPs from 1-Mb surrounding LD region of an index SNP (x-axis). Forth axis is $-log_{10}$ of the SNP-specific α-thresholds assigned by the quasi-adaptive method (green plus sign) to the candidate SNPs. red dashed line is $-log_{10}\,(5 \times 10^{-8})$. The secondary signal was found by the quasi-adaptive method is shown in red dot with corresponding $-log_{10}$ of the SNP-specific α-threshold assigned by the quasi-adaptive method (red plus sign).

Details of the quasi-adaptive method, the simulation study, and the power analysis were described in the respective publication[37].

1.5.1 Improved quasi-adaptive method

The quasi-adaptive method was developed to determine one independent signal (secondary signal) at each locus. We improved the method to identify plausible multiple independent signals at each locus (a tertiary signal, a signal of 4th, a signal of 5th, and beyond).

To detect independent tertiary signals, only loci with confirmed secondary signals (confirmed according to the quasi-adaptive method) were considered. We proceeded according to the idea of the main paper[37] but performed conditional analysis by adjusting for the primary index SNP and confirmed secondary signal for each locus. The number of loci with confirmed secondary signals is represented by N_3 and m_3 denotes the number of candidate SNPs (excluding index SNPs and secondary signals) from N_3 loci. Our method was applied on N_3 loci following the schema described above and the SNP-specific α-thresholds were assigned to each SNP(i) using equation (5).

$$G_i(\alpha,\, r^2, d\,) = \,1 - (1-\alpha)^{\frac{W_i}{m_3}}, \quad W_i = \frac{w_i \times m_3}{\sum_{i=1}^{m_3} w_i}, \qquad i = 1,2,\dots,m_3. \qquad (5)$$

The improved method is an iterative process that is subsequently performed to detect higher-order independent signals (applied to loci with confirmed independent signals from the previous steps) until no additional independent signals are found. Finding higher-order independent signals keeps the $FWER$ at $\alpha = 0.05$ because only the number of candidate SNPs and the LD structure have to be taken into account, where the LD structure does not change by analyzing higher-order independent signals. Details were described in the respective publication[38].

1.5.2 Conditional analysis in quasi-adaptive method

The conditional analysis in the quasi-adaptive method proceeds according to the standard conditional analysis approach. It is an interactive process that starts by adjusting for the primary index SNP at each locus. For each candidate SNP, the assigned SNP-specific α-level compared to the p-value from the conditional analysis, and one secondary signal with the smallest conditional p-value lower than the corresponding SNP-specific α-level is selected at a locus. When a secondary signal is identified at a locus, conditional analysis is performed by adjusting for the primary index SNP and the secondary signal to find the tertiary. Conditional analyses are performed by adjusting for the primary index SNP and any independent SNPs selected in previous steps by the quasi-adaptive method until no additional independent signals are found at a locus. Note that the stopping thresholds are the assigned SNP-specific α-levels by the quasi-adaptive method for each iteration. Approximate conditional analysis implemented in GCTA (GCTA COJO-cond algorithm) performs the conditional analysis while utilizing the summary statistics.

2 Summary

2.1 Summary of the thesis

This thesis focuses on identifying genetic factors associated with human kidney disease progression, with three articles presented. Article I describes the identification of loci associated with UACR through trans-ethnic, European-ancestry-specific, and diabetes-specific meta-analyses. An approximate conditional analysis was performed to identify additional independent UACR-associated variants within identified loci. The genome-wide significance level of $\alpha = 5 \times 10^{-8}$ is used for both primary GWAS association and conditional analyses. However, unlike primary association tests, conditional tests are limited to specific genomic regions surrounding primary GWAS index signals rather than being applied on a genome-wide scale.

In article II, we hypothesized that the application of $\alpha = 5 \times 10^{-8}$ is overly strict and results in a loss of power. To address this issue, we developed a quasi-adaptive method within a weighted hypothesis testing framework. This method exploits the type I error ($\alpha = 0.05$) by providing less conservative SNP specific α-thresholds to select secondary signals in conditional analysis. Through simulation studies and power analyses, we demonstrate that the quasi-adaptive method outperforms the established criterion $\alpha = 5 \times 10^{-8}$ as well as the equal weighting scheme (the Sidak-correction). Furthermore, our method performs well when applied to real datasets and can potentially reveal previously undetected secondary signals in existing data.

In article III, we extended our quasi-adaptive method to identify plausible multiple independent signals at each locus (a secondary signal, a tertiary signal, a signal of 4th, and beyond) and applied it to the publically available GWAS meta-analysis to detect additional multiple independent eGFR-associated signals. The improved quasi-adaptive method successfully identified additional novel replicated independent SNPs that would have gone undetected by applying too conservative genome-wide significance level of $\alpha = 5 \times 10^{-8}$. Colocalization analysis based on the novel independent signals identified potentially functional genes across the kidney and other tissues.

Overall, these articles contribute to the understanding of genetic factors associated with human kidney disease progression and provide novel methods for identifying secondary and multiple independent signals in conditional GWAS analyses.

2.2 Summary of the publications

2.2.1 Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria (Article I)

Higher levels of the urinary albumin-to-creatinine ratio (UACR) as a measure of kidney damage are associated with adverse clinical outcomes such as end-stage kidney disease, cardiovascular disease (CVD), and mortality. Despite the heritability of UACR, the underlying genetic mechanisms are not well understood, and identifying genetic loci for UACR through GWAS has been challenging. To this end, the CKDgen consortium (round 4) was established to increase statistical power and identify novel genetic markers for UACR through a meta-analysis of multiple studies with large sample sizes.

Data based on 564,257 individuals from 54 studies were combined in a trans-ethnic meta-analysis of UACR, including 547,361 of European ancestry (EA), 6,795 African Americans ancestry (AA), 6,324 of East Asian ancestry, 2,335 of South Asian ancestry, and 1,442 Hispanics. The GWAS results of 3,199 individuals from SHIP-START and 985 individuals from SHIP-Trend were included in the discovery stage of the meta-analysis. 59 UACR-associated loci were identified through the trans-ethnic meta-analysis, out of which 27 were considered known and 32 were considered novel (Figure 6).

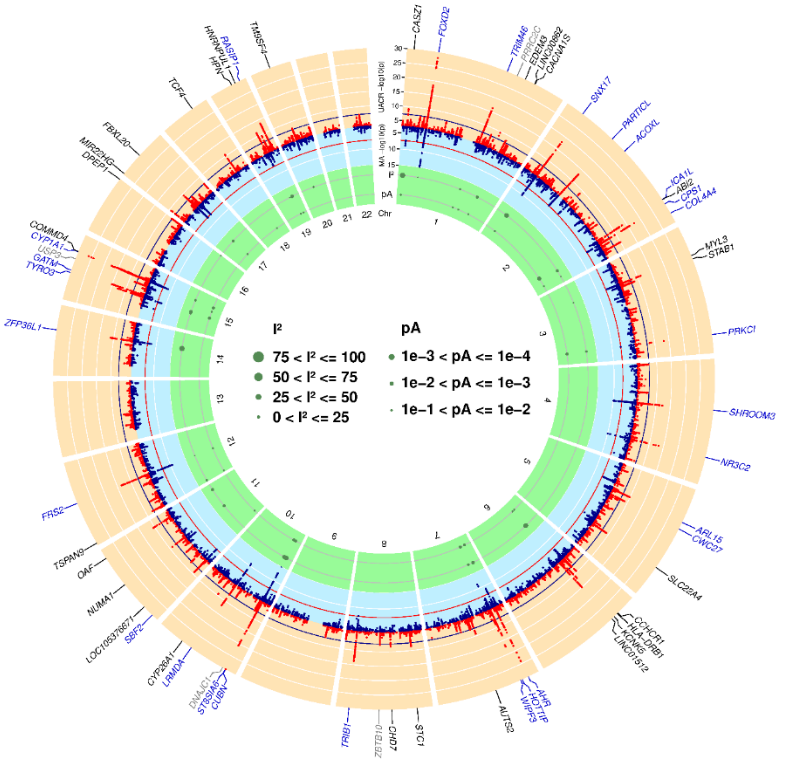


Figure 6. Genome-wide association results. The circos plot provides an overview of the association results: Red band: $-log_{10}(p)$ for association in the trans ethnic meta-analysis of urinary albumin-to-creatinine ratio (UACR), ordered by chromosomal position. The blue line indicates genome-wide significance (p = 5 × 10^{-8}). Black gene labels indicate novel loci, blue labels indicate known loci (known index SNP within ± 500 kb region of current index SNP), gray labels indicate loci not associated with UACR at the nominal significance level ($p \geq 0.05$) in the 53 CKDGen cohorts without UKBB. Blue band: $-log_{10}(p)$ for association with microalbuminuria (MA), ordered by chromosomal position. The red line indicates genome-wide significance (p =5 × 10^{-8}). Green band: measures of heterogeneity related to the UACR-associated index SNPs, where the dot sizes are proportional to two measures of heterogeneity, I^2 and the $-log_{10}(p)$ for heterogeneity attributed to ancestry (pA)

As secondary analyses, an ancestry-specific meta-analysis was performed for EA and 61 loci were identified, 56 of which overlapped with those from the primary trans-ethnic meta-analysis. Additionally, a diabetes-specific meta-analysis was performed on 51,541 individuals with diabetes, including 2,345 individuals from SHIP-START, and identified 8 loci, 4 of which were not detected in the primary meta-analysis (*KAZN*, *MIR4432HG-BCL11A*, *FOXP2*, and *CDH2*). In total, 68 UACR-associated genetic loci were identified including 34 novel loci through primary trans-ethnic GWAS meta-analysis and secondary analyses. The workflow of the project is illustrated in Figure 7.

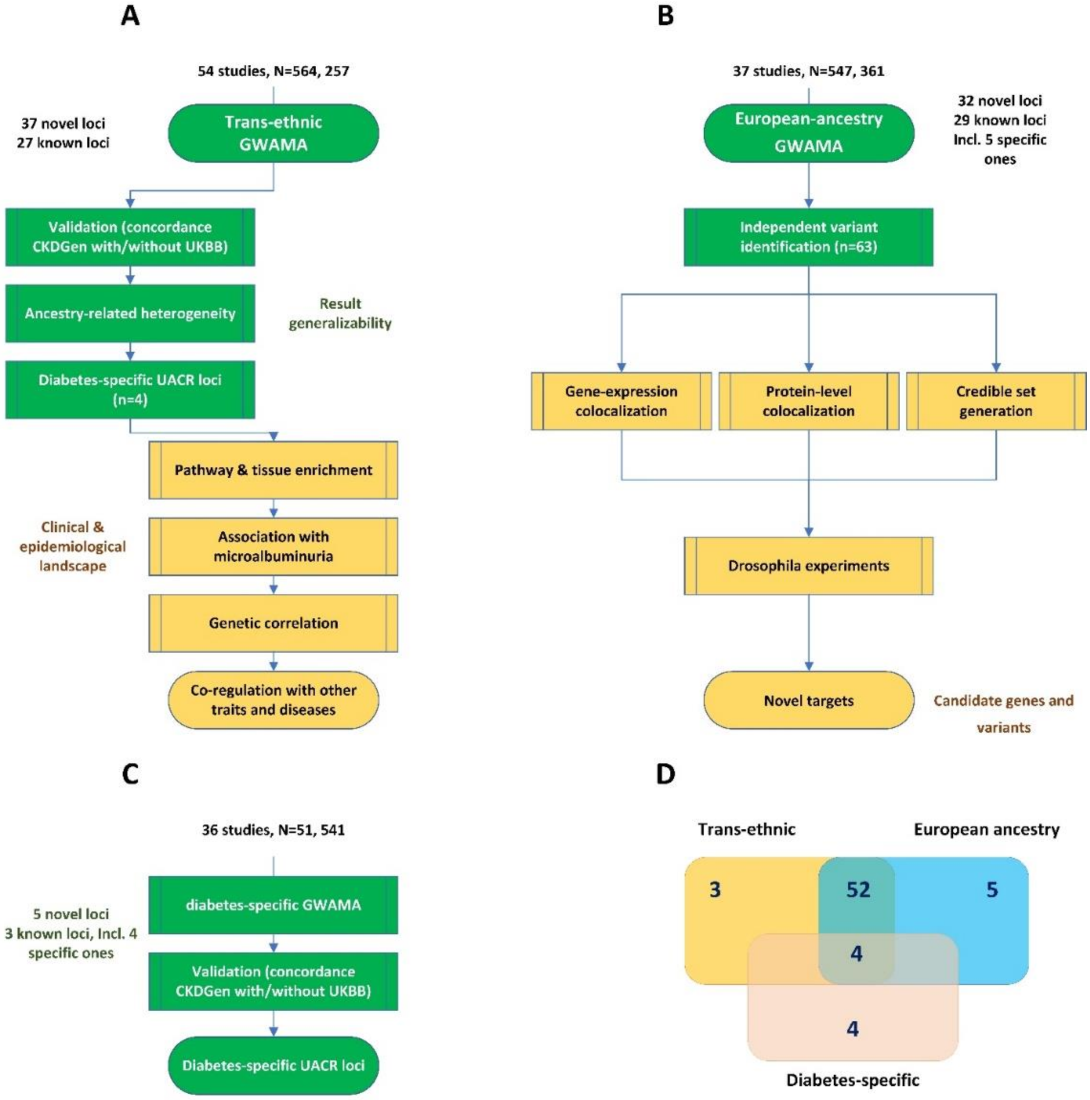

A
54 studies, N=564, 257
37 novel loci
27 known loci
Trans-ethnic GWAMA
Validation (concordance CKDGen with/without UKBB)
Ancestry-related heterogeneity
Result generalizability
Diabetes-specific UACR loci (n=4)
Pathway & tissue enrichment
Clinical & epidemiological landscape
Association with microalbuminuria
Genetic correlation
Co-regulation with other traits and diseases
B
37 studies, N=547, 361
European-ancestry GWAMA
32 novel loci
29 known loci
Incl. 5 specific ones
Independent variant identification (n=63)
Gene-expression colocalization
Protein-level colocalization
Credible set generation
Drosophila experiments
Novel targets
Candidate genes and variants
C
36 studies, N=51, 541
5 novel loci
3 known loci, Incl. 4 specific ones
diabetes-specific GWAMA
Validation (concordance CKDGen with/without UKBB)
Diabetes-specific UACR loci
D
Trans-ethnic
European ancestry
3
52
5
4
4
Diabetes-specific

Figure 7. Overview of the analysis workflow for trans-ethnic (a), European ancestry-specific (b), and diabetes-specific (c) genome-wide association meta-analyses (GWAMA). The Venn diagram in panel (d) shows the number of overlapping loci between the different GWAMA. The contribution of the author of the thesis, Sahar Ghasemi, is highlighted in green throughout the workflow, showcasing the specific responsibilities and tasks the author carried out.

Collaborators performed a series of analyses to determine the functional relevance of the identified loci. This included genetic correlation analyses and risk score associations in an independent electronic medical records database of 192,868 EA participants. The analysis revealed significant association between the identified genetic loci and several clinically relevant conditions such as proteinuria, hyperlipidemia, gout, and hypertension. Functional enrichment analyses, statistical fine-mapping and integrative trans-Omics analyses, including gene expression in 47 human tissues and plasma protein levels, were conducted to prioritize and characterize several of the revealed trait-associated loci in the EA. The analyses identified nine genes that colocalized with UACR-associated loci in the kidney tissues including *TGFB1*, *MUC1*, *PRKCI*, and *OAF*). These findings allow for a linking of UACR associations to alter plasma *OAF* concentrations. Several of the disease-associated loci were characterized for their likely causal variants or for functional effects using cell culture and animal models. For example, in vivo analyses of Drosophila orthologs supported a role of *OAF* in tubular protein endocytosis, and *PRKCI* in slit diaphragm formation potentially reflecting changes kidney structure.

2.2.2 Assessment of significance of conditionally independent GWAS signals (Article II)

As outlined in chapter 1.5, we developed the quasi-adaptive method with the goal of increasing power to detect independent secondary signals in conditional analyses. The method utilizes the LD structure (r^2, d) as prior knowledge to determine optimal weights for prioritizing SNPs and assign SNP-specific α-thresholds to candidate SNPs in conditional analysis. We have established a series of priority functions that are determined by a combination of pre-weights based on r^2 and d. It is important to note that r^2 and d operate to some extent in opposite directions. Although a high r^2 value may increase the biological prior (via haplotype effects), it reduces statistical power by decreasing the amount of independent information. To investigate the counter-running effect of r^2 and d, we evaluated several alternatives for the pre-weights on the r^2-component and d-component in a series of priority functions. The specifics of the SNP weighting schemes and the development of priority functions are outlined in the corresponding publication[37].

The validity of the method was assessed by evaluating the deviation of $FWER$ from desired level $\alpha = 0.05$ via simulation analysis based on imputed genotypes from the SHIP study. Under the null hypothesis of no secondary signals, the simulation analyses confirmed appropriate empirical $FWER$ for defined priority functions.

Power simulations were set up under the alternative hypothesis (pre-defined secondary signals) for three scenarios, scenario A (selecting secondary signal by random), scenario B (selecting secondary signal conditionally at random based on d, step-wide-moderate), and scenario C (selecting secondary signal conditionally at random based on d, step-wide-strong) to evaluate priority functions' power in detecting secondary signals. The quasi-adaptive method ($G5$-function) had the overall best power (mean power of 0.7289 and median power of 0.7305) among other priority functions (Figure 8). It showed improved power by 22 percentage points (median) over the established criterion $\alpha = 5 \times 10^{-8}$ ($G14$-function) as well as four percentage points (median) over the Sidak-correction ($G13$-function). It also showed the best results in application to real data sets.

As a proof of concept, the quasi-adaptive method was applied to GWAS on free thyroxine (FT4)[39], inflammatory bowel disease (IBD)[40], and human height[41]. Our algorithm revealed five secondary signals in GWAS on FT4, five in the IBD, and 19 in human height that would have gone undetected using the established genome-wide significance level of $\alpha = 5 \times 10^{-8}$.

We analyzed the impact of using different LD reference samples on our proposed method by using the SHIP and UKBB reference samples, which consist of 4070 and 13,558 individuals respectively, in our analysis of the FT4 and IBD GWAS examples. Secondary signals obtained with the SHIP reference sample showed overall good agreement with those obtained using UKBB. The results somewhat depend on the reference sample, which determines SNP availability and the LD structure estimated from it. In this context, we recommend using the reference LD panel from the same population the study data comes from and beyond a sample size of 5,000 for additional accuracy, as also proposed by Yang et al. (2011)[29].

The quasi-adaptive method is easy to use, operates directly with typically already existing GWAMA results, and makes use of existing analysis software for conditional analysis (GCTA). The method has the potential to reveal previously undetected secondary signals in already available data and to uncover plausible underlying gene mechanisms.

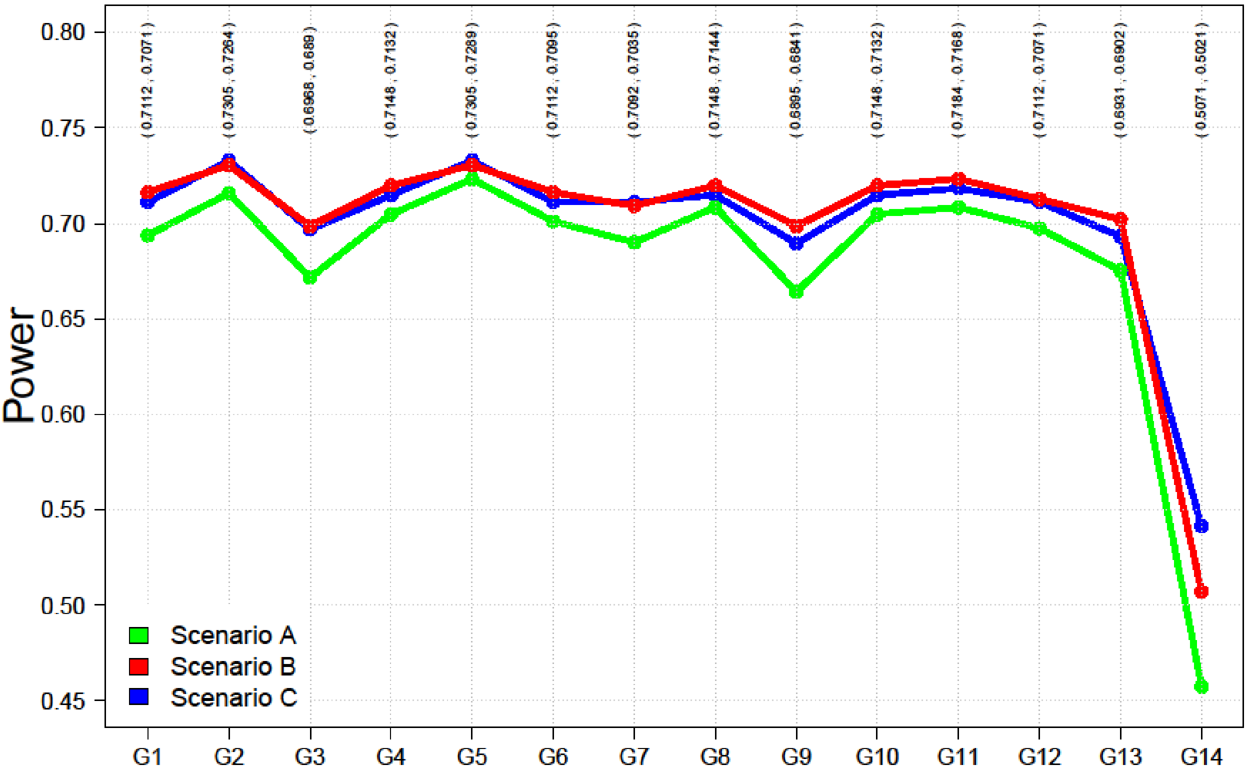


Figure 8: Power simulation analysis results across scenario A (selecting secondary signal by random), B (selecting secondary signal conditionally at random based on distance, step-wide-moderate), and C (selecting secondary signal conditionally at random based on distance, step-wide-strong) for 14 priority functions (G-functions). The numbers above show the median and the mean of power analyses across three scenarios for each G-function, respectively.

2.2.3 Discovery of novel eGFR-associated multiple independent signals using quasi-adaptive method (Article III)

The quasi-adaptive method was improved (as described in 1.5.1) and applied to the publically available GWAS meta-analysis of eGFR[4] from the CKDGen consortium to detect additional associated multiple independent signals. The quasi-adaptive method identified 87 multiple independent signals (without index SNPs from primary GWAS[4]), of which 27 were novel. The approach detected 60 known loci, of which 54 loci comprised the same independent signals identified in the previous GWAS and 6 loci with independent signals in high *LD* with the identified independent signals from the aforementioned GWAS. 19 of 27 new independent SNPs were subsequently replicated in an independent data set, UK Biobank genotype data among EA individuals (n = 408,608). These signals included 5 secondary signals, 5 tertiary signals, 6 signals of 4th, 2 signals of 5th, and one signal of 6th (Figure 9). These results would have gone undetected by conditional analysis applying the commonly used but too conservative genome-wide significance level of $\alpha = 5 \times 10^{-8}$.

Of note, the new independent signals rs3904600, rs13227214, rs81205, rs2075251, rs2695565, and rs6951593 (identified by the quasi-adaptive method based on meta-analysis of previous GWAS of eGFR[4]) showed smaller p-values in their unconditional eGFR-association analysis within the UKBB compared to their corresponding index SNP. Figure 10 shows regional association plot for index SNP rs3757387 and tertiary signal rs13227214 identified by the quasi-adaptive method.

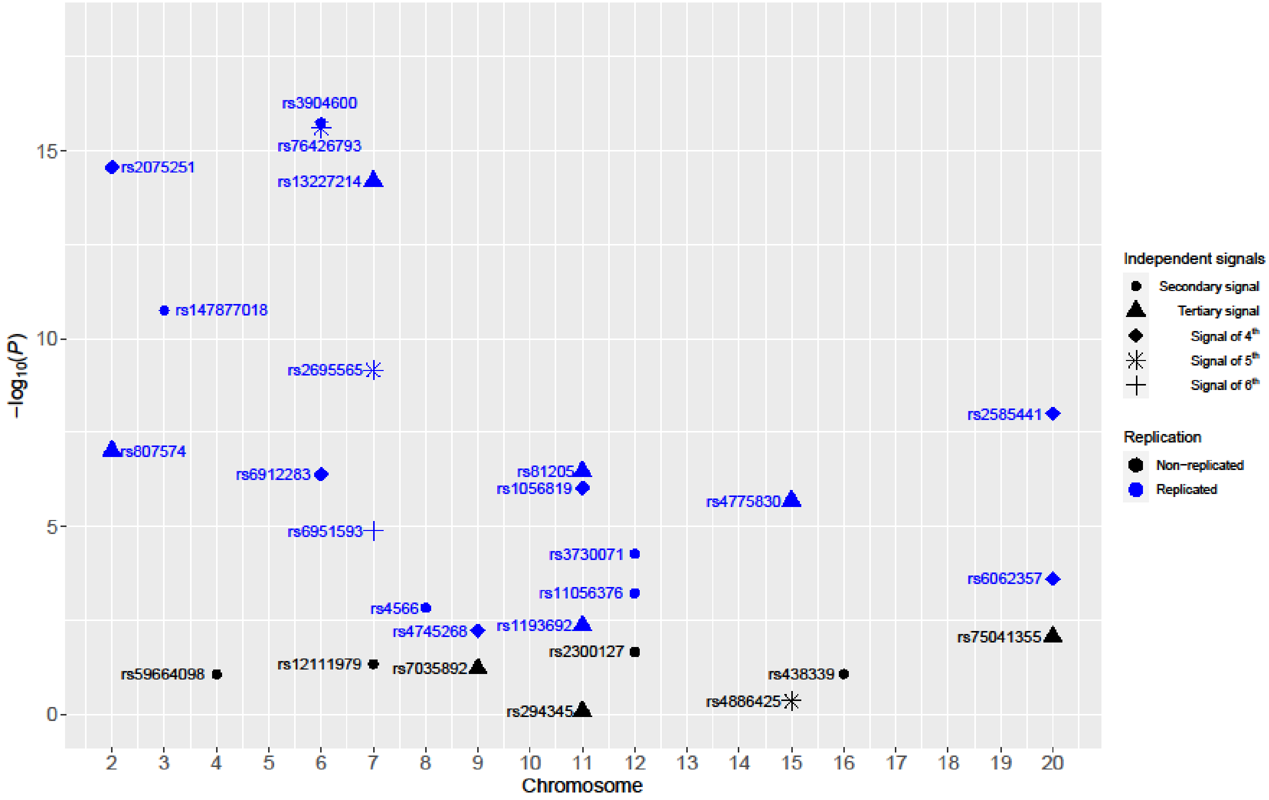


Figure 9. Replication of eGFR-associated multiple independent signals identified by the quasi-adaptive method using the UK Biobank (UKBB) genotype data among European-ancestry individuals. The x-axis shows the chromosome number, and the y-axis is the $-log_{10}(P)$ of the conditional GWAS of eGFR. Color coding reflects evidence of replication, which is coded as replicated (blue) and non-replicated (black). Different shapes showed multiple independent signals.

Colocalization based on the conditional and unconditional eGFR association results with *cis*-eQTL across 49 human tissues included in the GTEx project v8 release[42] as well as the microdissected human glomerular and tubulointerstitial kidney portions from 187 individuals from the NEPTUNE study[43] were conducted to characterize 17 known eGFR-associated index SNPs and 19 novel independent signals. Colocalization identified two potentially causal genes across kidney tissues: *TSPAN33* and *TFDP2*, and three candidate genes across other tissues: *SLC22A2*, *LRP2*, and *CDKN1C*. These results were not identified in the original report of eGFR[4]. Considering these signals in colocalization analyses can increase the precision of revealing potentially functional genes of GWAS loci.

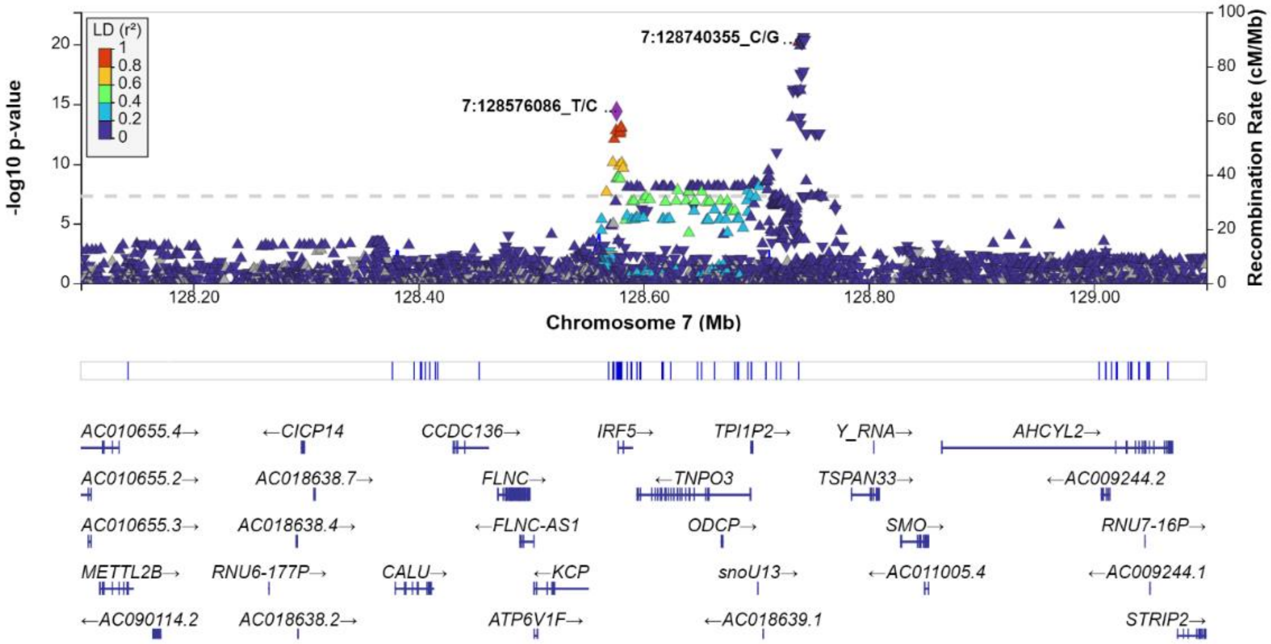


Figure 10. Regional association plot based on unconditional eGFR-association analysis within the UKBB with highlighted index SNP rs3757387 (7:128576086_T/C) and tertiary signal rs13227214 (7:128740355_C/G).

2.2.4 Association studies

The section on revealing disease associated loci provides examples in which I conducted GWAS projects including SHIP studies with specific focus on kidney function. The main findings of the kidney function projects include 34 novel loci for UACR[3], 166 new loci associated with eGFR[4], 147 new loci for serum urate concentrations[44], six new loci for rapid kidney function decline[45] . In addition, we identified 11 loci associated with heart failure[46] and 15 loci associated with muscle weakness[47].

In addition to GWAS, I conducted EWAS and EWAS meta-analysis projects including SHIP-Trend. The main finding include 69, 7, 100, and 1 CpG sites where DNA methylation associated with UACR[48], eGFR[48], serum urate levels[49], and common carotid intima-media thickness (cIMT)[50], respectively.

3 References

1. Goldstein DB. (2001). Islands of linkage disequilibrium. *Nat Genet*; Oct; **29(2)**:109-11.
2. Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics*; **49 (1)**, 49–67.
3. Teumer, A. *et al*. (2019). Genome-wide association meta-analyses and fine--mapping elucidate novel pathways influencing albuminuria. *Nat Commun*; **10**, 1–19.
4. Wuttke, M. *et al*. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. *NATURE GENETICS*; **51(6)**, 957.
5. Völzke, H. *et al*. (2022). Cohort Profile Update: The Study of Health in Pomerania (SHIP). *Int. J. Epidemiol*; **51(6)**, e372–e383.
6. Bycroft, C. *et al*. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature*; **562 (7726)**, 203–209.
7. Sudlow C., Gallacher J., Allen N. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*; 12(3).
8. Köttgen, A. & Pattaro, C. *et al*. (2020). The CKDGen Consortium: ten years of insights into the genetic basis of kidney function. *Kidney Int*; **97**, 236–242.
9. Bycroft C. *et al*. (2017). Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, 166298.
10. Marchini J, Howie B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*; **11 (7)**: 499–511.
11. Gibbs, R. *et al*. (2003). The international HapMap project. *Nature*; **426 (6968)**:789–796.
12. Frazer, K. A. *et al*. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*; **449(7164)**: 851–861.
13. Abecasis, G. *et al*. (2010). A map of human genome variation from populationscale sequencing. *Nature*; **467(7319)**: 1061–1073.
14. The Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet;* **48**, 1279–1283.
15. Das, S. *et al*. (2016). Next-generation genotype imputation service and methods. *Nat. Genet*. **48**, 1284–1287.
16. McCarthy, S. *et al*. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet*. **48**, 1279–83.
17. Loh, P.-R. *et al*. (2015). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics*; **47**, 284–290.
18. Zeggini E, Ioannidis JP. (2009). Meta-analysis in genome-wide association studies. *Pharmacogenomics*;**10(2)**:191-201.
19. Willer, C. J., Li, Y. & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genome wide association scans. *Bioinformatics*; **26**, 2190–1.

20. Higgins, J. P. T. *et al.* (2003). Measuring inconsistency in meta analyses. *BMJ*; **327**, 557–560.
21. Mägi, R. *et al.* (2017). Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet*. **26**, 3639–3650.
22. Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*; **21**:781-787.
23. Hochberg, Y. and Tamhane, A. C. (1987). Multiple comparison procedures. *New York: Wiley*.
24. Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc*; **62**:626-633.
25. Bonferroni, C. E. (1937). Teoria statistica delle classi e calcolo delle probabilita. In "Volume in Onore di Ricarrdo dalla Volta," Universita di Firenza, 1-62.
26. Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, **3**, 103-107.
27. Kang, G. *et al*. (2009) Weighted multiple hypothesis testing procedures. *Stat. Appl. Genet. Mol. Biol*; **8**, 1–22.
28. Genovese, C. *et al*. (2006). False discovery control with p-value weighting. *Biometrika*; **93**:509-524.
29. Yang, J. *et al*.; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet*; **44**, 369.
30. Yang, J. *et al*. (2011) A tool for genome-wide complex trait analysis. The American Journal of Human Genetics; 88(1), 76–82.
31. Nica AC, Dermitzakis ET. (2013). Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*; **368(1620)**, 20120362.
32. Nica AC, Dermitzakis ET. (2008). Using gene expression to investigate the genetic basis of complex disorders. *Human molecular genetics*; **17**: R129–R134.
33. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, *et al*. (2010) Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature*; **464**: 768–772.
34. Gillies, C. E. *et al*. (2018). An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet*; **103**, 232–244.
35. Zhu, Z. *et al.* (2006) Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–7.
36. Giambartolomei, C. *et al.* (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383.
37. Ghasemi, S. *et al*. Assessment of genome-wide significance of conditionally independent signals. *Bioinformatics*. 2021; **37(20)**, 3521–3529.
38. Ghasemi, S. *et al*. Discovery of novel eGFR-associated multiple independent signals using a quasi-adaptive method. Front Genet. 2022; **13**; 997302.

39. Teumer, A. *et al*. (2018). Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. *Nat. Commun*; **9**, 4455.
40. Liu,J.Z. *et al*.; International IBD Genetics Consortium. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet*; **47**, 979–986.
41. Wood, A. R. *et al*.; LifeLines Cohort Study. (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet*: **46**, 1173–1186.
42. Aguet F. *et al*. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. **bioRxiv**; 787903.
43. Gillies, C. E. *et al*. (2018). An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet*; **103**, 232–244.
44. Tin, A. *et al*. (2019). Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nat. Genet*. **51**, 1459–1474.
45. Gorski, M. *et al*. (2021). Meta-analysis uncovers genome-wide significant variants for rapid kidney function decline. *Kidney international*; **99(4)**, 926-939.
46. Shah ,S. *et al*. (2020). Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nature communications*; **11(1)**, 163.
47. Jones, G. *et al*. (2021). Genome-wide meta-analysis of muscle weakness identifies 15 susceptibility loci in older men and women. *Nature communications*;**12(1)**, 654.
48. Schlosser, P. *et al*. (2021). Meta-analyses identify DNA methylation associated with kidney function and damage. Nature communications; **12(1)**, 7174.
49. Tin, A. *et al*. (2021). Epigenome-wide association study of serum urate reveals insights into urate co-regulation and the SLC2A9 locus. *Nature communications*; **12(1)**, 7173.
50. Portilla-Fernández, E. *et al*. (2021). Meta-analysis of epigenome-wide association studies of carotid intima-media thickness. *Eur J Epidemiol*; **36**, 1143–1155.

Author contributions

Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria (Article I)

Alexander Teumer, Yong Li, **Sahar Ghasemi**. et al. **2019**;10 (1):1-19.
The project was part of CKDgen (round 4). Sahar Ghasemi was part of the project under the supervision of Alexander Teumer. Sahar Ghasemi contributed equally as the first author of this article. Sahar Ghasemi performed statistical analyses and interpreted the results as shown in green in Figure 7. Sahar Ghasemi was part of the writing group. All authors contributed to manuscript revision, read, and approved the submitted version. The manuscript was published by Nature communications journal.

Assessment of significance of conditionally independent GWAS signals (Article II)

Sahar Ghasemi, Alexander Teumer, Matthias Wuttke, Tim Becker. **2021**; 37(20), 3521–3529.
Sahar Ghasemi and Tim Becker contributed to conception and design of the study. Sahar Ghasemi performed the statistical analyses. Tim Becker supervised the project. Tim Becker acquired funding for the analyses. Sahar Ghasemi wrote the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. The manuscript was published by Bioinformatics journal.

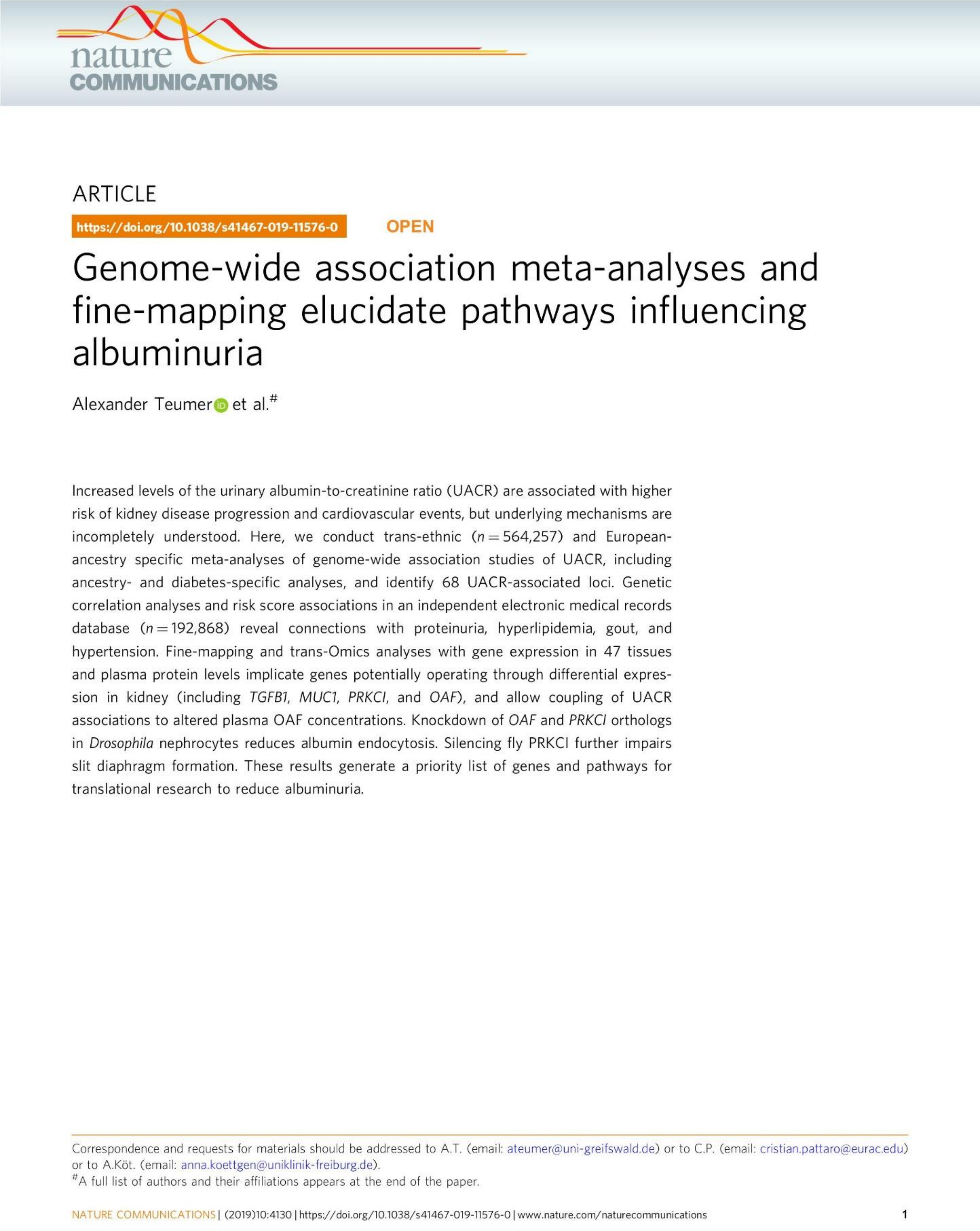
Discovery of novel eGFR-associated multiple independent signals using a quasi-adaptive method (Article III)

Sahar Ghasemi, Tim Becker, Hans J. Grabe, Alexander Teumer. **2022**; 13, 997302.
Sahar Ghasemi, Tim Becker, and Alexander Teumer contributed to conception and design of the study. Sahar Ghasemi performed the statistical analyses. Alexander Teumer supervised the project. Alexander Teumer and Hans J. Grabe acquired funding for the analyses. Sahar Ghasemi wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version. The manuscript was published by Frontiers in Genetics journal.

Sahar Ghasemi Prof. Dr. Volkmar liebscher

Articles

Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria (Article I)



nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-11576-0

OPEN

Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria

Alexander Teumer et al.[#]

Increased levels of the urinary albumin-to-creatinine ratio (UACR) are associated with higher risk of kidney disease progression and cardiovascular events, but underlying mechanisms are incompletely understood. Here, we conduct trans-ethnic ($n = 564,257$) and European-ancestry specific meta-analyses of genome-wide association studies of UACR, including ancestry- and diabetes-specific analyses, and identify 68 UACR-associated loci. Genetic correlation analyses and risk score associations in an independent electronic medical records database ($n = 192,868$) reveal connections with proteinuria, hyperlipidemia, gout, and hypertension. Fine-mapping and trans-Omics analyses with gene expression in 47 tissues and plasma protein levels implicate genes potentially operating through differential expression in kidney (including *TGFB1*, *MUC1*, *PRKCI*, and *OAF*), and allow coupling of UACR associations to altered plasma OAF concentrations. Knockdown of *OAF* and *PRKCI* orthologs in *Drosophila* nephrocytes reduces albumin endocytosis. Silencing fly PRKCI further impairs slit diaphragm formation. These results generate a priority list of genes and pathways for translational research to reduce albuminuria.

Correspondence and requests for materials should be addressed to A.T. (email: ateumer@uni-greifswald.de) or to C.P. (email: cristian.pattaro@eurac.edu) or to A.Köt. (email: anna.koettgen@uniklinik-freiburg.de).

[#] A full list of authors and their affiliations appears at the end of the paper.

NATURE COMMUNICATIONS | (2019)10:4130 | https://doi.org/10.1038/s41467-019-11576-0 | www.nature.com/naturecommunications

Higher levels of the urinary albumin-to-creatinine ratio (UACR) are associated with adverse clinical outcomes, such as end-stage kidney disease, cardiovascular disease (CVD), and mortality[1–5]. Elevated UACR is a measure of kidney damage that is used to diagnose and stage chronic kidney disease (CKD)[6], which affects >10% of adults worldwide[7], and represents a hallmark of diabetic kidney disease[8]. Even moderate elevations in UACR predict poorer health outcomes, independently of the glomerular filtration rate[4,5]. Lowering of UACR by pharmacological inhibition of the renin–angiotensin–aldosterone system (RAAS) is considered renoprotective standard of care to slow CKD progression.[9–11] RAAS blockage is associated with a reduction of albuminuria and lower risk of end-stage kidney disease[12] and CVD events[10,13–15]. However, the risk of CVD events among CKD patients remains high[3]. A better understanding of the pathways related to the development and consequences of albuminuria may facilitate the search for novel therapies to treat or prevent CKD progression and CVD.

Levels of UACR have a heritable component in population-based studies and groups at high risk of CKD, such as certain indigenous populations or persons with diabetes[16–20]. However, the identification of genetic loci for UACR through genome-wide association studies (GWAS) has proven difficult, and detected loci showed variable effects across ancestries or disease groups[21]. Initial GWAS of UACR identified only two genome-wide significant loci, *CUBN*[22,23] and *HBB*[24]. A complementary approach using admixture mapping also identified the *BCL2L11* locus[25]. One additional finding in patients with type I diabetes[26] was not detected in type II diabetes patients or the general population. Only very recently, a Mendelian Randomization study assessing a potentially causal effect of UACR on cardiometabolic traits based on data from the UK Biobank (UKBB) reported 33 genome-wide significant single-nucleotide polymorphisms (SNPs) associated with UACR[27]. The study supported a causal effect of higher UACR on elevated blood pressure and postulated that inhibition of UACR-increasing pathways could have anti-hypertensive effects and thereby reduce CVD risk.

In this project, we characterize known and identify additional novel genetic loci for UACR through trans-ethnic meta-analysis of GWAS from 564,257 participants, including an internal validation step and secondary analyses among participants with diabetes. To prioritize the most likely causal variants, genes, tissues, and pathways in associated loci, we perform functional enrichment analyses, statistical fine-mapping and integrative trans-Omics analyses, including with gene expression in 47 human tissues and plasma protein levels. Clinical correlates are identified through genome-wide genetic correlation analyses and a phenome-wide association scan of a genetic risk score for UACR in a large independent population. We evaluate translation to mechanistic insights in proof-of-concept studies for *OAF* and *PRKCI* using an experimental model of albuminuria. Together, the implicated variants, genes, proteins, tissues, and pathways provide a rich resource of new targets for translational research.

Results

The workflow of our study, which identified 68 UACR-associated loci across primary and secondary analyses, is illustrated in Supplementary Fig. 1.

Primary analysis: identification of 59 loci for UACR. The data based on 564,257 individuals from 54 studies were combined in a trans-ethnic meta-analysis of UACR, including 547,361 of European ancestry (EA), 6795 African Americans (AA), 6324 of East Asian ancestry, 2335 of South Asian ancestry, and 1442 Hispanics (Supplementary Data 1). The median of the median UACR across studies was 7.5 mg/g, and an average of 14.9% (range 3.2–70.9%) of participants had microalbuminuria (MA, UACR > 30 mg/g). Study-specific GWAS of UACR were carried out using imputed genotypes (Methods, Supplementary Data 2). We performed study-specific variant filtering and quality control (QC), followed by fixed-effects inverse-variance weighted meta-analysis. There was no evidence of unaccounted stratification (LD score regression intercept 0.95; genomic control (GC) parameter λ_{GC} 1.03). Downstream analyses were based on 8,034,757 SNPs available after variant filtering (Methods). Using SNPs of minor allele frequency (MAF) > 1% across the genome, the heritability of UACR was estimated as 4.3%.

We identified 59 UACR-associated loci, defined as 1 Mb genomic segments carrying at least one SNP associated with UACR with $p < 5 \times 10^{-8}$ (Methods; Fig. 1, Supplementary Data 3). The index SNP mapped within 500 kb of previously reported index SNPs for UACR at 27 loci, considered known, and the remaining 32 loci were considered novel. These 59 SNPs explained 0.69% of the variance of the inverse normal transformed UACR residuals. There was little evidence of between-study heterogeneity (median I^2 statistic 3.2%; Supplementary Data 3), with all index SNPs showing an I^2 of <50%. In meta-regression analysis (Methods), none of the 59 index SNPs showed evidence of ancestry-related heterogeneity after multiple testing correction ($p < 8.5 \times 10^{-4}$, Fig. 1; Supplementary Data 3)[28]. Regional association plots of all loci are displayed in Supplementary Fig. 2.

Some of the loci contain biologically plausible candidates in addition to the known *CUBN* (cubilin) locus: for example, rare mutations in *COL4A4* (Collagen Type IV Alpha 4 Chain) cause Alport syndrome, a monogenic disease of basement membranes that frequently leads to end-stage kidney disease. Recent sequencing studies show that the phenotypic spectrum of rare *COL4A4* mutations extends to focal segmental glomerulosclerosis, which typically presents with proteinuria[29,30]. Our study extends the genetic spectrum to common *COL4A4* variants associated with UACR in mostly population-based studies. Another example is *NR3C2* (Nuclear Receptor Subfamily 3 Group C Member 2), which encodes the mineralocorticoid receptor that mediates aldosterone action. Pharmacological inhibition of the RAAS is the mainstay treatment to lower albuminuria, illustrating the potential for pharmacological intervention on pathways identified in this project.

Lastly, we estimated the number of expected discoveries and the corresponding percentage of GWAS heritability explained in future studies of yet larger sample size (Methods)[31] and found that such studies can be expected to detect additional UACR loci (Supplementary Fig. 3).

Concordance between CKDGen cohorts and UK Biobank. To assess the influence of the UKBB, the largest study in the discovery sample ($n = 436,392$), we compared association statistics for the 59 index SNPs from the UKBB to the corresponding estimates from the 53 other studies participating in the CKDGen Consortium ($n \leq 127,865$). Effect direction was consistent for all 59 index SNPs ($p_{\text{binomial test}} = 3.5 \times 10^{-18}$; Fig. 2a), and 53 showed nominally significant associations in the CKDGen cohorts alone ($p < 0.05$; Supplementary Data 4). Two loci with strong effects in UKBB but not significant in CKDGen were *AHR* (aryl hydrocarbon receptor) and *CYP1A1* (Cytochrome P450 Family 1 Subfamily A Member 1), potentially reflecting factors related to standardized sample handling, storage, and measurements in the UKBB, or population-specific exposures.

Secondary ancestry-specific and diabetes-specific analyses. First, we conducted ancestry-specific meta-analyses for EA ($n =$

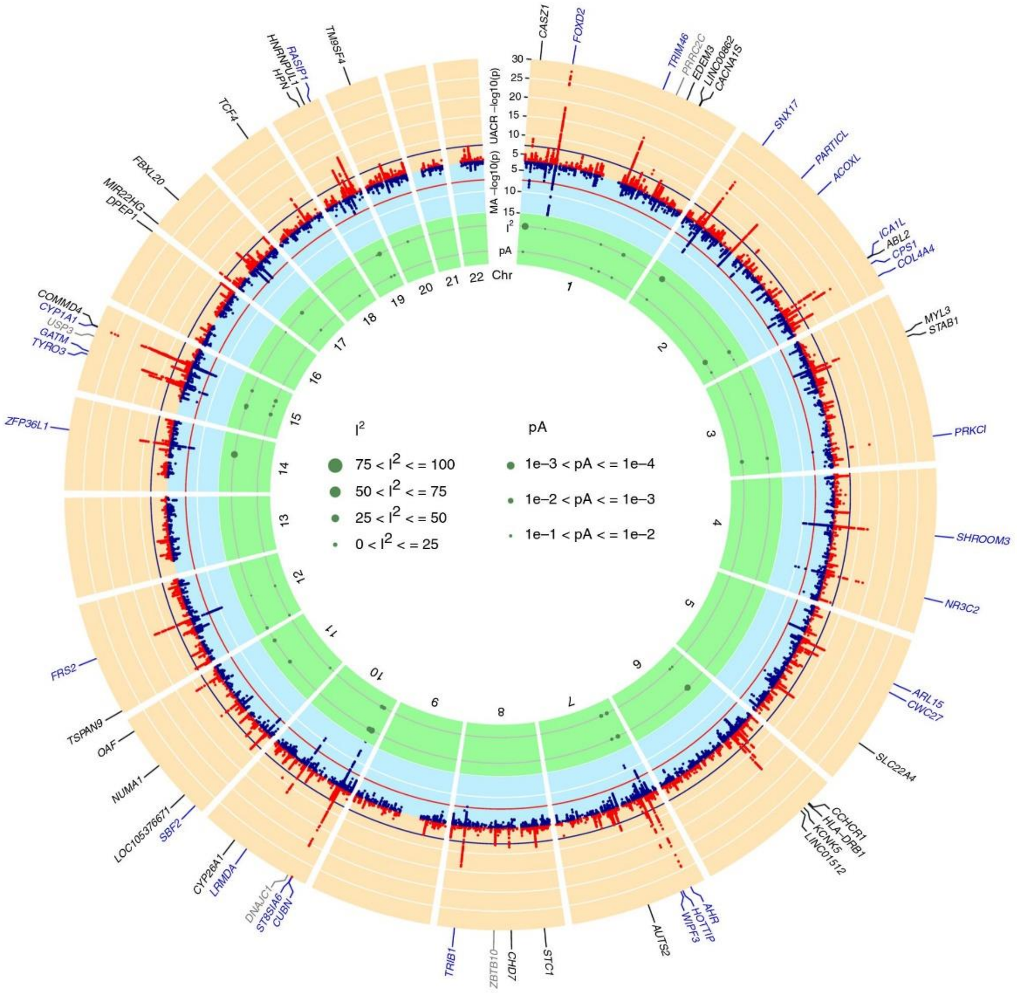


Fig. 1 Genome-wide association results. The circos plot provides an overview of the association results: Red band: $-\log_{10}(p)$ for association in the trans-ethnic meta-analysis of urinary albumin-to-creatinine ratio (UACR), ordered by chromosomal position. The blue line indicates genome-wide significance ($p = 5 \times 10^{-8}$). Black gene labels indicate novel loci, blue labels indicate known loci (known index SNP within ± 500 kb region of current index SNP), gray labels indicate loci not associated with UACR at the nominal significance level ($p \geq 0.05$) in the 53 CKDGen cohorts without UKBB. Blue band: $-\log_{10}(p)$ for association with microalbuminuria (MA), ordered by chromosomal position. The red line indicates genome-wide significance ($p = 5 \times 10^{-8}$). Green band: measures of heterogeneity related to the UACR-associated index SNPs, where the dot sizes are proportional to two measures of heterogeneity, I^2 and the $-\log_{10}$(p) for heterogeneity attributed to ancestry (pA)

547,361) and for AA ($n = 6795$), where ancestry-specific loci have been described[32,33]. There was little evidence of inflation of the results (λ_{GC} 1.06 for AA and 1.01 for EA; Methods). These meta-analyses identified 61 loci in EA, of which 56 overlapped with those from the primary trans-ethnic meta-analysis (Supplementary Data 5 and further discussed below), and no genome-wide significant loci in AA. The known UACR-associated sickle cell trait variant rs334 in *HBB* showed suggestive association in the AA-specific analysis ($p = 6.1 \times 10^{-8}$).

The other secondary analysis was restricted to 51,541 individuals with diabetes, in whom a larger effect of the known *CUBN* locus has been reported[23]. This analysis identified eight loci (Supplementary Fig. 4), four of which were not detected in the primary meta-analysis (*KAZN* [Kazrin, Periplakin Interacting Protein], *MIR4432HG-BCL11A*, *FOXP2*, and *CDH2*). Internal validation of the UKBB ($n = 21,703$) and CKDGen cohorts ($n \leq 29,812$) statistics found the effects to be direction consistent, of similar magnitude and at least nominally significant in both subsets at all eight loci (Supplementary Data 6). Index SNPs at *CUBN* and *HPN* (Hepsin) showed larger effect sizes among those with diabetes compared with the overall sample (Supplementary Data 6). Among the novel loci, it is noteworthy that *BCL11A*, a transcriptional regulator of insulin secretion[34], is involved in fetal-to-adult globin switching, as is the known UACR risk gene

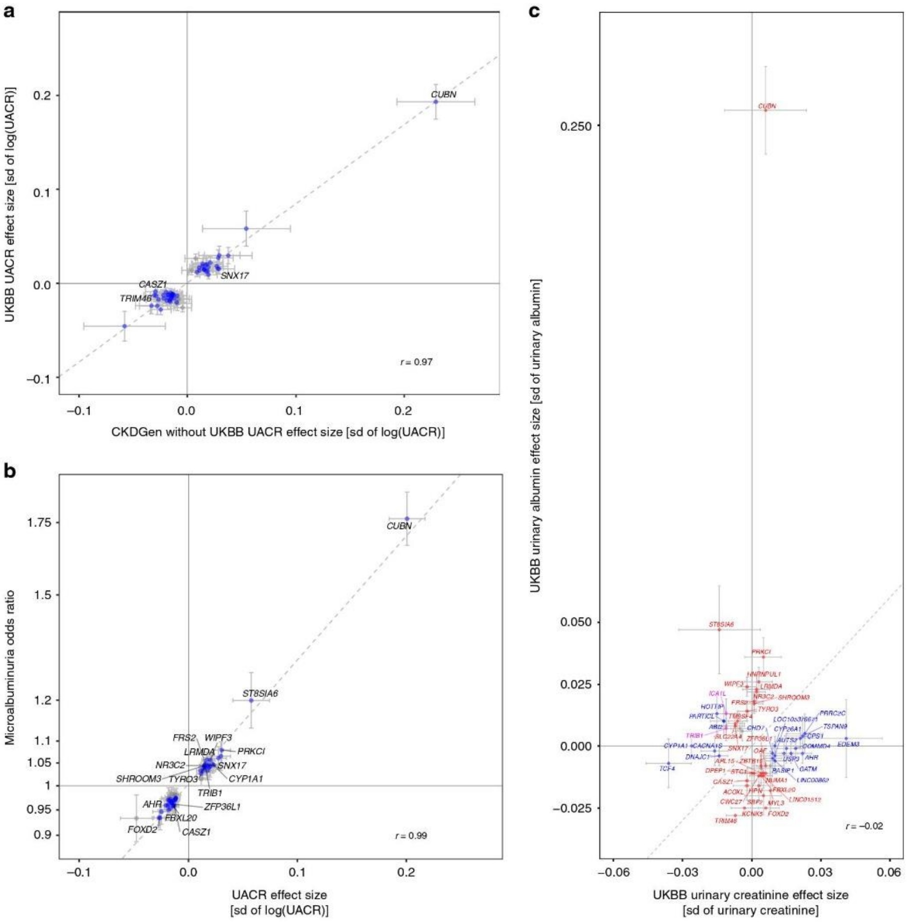


Fig. 2 Internal concordance of the urinary albumin-to-creatinine ratio (UACR) results, and association with microalbuminuria, urinary creatinine and albumin. **a** Comparison of effect estimates of the 59 genome-wide significant trans-ethnic UACR index SNPs in the UKBB (*x*-axis) and in the CKDGen cohorts without UKBB (*y*-axis). Blue dots indicate nominal significance ($p < 0.05$) in the CKDGen cohorts without UKBB, and loci at genome-wide significance ($p < 5 \times 10^{-8}$) in that meta-analysis are labeled with the closest gene. **b** Comparison of effect estimates of the 59 trans-ethnic UACR index SNPs (*x*-axis) with their corresponding estimate from the GWAS of microalbuminuria (MA; *y*-axis). Blue dots indicate significance in the MA results after multiple testing correction ($p < 0.05/59 = 8.5 \times 10^{-4}$), and loci that achieved genome-wide significance ($p < 5 \times 10^{-8}$) for MA are labeled. In both panels, the dashed line represents the line of best fit through the effect estimates. **c** Comparison of effect estimates of the 59 genome-wide significant trans-ethnic UACR index SNPs for their effect on urinary creatinine (*x*-axis) and urinary albumin levels (*y*-axis) in the UKBB sample. Blue, red, and purple color indicate significant associations after multiple testing correction ($p < 0.05/59 = 8.5 \times 10^{-4}$) with urinary creatinine, urinary albumin, and both, respectively. Significant associations are labeled with the closest gene name. The dashed line represents the median $y = x$. In all panels, error bars indicate 95% confidence intervals (CIs), and the Pearson correlation coefficient *r* between the effect estimates is shown. The effect directions correspond to the effect allele of the trans-ethnic UACR meta-analysis results

HBB. *KAZN* encodes for a protein with a role in actin organization and adhesion[35] that is highly abundant in glomeruli. QQ plots and Manhattan plots of the secondary meta-analyses are shown in Supplementary Figs. 5 and 6.

Functional enrichment and pathways. We searched for tissues, cell types, and systems that are enriched for the expression of genes mapping to the UACR-associated loci (Methods)[36]. Based on all SNPs with $p < 5 \times 10^{-8}$ from the trans-ethnic meta-analysis, there was no significant (false discovery rate [FDR] < 0.05) enrichment after correction for multiple testing (Supplementary Data 7). Nominally significant associations ($p < 0.05$) were observed for 37 annotations mapping into six systems (urogenital including kidney, endocrine, digestive including liver, musculoskeletal, respiratory, sense organs; Supplementary Fig. 7) and five tissues (exocrine glands, prostate, mucous membrane, membranes, and respiratory mucosa). These results reveal plausible enrichments although they did not reach significance after correction for multiple testing.

Next, we evaluated whether reconstituted gene sets were significantly (FDR < 0.05) enriched for genes mapping to UACR-associated loci, and identified three sets with FDR < 0.01 (embryonic development, partial embryonic lethality during organogenesis, abnormal placental labyrinth vasculature morphology). The remaining significant gene sets included terms that can be reconciled with existing knowledge about albuminuria, including "tube development", "abnormal kidney morphology", and several terms related to vascular development and morphology (Supplementary Data 8).

UACR-associated loci are associated with MA. Clinical MA (UACR > 30 mg/g) is associated with increased risk for adverse kidney and cardiovascular outcomes, as well as mortality[3]. We therefore evaluated the association of the 59 UACR index SNPs with MA by meta-analyzing data from 36 cohorts and 347,283 individuals (Supplementary Data 1; Fig. 1). Figure 2b shows that for all UACR index SNPs, the allele associated with higher UACR was associated with an increased risk of MA (Supplementary Data 3). Of the 59 SNPs, 49 were significantly associated with MA after correction for multiple testing ($p < 0.05/59 = 8.5 \times 10^{-4}$), including 17 that reached genome-wide significance. The low-frequency missense SNP rs45551835 in *CUBN* showed the largest effect with an odds ratio (OR) of 1.76 (95% CI 1.67–1.87) per minor allele. When 232,751 UKBB participants were grouped into quartiles based on a UACR genetic risk constructed from the 59 index SNPs, each quartile showed a significantly higher OR for MA compared with the lowest quartile (e.g., OR of 1.69 for quartile 4 vs. 1, $p = 3.0 \times 10^{-191}$, Supplementary Table 1).

UACR loci: association with urinary albumin and creatinine. The UACR is a ratio. Understanding whether a genetic locus is more strongly associated with its numerator, albumin, or with its denominator, creatinine, may provide important physiological insights. We therefore performed separate tests for urinary albumin and creatinine in the UKBB sample ($n_{\text{Ualbumin}} = 436{,}398$; $n_{\text{Ucreatinine}} = 436{,}412$). Of the 59 index SNPs, 31 were significantly associated with urinary albumin ($p < 8.5 \times 10^{-4}$), 21 with urinary creatinine, and two with both. The *CUBN* locus showed the largest effect on urinary albumin, and was not significantly associated with urinary creatinine levels (Fig. 2c), followed by *ST8SIA6* (ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 6), *PRKCI* (protein kinase C iota), *TRIM46*/*MUC1* (Mucin 1, cell surface associated), *HNRNPUL1*/*TGFB1* (transforming growth factor beta 1), *FOXD2*, *KCNK5*, *WIPF3* (WAS/WASL interacting protein family member 3), *LRMDA*, and *NR3C2*.

A genetic UACR score is associated with medical diagnoses. Next, we evaluated whether a weighted genetic risk score (GRS) composed of UACR-increasing alleles was associated with clinical endpoints in a large, independent electronic medical record database to detect diagnoses with potentially shared genetic components or co-regulation. We tested associations with 1422 billing code-based phenotypes of up to 192,868 EA participants of the Million Veteran Program (MVP) from US Veterans' Administration facilities[37]. Significant associations ($p < 3.5 \times 10^{-5}$, 0.05/1,422) were detected with 10 diagnoses: proteinuria, four related to hyperlipidemia, two related to hypertension, two related to gout, as well as Fuchs' dystrophy (Fig. 3). While the association with disorders of lipoid metabolism had the lowest *p*-value ($p = 4.1 \times 10^{-11}$), the association with Fuchs' dystrophy showed the greatest magnitude (OR = 6.68 per SD increase of log [UACR], 95% CI 3.06–14.59, $p = 1.9 \times 10^{-6}$), followed by proteinuria (OR = 2.7, 95% CI 1.76–4.14, $p = 5.0 \times 10^{-6}$). Many other associations that approached statistical significance were related to the kidney and metabolic diseases (Supplementary Data 9).

The association with Fuchs' disease, a dystrophy of the corneal endothelium, was unexpected and assessed in greater detail. Autosomal-dominant forms of Fuchs' dystrophy have been attributed to genetic variation in *TCF4* (transcription factor 4)[38], a novel UACR-associated locus identified here (index rs11659764, $p = 2.8 \times 10^{-11}$; $r^2 = 0.21$, D' = −0.97 with rs613872, a previously reported Fuchs index SNP[39]). After exclusion of the *TCF4* index SNP, the GRS was still significantly associated with proteinuria, hyperlipidemia codes, gout, and hypertension with nearly identical ORs, but the association with Fuchs' dystrophy disappeared ($p = 0.2$). This illustrates that unexpected significant associations from PheWAS require careful evaluation.

We also evaluated an association of the GRS with cardiovascular outcomes based on published GWAS and the UKBB (Supplementary Table 2). This revealed significant ($p < 0.007$, Methods) positive associations of the GRS with an increased risk of hypertension ($p = 2.4 \times 10^{-21}$). Conversely, weighted genetic risk scores based on recently published GWAS of systolic and diastolic blood pressure as well as of type 2 diabetes were positively associated with UACR ($p = 3.5 \times 10^{-63}$ for systolic and $p = 1.2 \times 10^{-24}$ for diastolic blood pressure, $p = 1 \times 10^{-10}$ for type 2 diabetes; Supplementary Table 2).

Genome-wide genetic correlations of UACR. Albuminuria is associated with multiple cardiovascular and metabolic traits and diseases[4,40–42]. In addition to the GRS analyses, we thus also assessed genome-wide genetic correlations between the EA-specific UACR association statistics and 517 traits and diseases (Methods; Supplementary Data 10). Significant genetic correlations ($p < 9.7 \times 10^{-5}$ [0.05/517]) were observed for 67 traits (Fig. 4). The strongest negative correlations were observed for urinary creatinine and other urinary parameters, and the largest positive genetic correlations with different measures of hypertension. These findings provide support for the observational association between albuminuria and blood pressure on a genetic level, the significant associations between the UACR GRS and hypertension in the MVP population, and the recent Mendelian Randomization study of UACR[27]. Negative genetic correlations with anthropometric measures are potentially explained by their positive associations with muscle mass, and hence creatinine concentrations.

Statistical fine-mapping and secondary signal analysis. Statistical fine-mapping was performed using summary statistics to prioritize SNPs or sets of SNPs (credible set) driving each association signal (Methods). These analyses were limited to EA, comprising > 97% of the total sample, for whom large data sets to estimate reference LD for summary statistics-based fine-mapping were publicly accessible[43,44]. Based on 57 combined genomic regions from the 61 genome-wide significant loci in EA (Methods, Supplementary Data 5), we identified 63 independent SNPs (Supplementary Data 11). Next, 99% credible sets were computed based on Approximate Bayes Factors, resulting in a set of SNPs that with 99% posterior probability (PP) contained the variant(s) driving the association signal for each of the 63 conditionally independent signals[45]. The credible sets contained a median of 25 SNPs (Quartile 1: 10; Quartile 3: 74). Two credible sets at *CUBN* and one at *PRKCI* consisted of a single SNP (Supplementary Data 12). The previously described *CUBN* missense SNP rs45551835 (p.A2914V) had a PP of causing the association signal of >99.9%. There were 11 small credible sets with ≤5 SNPs, representing candidate causal variants for further study.

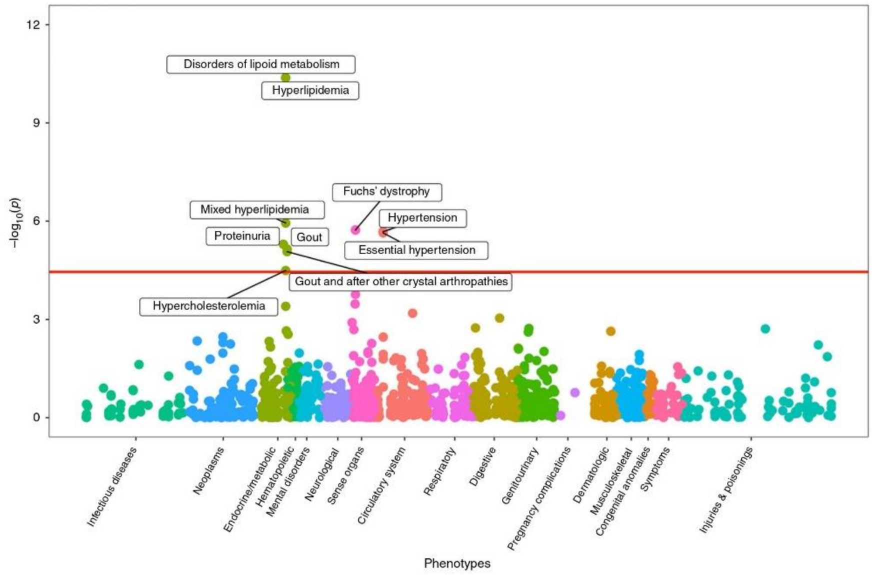


Fig. 3 Phenome-wide association scan of a genetic urinary albumin-to-creatinine ratio (UACR) risk score. PheWAS association results were obtained from EA participants of the Million Veteran Program. Association test -$\log_{10}$(p-values) are plotted on the y-axis, and the corresponding trait or disease category on the x-axis. Significant results, after correcting for the 1422 phenotypes tested ($p < 0.05/1422 = 3.5 \times 10^{-5}$), are labeled in the figure

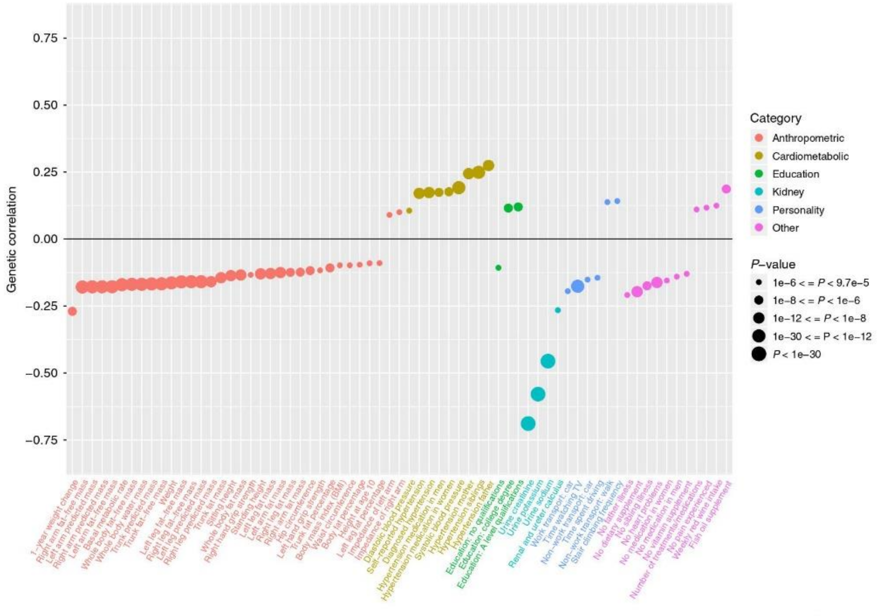


Fig. 4 Genetic correlation of urinary albumin-to-creatinine ratio (UACR) with other traits and diseases. Significant ($p < 9.7 \times 10^{-5}$) genetic correlations based on the genome-wide summary statistics from the EA UACR GWAS and 517 pre-computed and publicly available GWAS summary statistics of UKBB traits and diseases, available through LDHub. Traits are shown on the x-axis, and colored according to broad physiological categories. Genetic correlations between traits and UACR are reported on the y-axis. Dot size is proportional to the -$\log_{10}(p)$ of the corresponding genetic correlation

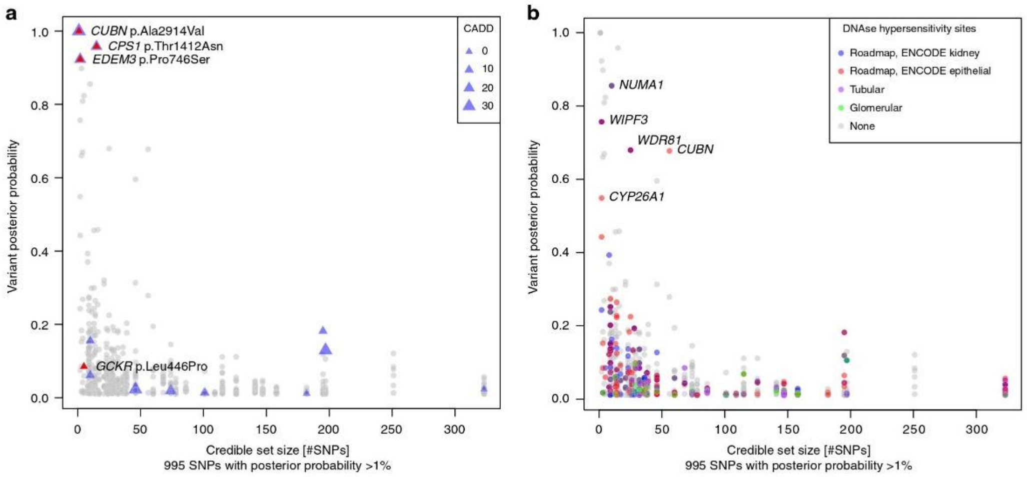


Fig. 5 Fine-mapping and functional annotation of potentially causal variants. Overview of 995 SNPs with a posterior probability of association with urinary albumin-to-creatinine ratio (UACR) of >1%. The *x*-axis indicates the 99% credible set size and the *y*-axis the SNPs' posterior probability of association. In panel **a**, missense SNPs are marked by triangles, with size proportional to the SNP CADD score. In panel **b**, SNPs are color-coded with respect to location in regulatory regions of specific kidney tissues. The labels show the closest gene, and are restricted to variants mapping to small credible sets (≤5 SNPs), or to variants with high individual posterior probability (>0.5) of driving the association signal. For the *CUBN* locus, a credible set was computed for each independent SNP

All 995 SNPs with PP > 1% were annotated. Regulatory potential was assessed via mapping into regions of open chromatin identified from primary cultures of human tubular and glomerular cells (GEO accession number GSE115961)[46] and from publicly available kidney cells types (ENCODE and Roadmaps Projects; Methods). Supplementary Data 12 summarizes annotation information for all variants with PP > 1% that mapped into small credible sets or those containing a SNP with PP > 50%. Among these, there were four missense SNPs in *CUBN*, *CPS1*, *EDEM3*, and *GCKR* (Fig. 5a; Supplementary Table 3). One non-exonic SNP near *NUMA1* with PP > 50% mapped into open chromatin in both glomerular and tubular primary cell cultures, and four other SNPs in or near *WIPF3*, *WDR81*, *CUBN*, and *CYP26A1* mapped into putative regulatory regions in other kidney tissues or cell lines (Fig. 5b, Supplementary Data 12).

Association with gene expression and co-localization. We investigated whether the UACR-association signals co-localized with association signals for transcript abundance of any genes in *cis* across 47 tissues, thereby implicating effector genes at associated loci (Methods). Gene expression was quantified via RNA-seq in 44 tissues from the GTEx Project [https://gtexportal.org/] and in kidney cortex from The Cancer Genome Atlas[47], and via microarray from microdissected glomerular and tubulointerstitial portions of kidney biopsies from participants of the NEPTUNE study[48] (Methods).

We identified nine genes for which *cis* eQTLs in kidney tissues co-localized with the UACR association signals with a high PP (≥80%), implicating a shared underlying variant (Fig. 6). These represent candidate causal genes for further investigation (Table 1). Alleles associated with higher UACR were associated with higher expression of *MUC1* and *PRKCI* across a range of tissues. This observation is consistent with a gain-of-function mechanism proposed for the monogenic kidney disorder caused by *MUC1* variation[49]. Conversely, alleles associated with higher UACR were associated with lower *OAF* and *TGFB1* expression. The co-localization with expression of *WIPF3* in glomerular kidney portions illustrates an example of a potentially regulatory causal variant, rs17158386, which maps into open chromatin in kidney tissue (Figs. 5b, 6). Across kidney tissues, co-localization was most often observed in glomerular kidney portions, consistent with the prominent role of the glomerular filtration barrier in albuminuria. Altogether, there were 90 significant co-localizations in at least one of the 47 evaluated tissues (Supplementary Fig. 8).

Association with gene expression in *trans* requires large sample sizes and was thus evaluated for all index SNPs in whole blood. Excluding the extended MHC region, there was one SNP associated with expression of one or more transcripts in *trans* in more than one study (Supplementary Table 4): genotype at rs12714144, upstream of *PARTICL* on chromosome 2, was associated with the expression of *DPEP3*, encoded on chromosome 16.

Association with protein levels and co-localization analyses. Recently, large GWAS of plasma protein levels have been published, which allow for systematic investigations of associated variants (pQTLs). Using these data, we investigated the association of the 61 EA index SNPs in a pQTL study of 3301 healthy EA participants of the INTERVAL study[50]. Genome-wide significant associations were identified between 17 UACR-associated SNPs and plasma levels of 53 unique proteins, for a total of 56 associations (Supplementary Data 13). Interestingly, concentrations of three proteins each showed associations with two UACR-associated index SNPs on different chromosomes, thereby connecting the two genetic loci through association with plasma concentrations of the same protein: SNPs rs34257409 on chromosome 1 and rs838142 on chromosome 19 with plasma gastrokine-2 (GKN2) concentrations, rs12714144 on chromosome 2 and rs1010553 on chromosome 3 with concentrations of Janus kinase and microtubule interacting protein 3 (JAKMIP3), and rs1010553 on chromosome 3 and rs2954021 on chromosome

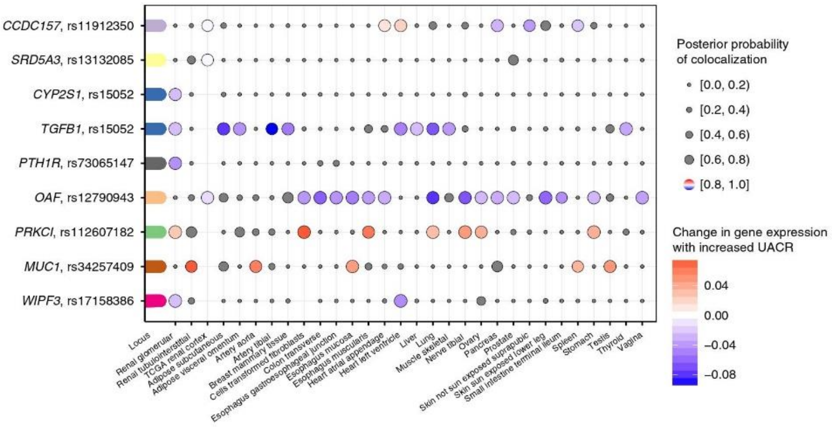


Fig. 6 Co-localization of associations signals for urinary albumin-to-creatinine ratio (UACR) and gene expression in kidney tissues. The plot shows the nine genes for which there is a high likelihood (posterior probability ≥ 80%) of a shared causal signal for gene expression in at least one of three kidney tissues and UACR. The loci are colored-coded and shown on the *y*-axis with the closest gene next to the index SNP. Co-localization with gene expression across all tissues (*x*-axis) is shown as dots, where the size of the dots (implying that eQTL data were available) corresponds to the posterior probability of the co-localization. The change in UACR is color-coded relative to the change in gene expression, or gray in case of a posterior probability < 80%

8 with inter-alpha-trypsin inhibitor heavy chain 1 (ITIH1) concentrations.

Co-localization of UACR association signals with those for pQTLs of 38 proteins (Methods, Supplementary Table 5) provided evidence for a shared underlying SNP for plasma concentrations of the Out At First Homolog (OAF) protein. This was consistent with the eQTL co-localization analyses, with the minor T allele at rs12790943 associated with higher levels of UACR as well as with both lower *OAF* transcript levels in multiple tissues and lower OAF plasma levels (Fig. 7). Association patterns with UACR (Fig. 7a) and *OAF* transcript levels (Fig. 7b) looked similar, as expected for a shared underlying variant. The pattern looked different for OAF plasma levels, and conditional analyses revealed two independent SNPs (rs117554512 and rs508205; $r^2 = 0$, $D' = 0.02$ in the 1000 Genomes Project EUR sample). There was no evidence for a shared variant underlying the associations of UACR and OAF plasma levels for the signal tagged by the initial index SNP for OAF plasma levels, rs117554512 (PP H4 = 0; Fig. 7c), which was also significantly associated with plasma levels of IL25 in *trans* ($p = 1.3 \times 10^{-12}$, Supplementary Data 13). Conversely, there was strong evidence for a shared variant underlying associations with UACR and OAF plasma levels tagged by the second, independent signal at rs508205 (PP H4 = 0.99; Fig. 7d), allowing to follow associations from genetic variants to transcript, protein, and phentoype. The SNP rs508205 is located upstream of *OAF*, and was also the index variant identified in the trans-ethnic meta-analysis of UACR ($r^2 = 0.94$ with rs12790943 in the 1000 Genomes Project EUR sample). It represents an interesting regulatory candidate variant because of its relatively small credible set of eight SNPs, a CADD score of 13, and its localization in open chromatin in kidney tissue.

In vivo analyses of Drosophila orthologs. Finally, we used a *Drosophila* model to establish proof-of-principle that prioritized candidates can be used to gain mechanistic insights into albuminuria. *Drosophila* nephrocytes are specialized cells that harbor a slit diaphragm formed by the orthologs of the mammalian slit diaphragm proteins. These cells exhibit size-dependent molecule filtration across the slit diaphragm, followed by endocytosis via the scavenger receptor Cubilin and finally lysosomal degradation or storage. Protein endocytosis mainly occurs within a network of membrane invaginations, the labyrinthine channels. Formation of the labyrinthine channels depends on presence of functional slit diaphragms. Thus, these cells reflect aspects of glomerular (slit diaphragm) and proximal tubular function (protein endocytosis)[51]. Studying endocytosis of a tracer molecule able to pass the slit diaphragm, such as albumin, renders an integrative read-out of nephrocyte function[52]: FITC-albumin uptake declines both through loss of slit diaphragms and also through impaired protein endocytosis. We selected three candidates for functional study, based on their associations with urinary albumin (Fig. 2c), support from downstream fine-mapping and co-localization analyses (Table 1), and degree of conservation and availability of at least two independent *Drosophila* RNAi lines per gene: *OAF*, *PRKCI*, and *WIPF3*. Orthologs of *OAF* (*oaf*), *PRKCI* (*aPKC*), and *WIPF3* (*Vrp1*) were silenced specifically in nephrocytes by crossing *Dorothy-GAL4* with the respective UAS-RNAi line.

Nephrocytes stained with an available antibody for aPKC showed a strongly reduced signal using two independent *aPKC*-RNAi lines (Supplementary Fig. 9A–C). We observed no effect of *Vrp1*-RNAi on nephrocyte function studying FITC-albumin endocytosis (Supplementary Fig. 9D, E). In contrast, we detected a significant reduction of tracer endocytosis upon silencing *oaf* and *aPKC* (Fig. 8a, b). This indicates a functional requirement of these genes within nephrocytes and supports a role of their human orthologs in glomerular filtration or tubular re-uptake of albumin. To distinguish between these roles, we studied immunofluorescence of the *Drosophila* slit diaphragm proteins, whose staining patterns remain unaltered in isolated defects of protein endocytosis. Despite the significant impairment of nephrocyte function, we observed a slit diaphragm staining pattern comparable to control conditions for *oaf*-RNAi (Fig. 8c–f). This suggests that oaf may be dispensable for slit diaphragm formation, but likely is involved in protein reabsorption. Accordingly, co-localization with *OAF* gene expression in human kidney was observed in the renal cortex, reflecting largely tubulointerstitial portions, and protein staining in the Human Protein Atlas is observed in tubules but not glomeruli. Conversely, silencing the ortholog of *PRKCI* entailed an extensive

Table 1 Evidence for candidate causal genes at UACR-associated variants

Gene	SNP	H4 coloc	Credible set size	SNP PP	Functional consequence	CADD	DHS	Brief summary of literature and gene function
PRKCI	rs112607182	1.00	1	1.00	Intergenic, downstream	1.9	-	*PRKCI* encodes a serine/threonine protein kinase that plays a role in microtubule dynamics. Has been identified as an important factor for actin cytoskeletal regulation in podocytes (PMID: 24096077). Podocyte-specific deletion of aPKClambda/iota in mice results in severe proteinuria (PMID: 19279126).
TGFB1	rs15052	1.00	3	0.75	3′UTR (*HNRNPUL1*)	9.9	-	*TGFB1* encodes a transcription factor that controls proliferation, differentiation and other functions in many cell types. Has been implicated as a cause of fibrosis in most forms of experimental and human kidney disease (PMID 10793168). Numerous publications and animal models connect it to diabetic kidney disease, as well as numerous animal models.
WIPF3	rs17158386	1.00	2	0.81	Intergenic	11.6		1*, 2*, 3* The protein encoded by *WIPF3* is involved in the Cdc42/N-WASP/Arp2/3 signaling pathway-mediated remodeling of the actin cytoskeleton (PMID: 11553796).
PTH1R	rs73065147	0.98	14	0.20	Intergenic	15.1	-	*PTH1R* encodes for a receptor for parathyroid hormone, with high expression only in kidney cortex. The PTHrP/PTH1R system appears to adversely affect the outcome of diabetic and other renal diseases (PMID: 16783882, 21052497). Rare mutations have been reported to cause multiple aut-rec (#215045, #600002), or aut-dom (#125350, #156400) chondrodysplasias or tooth eruption phenotypes.
CYP2S1	rs15052	0.95	3	0.75	3′UTR (*HNRNPUL1*)	9.9	-	*CYP2S1* encodes for a member of the cytochrome P450 enzyme family, which catalyze many reactions involved in drug and lipid metabolism. It is transcriptionally regulated by AHR, also identified in the present GWAS meta-analysis, in rats (PMID: 19883719).
MUC1	rs34257409	0.89	25	0.10	Intergenic	3.1	1*	*MUC1* encodes for a membrane-bound member of the mucin family that play an essential role in forming protective mucous barriers on epithelial surfaces. Rare mutations cause medullary cystic kidney disease 1 (#174000), an autosomal-dominant tubulo-interstitial kidney disease. Patients show minimal to mild proteinuria in addition to decreased eGFR and renal cysts (PMID: 29217307).
OAF	rs12790943	0.97	7	0.47	Intergenic	1.8	1*	The *OAF* gene encodes for a transcription factor of the basic helix-loop-helix family. Relatively little is known about its function in humans.
SRD5A3	rs13132085	0.92	183	0.03	Intergenic	4.0	-	The protein encoded by *SRD5A3* gene is involved in the production of androgen 5-alpha-dihydrotestosterone, and in the conversion of polyprenol into dolichol and thereby N-linked glycosylation of proteins (PMID: 20852264). Rare mutations cause autosomal-recessive disorders of glycosylation, type Iq ((#612379) or Kahrizi syndrome (#612713).
CCDC157	rs11912350	0.88	85	0.05	Intron *SF3A1*	0.1	-	Very little is known about the role of the *CCDC157* gene, there are no specific publications. Co-localization is observed with multiple other transcripts at this locus.

PP posterior probability, *DHS* DNAse I hypersensitivity site, *SNP* index SNP from the EA-specific meta-analysis
This table includes all genes with high posterior probability ($H4 \geq 0.8$) of co-localization of the UACR association signal and gene expression in kidney tissues.
1*: ENCODE kidney, 2* ENCODE epithelial, 3* Roadmap kidney

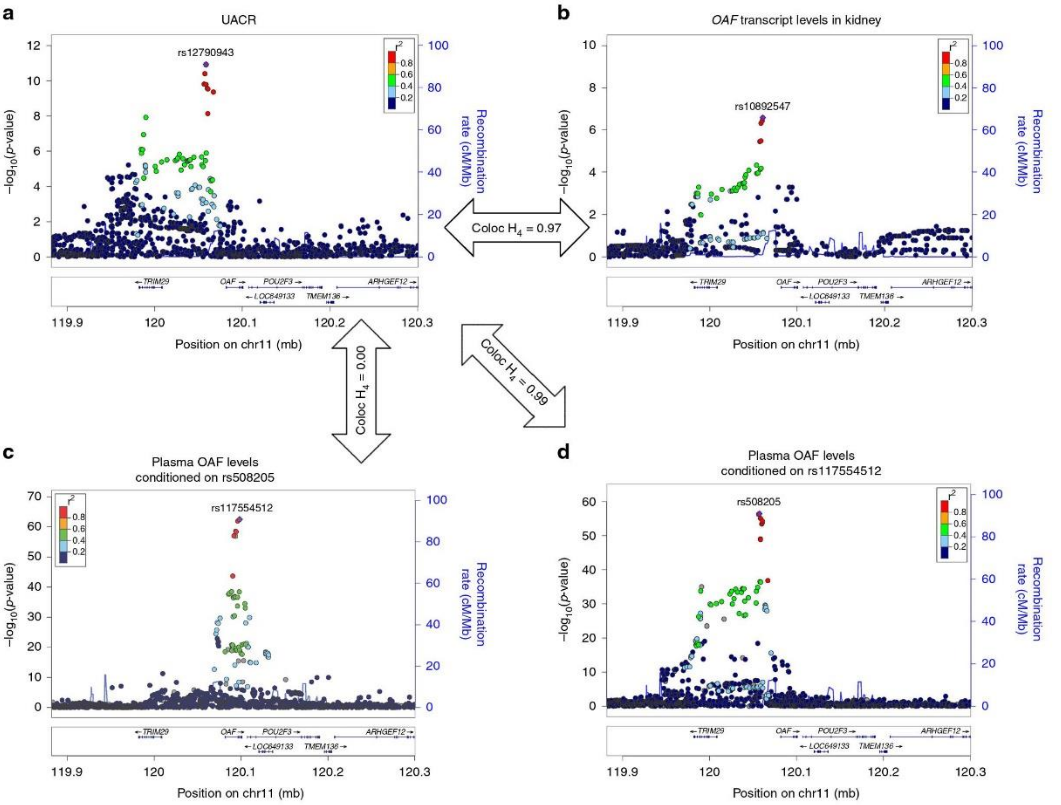


Fig. 7 Co-localization of association signals of the *OAF* locus. Regional association plots of the *OAF* locus in the European ancestry urinary albumin-to-creatinine ratio (UACR) GWAS (**a**), with *OAF* gene-expression levels in healthy kidney tissue sections (**b**), and with OAF plasma levels (**c**, **d**). The dots are colored according to their correlation r^2 with the index SNP estimated based on the 1000 Genomes EUR reference samples (gray for missing data). This locus has two independent pQTLs for OAF levels, where panel **c** shows the association between the index pQTL at the locus (rs117554512) conditioned on its secondary signal (indexed by rs508205), and panel **d** shows the association with a conditionally independent SNP (rs508205, $r^2 < 0.01$ in 1000 Genomes EUR). The secondary signal rs508205 has strong evidence of co-localization with the UACR association signal (posterior probability $H_4 = 0.99$, Methods), while the signal rs117554512 has not (posterior probability $H_4 = 0$). There was strong evidence of co-localization between the UACR association signal and *OAF* expression in kidney tissue (posterior probability $H_4 = 0.97$)

loss of slit diaphragm proteins (Fig. 8g, h; 3D reconstruction Supplementary Fig. 9K). This implies that the polarity factor aPKC is directly involved in slit diaphragm formation, consistent with studies in murine podocytes[53]. Staining patterns were comparable when silencing *oaf* and *aPKC* using second RNAi lines (Supplementary Fig. 9F–I). In summary, the *Drosophila* data support a role of *OAF* in tubular protein endocytosis and *PRKCI* in slit diaphragm formation.

Discussion

In this GWAS meta-analysis of UACR, we identified 68 loci in total, the majority of which was associated with urinary albumin concentrations and MA. Statistical fine-mapping and co-localization analyses with gene expression across 47 human tissues and with plasma protein levels resolved GWAS loci into novel driver genes and variants. This approach allowed for translating two genes prioritized in our workflow, *OAF* and *PRKCI*, into mechanistic insights in an in vivo experimental model of proteinuria. Genome-wide genetic correlation analyses and a phenome-wide association study of a genetic risk score for UACR in a large independent population highlighted a common genetic component or co-regulation with traits and diseases with renal, hepatic, or endothelial components. Together, these results represent a comprehensive resource for translational research into albuminuria.

Until recently, GWAS of UACR in mostly population-based studies only identified and replicated two loci: *CUBN*[22,54] and *HBB*[24], detected through an earlier candidate gene study[33]. In addition to these two loci, we also identified the *BCL2L11* locus, reported in an earlier admixture mapping study[25], with the index SNP mapping to the neighboring *ACOXL* gene. Our fine-mapping workflow did not provide strong evidence for either *ACOXL* or *BCL2L11* as the likely causal gene. We did not identify genome-wide significant signals at *RAB38* and *HS6ST1* among persons with diabetes, which we reported in an earlier study at suggestive significance[23]. Potential reasons include differences in quantification and statistical transformation of UACR, different participating studies, and false-positive results in the initial report. Twenty-eight of the 61 loci detected in EA individuals

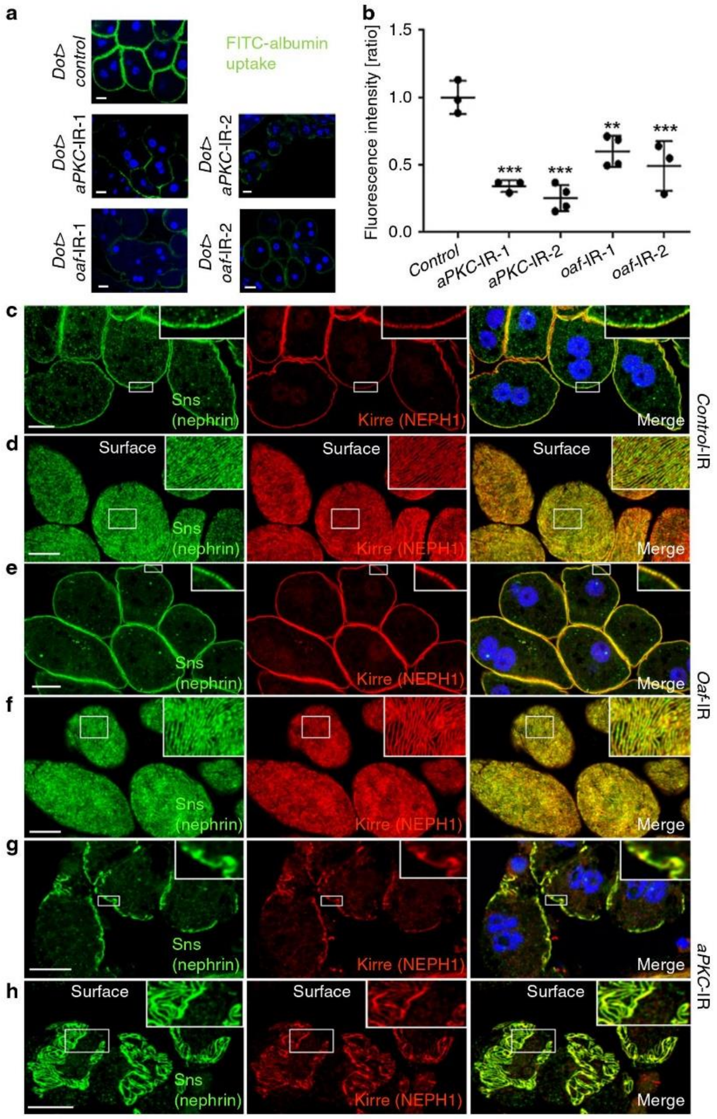


Fig. 8 In vivo results of *Drosophila* orthologs. The *Drosophila* orthologs of *OAF* and *PRKCI* (*aPKC*) are both required for nephrocyte function and *aPKC*-RNAi affects slit diaphragm formation. **a** Garland cell nephrocytes were exposed to FITC-albumin. Nephrocytes expressing control RNAi exhibit intense endocytosis, while expression of RNAi directed against *oaf* and *aPKC* (ortholog of *PRKCI*) decreases tracer uptake. **b** Quantitation of fluorescence intensity from FITC-albumin uptake is shown for the indicated genotypes. Values are presented as mean ± standard deviation of the ratio to a control experiment. Statistical significance was calculated using ANOVA and Dunnett's post hoc analysis. A statistically significant difference (defined as $p < 0.05$) is observed for *oaf*-RNAi-1 ($N = 4$), *oaf*-RNAi-2 ($N = 3$), *aPKC*-RNAi-1 ($N = 3$), and *aPKC*-RNAi-2 ($N = 4$), where ** indicate $p < 0.01$ and ***$p < 0.001$. **c** Staining the slit diaphragm proteins Sns (ortholog of nephrin) and Kirre (ortholog of NEPH1) in control nephrocytes shows regular formation of slit diaphragms. Airyscan technology partially allows for distinguishing individual slit diaphragms (insets). **d** Tangential sections through the surface of control nephrocytes reveals the regular fingerprint-like pattern of slit diaphragm proteins. **e**, **f** Expression of *oaf*-RNAi-1 does not entail an overt phenotype, suggesting reduced nephrocyte function may be a consequence of impaired protein reabsorption while slit diaphragm formation is not affected. **g**, **h** Expression of *aPKC*-RNAi-1 results in a clustered and irregular pattern of slit diaphragm proteins (insets in **g**) and a complete loss of slit diaphragm protein distinct areas on the cell surface. This suggests the loss of nephrocyte function is a consequence of impaired slit diaphragm formation. All scale bars represent 10 μm

were also reported in the recent Mendelian Randomization study of albuminuria[27], which is not surprising given the inclusion of UKBB data in our meta-analysis. Still, our study identifies 32 additional loci for UACR in the overall sample, as well as four among people with diabetes. Moreover, results allow for prioritization of loci with respect to their association with urinary albumin, whereas previous studies have not evaluated whether UACR-associated loci were driven by associations with urinary albumin, creatinine, or both.

Previous GWAS of albuminuria have not resolved associated loci into underlying genes and variants. Our workflow identified co-localization of UACR-associations with differential gene expression of *PRKCI*, *TGFB1*, *WIPF3*, *PTH1R*, *CYP2S1*, and *MUC1* in glomerular kidney portions and *OAF*, *SRD5A3*, and *CCDC157* in tubulointerstitial tissue. Some of these genes already have established roles in the function of the glomerular filter in diabetic (*TGFB1*)[55,56] and monogenic kidney disease (*MUC1*)[49], while others such as *OAF or WIPF3* represent novel candidates or, as for *PRKCI*, have not yet been implicated in humans[53]. Our combination of human and *Drosophila* studies support a role of *PRKCI* in glomerular filtration function and of *OAF* in tubular protein reabsorption, where reduced endocytosis upon gene silencing reflects the human allele associated with higher UACR and lower *OAF* expression and plasma levels. The lack of a phenotype upon silencing of the *WIPF3* ortholog may reflect the unclear state of orthology, a lack of evolutionary conservation, or potentially an insufficient knockdown.

Several insights from our study are of clinical interest. First, the clinical relevance of genes detected in our screen, *CUBN* and *COL4A4*, is underscored by a respective monogenic disease featuring albuminuria and kidney disease, Imerslund-Grasbeck (MIM 261100) and Alport syndrome (MIM 203780). Second, the identification of *NR3C2*, encoding an essential component of the RAAS, links this pathway to both albuminuria and adverse clinical outcomes. Pharmacological inhibition of the RAAS has been shown to be associated with reduced risk of end-stage kidney disease[12] and cardiovascular events[10,13–15], suggesting that genetic studies of UACR in large human populations may identify pathways amenable to pharmacological intervention that reduce both albuminuria and CVD risk. Third, the genome-wide genetic correlations of UACR and the UACR GRS associations may point toward diseases with a common genetic basis or to co-regulation of disease-relevant cell types. The latter could be reflected in the role of the liver in lipid metabolism and albumin production, the role of the kidney in urate metabolism and albumin excretion, and the role of the endothelium in hypertension and glomerular filtration. A potential role of the endothelium and the vasculature is further corroborated by the significantly enriched pathway "abnormal placental labyrinth vasculature morphology" and many other nominally enriched pathways related to angiogenesis, as well as the identification of the *VEGFA* (Vascular Endothelial Growth Factor A; *LINC01512*) locus, an important growth factor for vascular endothelial cell migration and proliferation. Interestingly, a recent Mendelian Randomization analysis of UACR and blood pressure supported a causal relationship between the two, but reported that SNPs in *CUBN* and *CYP1A1* were only associated with UACR and not blood pressure. We find that the index SNPs in *CUBN* and *CYP1A1* are related to UACR via tubular albumin reabsorption and an association with urinary creatinine but not albumin, respectively. This may indicate that the increased filtration of albumin in the glomerulus, potentially as a result of endothelial damage, and not albuminuria per se may link albuminuria to hypertension and increased CVD risk. Fourth, albuminuria is a hallmark of diabetic kidney disease and associated with unfavorable outcomes. Understanding pathways underlying albuminuria in diabetes may therefore be of particular relevance, and the four novel diabetes-specific loci identified in our study may represent a first step into this direction. Lastly, translation of GWAS loci into differential plasma protein levels as observed for OAF is of particular interest, as plasma protein levels represent both potential biomarkers and interventional targets.

Strengths of our study include its standardized approach to phenotype definition, its large samples size, internal locus validation, and the study of participants with diabetes. The identification of a previous Amerindian-specific locus[25] in our trans-ethnic analysis underscores the value of studying diverse ancestries, but EA individuals are still strongly overrepresented, which limits the power to detect heterogeneity correlated with ancestry. Limitations that are not specific to our study are related to the accurate quantification of UACR, which is influenced by biologic variation of urinary albumin, by the sensitivity and variation of albumin assays, and by standardization to urinary creatinine to account for urine dilution[23]. We addressed these issues by harmonizing UACR calculation across cohorts, and by separate assessment of associations with urinary albumin and creatinine. Across-cohort variation was overcome to some degree by the use of a central lab in the large UKBB, but may also introduce findings related to UKBB-specific sample handling, storage, measurement, or exposures. The statistical fine-mapping focused on SNPs available in the majority of studies, which might have limited the discovery of novel associations or the fine-mapping of population-specific or low-frequency variants. Such analyses represent avenues for future research. Other fine-mapping methods such as Bayesian approaches that incorporate priors based on variant annotation exist, but ultimately all statistically prioritized variants need to be experimentally validated.

In summary, we identified and characterized 68 loci associated with UACR and highlight potential causal genes, driver variants, target tissues, and pathways. These findings will inform experimental studies and advance the understanding of albuminuria and correlated traits, an essential step for the development of novel therapies to reduce the burden of CKD and potentially CVD.

Methods

We set up a collaborative meta-analysis based on a distributive data model. An analysis plan was developed and circulated to all participating studies via a Wiki system [https://ckdgen.eurac.edu/mediawiki/index.php/CKDGen_Round_4_EPACTS_analysis_plan]. Phenotypes were generated and quality checks performed within each study in a standardized manner through scripts provided to all study centers. Before conducting the analyses, studies uploaded automatically generated PDF and text files. After approval of the phenotype quality, ancestry-specific GWAS were performed in each study and uploaded centrally. Files were quality controlled using GWAtoolbox[57] and customized scripts, harmonized, and meta-analyzed. Details regarding each step are provided below. Each study was approved by the respective ethics committee, and all participants provided written informed consent. Drosophila research was carried out in compliance with all relevant ethical regulations. Drosophila experiments are exempt from a specific regulatory approval.

Phenotype definition. Methods for the measurement of urinary albumin and creatinine in each study are reported in Supplementary Data 1. Urinary albumin values below the detection limit of the used assays were set to the lower limit of detection, and the UACR was assessed in mg/g and calculated as urinary albumin (mg/l)/urinary creatinine (mg/dl) × 100. MA cases were defined as UACR > 30, and controls as UACR < 10 mg/g, no other exclusions were applied. These steps were all included in the distributed phenotyping script. MA GWAS analyses were limited to studies with ≥100 MA cases.

GWAS in individual studies. In each study, genotyping was performed using genome-wide arrays followed by application of study-specific quality filters prior to phasing and imputation. Genome-wide data were imputed to the Haplotype Reference Consortium (HRC) version 1.1, 1000 Genomes Project (1000G) phase 3 v5 ALL, or the 1000G phase 1 v3 ALL reference panels using the Sanger [https://imputation.sanger.ac.uk/] and Michigan Imputation Server [https://imputationserver.sph.umich.edu/]. Detailed information on study-specific genotyping, imputation, and QC is provided in Supplementary Data 2. Unless indicated differently, variants are annotated according to the GRCh37 (hg19) reference build.

The inverse normal transformed age-adjusted and sex-adjusted residuals of log-transformed UACR, as well as urinary albumin and urinary creatinine levels separately for the sensitivity analysis in the UKBB sample, were used as the dependent variable in a linear regression model fitted in each study-specific GWAS. For MA, a logistic regression model adjusted for sex and age was used. The models were adjusted for study-specific covariates, such as recruitment site and genetic principal components where applicable. Family-based studies used mixed-effect models by including the relationship of the individuals as a variance component. Additive genetic models were fitted using the SNP's allele dosage as an independent variable. The analysis programs used for the GWAS are provided in Supplementary Data 2.

GWAS meta-analysis. For UACR, studies contributed a total of 54 GWAS summary statistics files. After QC, the total samples size was 564,257 (547,361 individuals of European ancestry [EA], 6324 of East Asian ancestry [EAS], 6795 African Americans [AA], 2335 of South Asian ancestry [SA], and 1442 Hispanics; Supplementary Data 1). For MA, a total of 38 GWAS summary files were contributed, totaling a post-QC samples size of 348,954 (51,861 cases; Supplementary Data 1). Both meta-analyses included individuals with and without diabetes.

Before meta-analysis, study-specific GWAS files were filtered to retain only SNPs with imputation quality (IQ) score > 0.6 and MAC > 10, effective sample size ≥ 100, and a $|\text{beta}| < 10$ to remove implausible outliers. Within study, we estimated the genomic inflation factor λ_{GC} and applied GC correction when λ_{GC} was >1. Fixed effects inverse-variance weighted meta-analysis of the study-specific GWAS result files was performed using METAL[58], which was adapted to obtain effects and standard errors of higher precision if required (seven decimal places instead of four). After meta-analysis of 37,915,339 SNPs, we retained only variants that were present in ≥50% of the GWAS data files (27 studies) and had a total MAC of ≥400. Across ancestries, this yielded 8,034,757 variants for UACR (8,603,712 in EA with an observed MAF > 0.3%), and 8,326,000 variants for MA.

The inflation of *p*-values attributed to reasons other than polygenicity was assessed by LD score regression.[59] The intercept was estimated as 0.95, and thus ≤1, indicating that any residual inflation was likely due to polygenicity rather than confounding. Therefore, *p*-values were not corrected for a second round of genomic control after the meta-analysis.

The genome-wide significance level was set at 5×10^{-8}. Between-study heterogeneity was assessed using the I^2 statistic[60]. Variants were assigned to loci by selecting the SNP with the lowest *p*-value genome-wide as the index SNP, defining the corresponding locus as the ±500 kb region around it, and repeating the procedure until no further genome-wide significant SNP remained. A locus was considered novel if it did not contain any variant identified by previous GWAS of UACR. The loci were named according to the nearest gene of the index SNP, the SNP with the lowest *p*-value within a locus.

For UACR, we evaluated heterogeneity correlated with ancestry using study-specific GWAS files filtered for polymorphic SNPs with an IQ score > 0.3, an effective sample size ≥ 100, and a $|\text{beta}| < 10$. Analysis was performed using the software Meta-Regression of Multi-Ethnic Genetic Association (MR-MEGA v0.1.2.25)[28], where the meta-regression model included the three axes explaining the largest genetic variation estimated from allele frequencies provided in the study-specific GWAS files.

The narrow-sense heritability of the trait based on all SNPs with a MAF > 1% was estimated using the genome-wide summary statistics for UACR with the MHC region removed as input for the LD score regression software[59], using the 1000 Genomes phase 3 EUR reference panel for estimating LD. The proportion of phenotypic variance explained by the index SNPs was estimated as $\beta^2{*}2{*}\text{MAF}{*}(1-\text{MAF})$, with β representing the SNP effect and accounting for a trait variance of 1 due to the inverse normal transformation of the analyzed trait. Thus, the estimates provide the proportion of the variance of sex- and age-adjusted log-transformed UACR that is explained by the respective SNPs. The expected number of discoveries in future, larger studies and the corresponding percentage of GWAS heritability explained with increases in sample size was estimated using a recently published method[31]. The summary statistics of the UACR trans-ethnic meta-analysis were used as input.

Functional enrichment. We used DEPICT[36] version 1 release 194 to identify gene sets and tissue/cell types enriched in UACR-associated loci. DEPICT performs gene set and tissue-/cell-type enrichment analysis by testing whether genes in GWAS-associated loci are enriched in 14,461 reconstituted gene sets. These reconstituted gene sets were generated based on a large number of predefined gene sets from diverse molecular pathway databases including protein–protein interactions, and gene sets from mouse gene knockout studies. The function of each gene in 14,461 reconstituted gene sets was predicted from co-regulation analyses of 77,840 expression microarray samples. Tissues and cell-type enrichment was conducted in DEPICT by testing whether the genes in associated regions were highly expressed in any of 209 MeSH annotations for 37,427 microarrays. We included all variants that reached a genome-wide significant *p*-value of association with UACR ($p < 5 \times 10^{-8}$) from the trans-ethnic meta-analysis. DEPICT analysis was conducted with 500 repetitions to compute FDR and 5000 permutations to compute enrichment test *p*-values adjusted for gene length by using 500 null GWAS.

Phenome-wide association study. All analyses were conducted using standard PheWAS coding methodologies[37] using the R-package "PheWAS". Models were adjusted for ten genetic principal components and sex, when appropriate. All analyses were conducted among 192,868 participants of European ancestry in the Million Veteran Program sample. A weighted genetic risk score was first built using the 59 UACR-associated SNPs (Supplementary Data 3) where the UACR-increasing allele was coded as the effect allele. Based on the number of covariates included in the model, only traits with ≥100 cases were included in the analysis resulting in evaluation of 1422 traits. A Bonferroni threshold of 3.5×10^{-5} (0.05/1422) was applied for assessing significance of the association test.

The genetic UACR risk score was also tested for association with additional outcomes using GWAS summary statistics with association testing implemented in the function *grs.summary()* of the R-package "gtx". The summary statistics for hypertension and heart failure were calculated in the UKBB prior to the risk score association analysis. Hypertension cases were defined based on ICD-10 codes (I10, I11, I11.0, I11.9, I12, I12.0, I12.9, I13, I13.0, I13.1, I13.2, I15, I15.0, I15.1, I15.2, I15.8, and I15.9), as self-reported hypertension or essential hypertension, by measured systolic blood pressure ≥ 140 mmHg, diastolic blood pressure ≥ 90 mmHg, or by taking blood pressure medication. Hear failure cases were defined based on ICD-10 codes (I11.0, I13.0, I13.2, I25.5, I42.0, I42.5, I42.8, I42.9, I50, I50.0, I50.1, and I50.9), or by self-reported cardiomyopathy, excluding hypertrophic cardiomyopathy. The summary statistics for other outcomes were based on results from published GWAS meta-analyses with references provided in Supplementary Table 2. Statistical significance was defined as $p < 0.007$ of the association test after correction for the number of evaluated associations (0.05/7).

Genetic correlation with other traits. Genome-wide genetic correlations between UACR and UK Biobank traits and diseases were evaluated to investigate whether there was evidence of co-regulation or a shared genetic basis, both known and novel. Using LD score regression that can account for overlapping samples[61] and the EA association summary statistics as input, we evaluated pair-wise genetic correlations between UACR and each of 517 pre-computed GWAS summary statistics of UKBB traits and diseases available through the web-platform LDHub. An overview of the sources of these summary statistics and their corresponding sample sizes is available at [http://ldsc.broadinstitute.org]. Statistical significance was assessed at the Bonferroni corrected level of 9.7×10^{-5} (0.05/517).

Second signals within identified loci. To identify additional, independent UACR-associated variants within the identified loci, approximate conditional analyses were carried out that incorporated LD information from an ancestry-matched reference population. We used the genome-wide UACR summary statistics from the EA meta-analysis as input, because an LD reference sample scaled to the size of our meta-analysis was only available for EA individuals[43,44]. We randomly selected 15,000 participants from the UK Biobank data set (UKBB; application ID 2027, data set ID 8974). Individuals who withdrew consent and those not meeting data cleaning requirements were excluded, keeping only those who passed sex check, had a genotyping call rate of ≥95%, and did not represent outliers with respect to SNP heterozygosity. For each pair of individuals, the proportion of variants shared identical-by-descent (IBD) was computed using PLINK [https://www.cog-genomics.org/plink/]. We retained only one member of each pair with an IBD coefficient of ≥0.1875. Individuals were restricted to those of EA by excluding outliers along the first two PCs from a principal component analysis using the HapMap phase 3 release 2 populations as reference. The final data set to estimate LD included 13,558 EA individuals and 16,969,363 SNPs.

Basis for statistical fine-mapping were the 61 1-Mb genome-wide significant loci identified in the EA meta-analysis, clipping at chromosome borders. Overlapping loci as well as pairs of loci whose respective index SNPs were correlated ($r^2 > 0.1$ in the UKBB data set described above) were merged, resulting in a final list of 57 regions prior to fine-mapping. Within each region, the GCTA stepwise model selection procedure (cojo-slct algorithm) was used to identify independent variants employing a stepwise forward selection approach[44]. We used the default collinearity cutoff of 0.9 and set the significance threshold to identify independent SNPs to 5×10^{-8}.

Estimation of credible sets. Statistical fine-mapping was carried out for each of the 57 merged regions used as input for GCTA cojo-slct. For each region that contained multiple independent SNPs identified by the GCTA stepwise forward selection approach, approximate conditional analyses conditioned on all remaining independent SNP of this region were carried out using the GCTA cojo-cond algorithm to estimate conditional effect sizes. The derived effect estimates were used in the Wakefield's formula as implemented in the R-package'gtx' version 2.0.1 [https://github.com/tobyjohnson/gtx] to derive approximate Bayes factors (ABF) from conditional estimates in regions with multiple independent SNPs, and from the original estimates for regions with a single independent SNP. Given that 95% of the SNP effects from the UACR GWAS were within ±0.03, the standard deviation prior was chosen as 0.0153 based on formula (8) in the original publication[45]. For each variant within an evaluated region, the Approximate Bayes Factor obtained from the effect and its standard error of the marginal (single signal region) or conditional estimates (multi-signal regions) was used to calculate the PP for the variant driving the association signal (causal variant). For each region, 99% credible sets, representing the set of SNPs that contain with a 99% PP the variant causing the association, were calculated by summing up the PP-ranked variants until the cumulative PP was >99%.

Functional annotation of identified variants. Functional annotations of index variants of associated loci and credible set variants were performed by querying the SNiPA database v3.2 (March 2017) [https://snipa.helmholtz-muenchen.de/snipa/]. SNiPA includes extensive annotations ranging from regulatory elements, over gene annotations to variant annotations and published GWAS associations. SNiPA release v3.2 is based on 1000 the Genomes phase 3 version 5 and Ensembl version 87 data sets. The Ensembl VEP tool [https://www.ensembl.org/info/docs/tools/vep/] was used for primary effect prediction of SNPs. The CADD score[62] provided by SNiPA is based on CADD release v1.3 and presented as PHRED-like transformation of the C score.

Co-localization of UACR and cis-eQTL associations. Co-localization analysis was based on the genetic associations with UACR in the EA sample (because the great majority of gene expression data sets was generated from EA). Gene expression was quantified from microdissected human glomerular and tubulointerstitial kidney portions from 187 individuals participating in the NEPTUNE study[48], as well as from the 44 tissues included in the GTEx Project version 6p release [https://gtexportal.org/]. The eQTL and GWAS effect alleles were harmonized. For each locus, we identified tissue–gene pairs with reported eQTL data within ±100 kb of each GWAS index variant. The region for each co-localization test was defined as the eQTL *cis* window defined in the underlying GTEx and NephQTL studies. We used the default parameters and prior definitions set in the "coloc.fast" function from the R-package "gtx" version 2.0.1 [https://github.com/tobyjohnson/gtx], which is an adapted implementation of Giambartolomei's co-localization method[63]. The same package was also used to estimate the direction of effect as the ratio of the average PP (that was obtained from credible set estimation) weighted GWAS effects over the PP weighted eQTL effects.

An additional co-localization analysis was performed using a complementary gene-expression data set derived from healthy human kidney tissue. The corresponding eQTL data set was generated by correlating genotype with RNA-seq-based gene expression levels from 96 human kidney samples[47]. Co-localization analysis of GWAS signals and eQTL signals was performed using *Coloc*[63], using the same distance criteria to identify shared eQTL and GWAS regions as above, including variants within the *cis*-window (±1 Mb from TSS) of each identified candidate gene, and the parameters $p1 = 1 \times 10^{-4}$, $p2 = 1 \times 10^{-4}$, and $p12 = 1 \times 10^{-5}$.

For all co-localization analyses, a PP ≥ 0.8 of the H4 test (one common causal variant underlying UACR and eQTL association signal) was applied to select a significant result.

Trans-eQTL analysis. We performed *trans*-eQTL annotation through LD mapping based on the 1000 Genomes phase 3 version 5 European reference panel with a r^2 cutoff of >0.8. We limited annotation to index SNPs with a fine-mapping PP ≥1% in at least one fine-mapped-region. Due to expected small effect sizes, only available genome-wide *trans*-eQTL studies of either peripheral blood mononuclear cells or whole blood with a sample size of ≥1000 individuals were considered, resulting in five non-overlapping studies[64–68]. For the study by Kirsten et al.[68], we had access to an update with larger sample size combining two nonoverlapping studies (LIFE-Heart and LIFE-Adult) resulting in a total sample size of 6645. To improve stringency of results, we focused the analysis on inter-chromosomal *trans*-eQTLs with association test *p*-values of $p < 5 \times 10^{-8}$ reported by ≥2 studies (Supplementary Table 4).

pQTL lookup and co-localization. The pQTL data were generated using an aptamer-based multiplex protein assay (SOMAscan) to quantify 3622 proteins from stored EDTA plasma of 3301 healthy participants of the INTERVAL study, which were genotyped on the Affymetrix Axiom UK Biobank genotyping array and imputed to a combined 1000 Genomes Phase 3-UK10K reference panel[50]. For this lookup, all pQTLs with $p < 1 \times 10^{-4}$ were selected.

Co-localization analysis for pQTL data was performed using the same analysis approach as described for eQTL co-localization. For associations with plasma protein concentrations, pQTL results of 1927 genetic associations with 1478 proteins obtained by the Somalogic proteomics platform GWAS[50] were included. In a first instance, pQTLs within a ± 500 kb region of each UACR-associated SNP (Supplementary Data 5) were identified. In case a pQTL region contained multiple independent index SNPs, additional pQTLs were calculated conditioning on the respective index SNP. Next, the conditional and unconditional pQTLs ($n = 38$) were included in the co-localization analysis using the coloc.abf() function with default priors of the R-package "coloc" implementing the co-localization method of Giambartolomei[63].

The intra-assay coefficient of variation for the OAF protein, for which evidence for co-localization of the UACR association and OAF plasma levels was identified, was 5.7% and 16.9% in the two batches of SOMAscan measurements[50].

Drosophila experiments. Transgenic RNAi studies were performed using the UAS/GAL4 system, flies were raised on standard agar cornmeal molasses. RNAi crosses were grown at 30 °C. The RNAi stocks were obtained from the Bloomington Drosophila Stock Center at Indiana University (*oaf*-RNAi-1 BDSC #40926, *aPKC*-RNAi-1 BDSC # 35001, *aPKC*-RNAi-2 BDSC #34332) or the Vienna Drosophila Resource Center respectively (*oaf*-RNAi-2 VDRC #38257, *Vrp1*-RNAi-1 VDRC #102253, *Vrp1*-RNAi-2 VDRC #23888). Control RNAi was directed against EGFP (BSDC# 41553). *Dorothy*-GAL4 (BDSC #6903) was used to drive expression in nephrocytes.

To perform the FITC-albumin endocytosis assay, garland cell nephrocytes were dissected from wandering third instar larvae in PBS and incubated with 0.2 mg/ml FITC-albumin (Sigma) for 30 s. Cells were rinsed briefly with ice-cold PBS four times and fixed immediately for 5 min in 8% paraformaldehyde in presence of Hoechst 33342 (1:1000). Cells were mounted in Roti-Mount FluorCare (Carl Roth GmbH) and imaged using a Zeiss LSM 880 confocal microscope. Quantification of fluorescent tracer uptake was performed with ImageJ software. Average fluorescence of the three brightest cells was measured and intensity of the background subtracted. The results are expressed as a ratio to a control experiment with EGFP-RNAi that was performed in parallel.

For immunohistochemistry, garland cell nephrocytes were dissected from wandering third instar larvae, fixed for 20 min in PBS containing 4% paraformaldehyde, and stained according to the standard procedure. The following primary antibodies were used: rabbit anti-sns (1:500, gift from S. Abmayr), guinea pig anti-Kirre (1:200, gift from S. Abmayr), and rabbit anti anti-PKCζ (C20) (1:200, sc-216-G, Santa Cruz Biotechnology) that was previously shown to detect *Drosophila* aPKC[69]. For imaging, a Zeiss LSM 880 confocal microscope was used. Image processing was done by ImageJ and Gimp software. Three-dimensional reconstruction of confocal images was done using Imaris software.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Summary genetic association results are freely available on the CKDGen Consortium website [https://ckdgen.imbi.uni-freiburg.de/]. The source data underlying Figs. 1, 2, 5–8 and Supplementary Figs. 8 and 9 are provided as a Source Data file. The source data underlying Figs. 3, 4, and Supplementary Fig. 7 are provided in Supplementary Data 9, 10, and 7, respectively, and the data underlying the Supplementary Figs. 2–6 are based on the respective downloadable summary genetic association results.

Code availability

The script for generating the phenotypes used in the GWAS is available via GitHub [https://github.com/genepi-freiburg/ckdgen-pheno].

Received: 7 December 2018 Accepted: 23 July 2019
Published online: 11 September 2019

References

1. Hemmelgarn, B. R. et al. Relation between kidney function, proteinuria, and adverse outcomes. *JAMA* **303**, 423–429 (2010).
2. Bello, A. K. et al. Associations among estimated glomerular filtration rate, proteinuria, and adverse cardiovascular outcomes. *Clin. J. Am. Soc. Nephrol.* **6**, 1418–1426 (2011).
3. Gansevoort, R. T. et al. Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. *Kidney Int.* **80**, 93–104 (2011).
4. Matsushita, K. et al. Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: a collaborative meta-analysis of individual participant data. *Lancet Diabetes Endocrinol.* **3**, 514–525 (2015).
5. Chronic Kidney Disease Prognosis Consortium et al. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. *Lancet* **375**, 2073–2081 (2010).
6. Inker, L. A. et al. KDOQI US commentary on the 2012 KDIGO clinical practice guideline for the evaluation and management of CKD. *Am. J. Kidney Dis.* **63**, 713–735 (2014).
7. Eckardt, K.-U. et al. Evolving importance of kidney disease: from subspecialty to global health burden. *Lancet* **382**, 158–169 (2013).
8. Tuttle, K. R. et al. Diabetic kidney disease: a report from an ADA Consensus Conference. *Am. J. Kidney Dis.* **64**, 510–533 (2014).

9. Brenner, B. M. et al. Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N. Engl. J. Med.* **345**, 861–869 (2001).
10. de Zeeuw, D. et al. Albuminuria, a therapeutic target for cardiovascular protection in type 2 diabetic patients with nephropathy. *Circulation* **110**, 921–927 (2004).
11. Lea, J. et al. The relationship between magnitude of proteinuria reduction and risk of end-stage renal disease: results of the African American study of kidney disease and hypertension. *Arch. Intern. Med.* **165**, 947–953 (2005).
12. Heerspink, H. J. L., Kröpelin, T. F., Hoekman, J. & de Zeeuw, D. & Reducing Albuminuria as Surrogate Endpoint (REASSURE) Consortium. Drug-induced reduction in albuminuria is associated with subsequent renoprotection: a meta-analysis. *J. Am. Soc. Nephrol.* **26**, 2055–2064 (2015).
13. Ibsen, H. et al. Reduction in albuminuria translates to reduction in cardiovascular events in hypertensive patients: losartan intervention for endpoint reduction in hypertension study. *Hypertens.* **45**, 198–202 (2005).
14. Asselbergs, F. W. et al. Effects of fosinopril and pravastatin on cardiovascular events in subjects with microalbuminuria. *Circulation* **110**, 2809–2816 (2004).
15. Holtkamp, F. A. et al. Albuminuria and blood pressure, independent targets for cardioprotective therapy in patients with diabetes and nephropathy: a post hoc analysis of the combined RENAAL and IDNT trials. *Eur. Heart J.* **32**, 1493–1499 (2011).
16. Duffy, D. L. et al. Familial aggregation of albuminuria and arterial hypertension in an Aboriginal Australian community and the contribution of variants in ACE and TP53. *BMC Nephrol.* **17**, 183 (2016).
17. MacCluer, J. W. et al. Heritability of measures of kidney disease among Zuni Indians: the Zuni Kidney Project. *Am. J. Kidney Dis.* **56**, 289–302 (2010).
18. Forsblom, C. M., Kanninen, T., Lehtovirta, M., Saloranta, C. & Groop, L. C. Heritability of albumin excretion rate in families of patients with Type II diabetes. *Diabetologia* **42**, 1359–1366 (1999).
19. Fox, C. S. et al. Genome-wide linkage analysis to urinary microalbuminuria in a community-based sample: the Framingham Heart Study. *Kidney Int.* **67**, 70–74 (2005).
20. Langefeld, C. D. et al. Heritability of GFR and albuminuria in Caucasians with type 2 diabetes mellitus. *Am. J. Kidney Dis.* **43**, 796–800 (2004).
21. Pattaro, C. Genome-wide association studies of albuminuria: towards genetic stratification in diabetes? *J. Nephrol.* **31**, 475–487 (2018).
22. Böger, C. A. et al. CUBN is a gene locus for albuminuria. *J. Am. Soc. Nephrol.* **22**, 555–570 (2011).
23. Teumer, A. et al. Genome-wide association studies identify genetic loci associated with albuminuria in diabetes. *Diabetes* **65**, 803–817 (2016).
24. Kramer, H. J. et al. African ancestry-specific alleles and kidney disease risk in Hispanics/Latinos. *J. Am. Soc. Nephrol.* **28**, 915–922 (2017).
25. Brown, L. A. et al. Admixture mapping identifies an Amerindian ancestry locus associated with albuminuria in hispanics in the United States. *J. Am. Soc. Nephrol.* **28**, 2211–2220 (2017).
26. Sandholm, N. et al. Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes. *Diabetologia* **57**, 1143–1153 (2014).
27. Haas, M. E. et al. Genetic association of albuminuria with cardiometabolic disease and blood pressure. *Am. J. Hum. Genet.* **103**, 461–473 (2018).
28. Mägi, R. et al. Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution. *Hum. Mol. Genet.* **26**, 3639–3650 (2017).
29. Voskarides, K. et al. COL4A3/COL4A4 mutations producing focal segmental glomerulosclerosis and renal failure in thin basement membrane nephropathy. *J. Am. Soc. Nephrol.* **18**, 3004–3016 (2007).
30. Malone, A. F. et al. Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. *Kidney Int.* **86**, 1253–1259 (2014).
31. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).
32. Genovese, G. et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
33. Naik, R. P. et al. Association of sickle cell trait with chronic kidney disease and albuminuria in African Americans. *JAMA* **312**, 2115–2125 (2014).
34. Peiris, H. et al. Discovering human diabetes-risk gene function with genetics and physiological assays. *Nat. Commun.* **9**, 3855 (2018).
35. Sevilla, L. M., Nachat, R., Groot, K. R. & Watt, F. M. Kazrin regulates keratinocyte cytoskeletal networks, intercellular junctions and differentiation. *J. Cell Sci.* **121**, 3561–3569 (2008).
36. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
37. Denny, J. C. et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1110 (2013).
38. Sundin, O. H. et al. A common locus for late-onset Fuchs corneal dystrophy maps to 18q21.2-q21.32. *Invest. Ophthalmol. Vis. Sci.* **47**, 3919–3926 (2006).
39. Afshari, N. A. et al. Genome-wide association study identifies three novel loci in Fuchs endothelial corneal dystrophy. *Nat. Commun.* **8**, 14898 (2017).
40. Brantsma, A. H. et al. Urinary albumin excretion and its relation with C-reactive protein and the metabolic syndrome in the prediction of type 2 diabetes. *Diabetes Care* **28**, 2525–2530 (2005).
41. Wang, T. J. et al. Low-grade albuminuria and the risks of hypertension and blood pressure progression. *Circulation* **111**, 1370–1376 (2005).
42. Gerstein, H. C. et al. Albuminuria and risk of cardiovascular events, death, and heart failure in diabetic and nondiabetic individuals. *JAMA* **286**, 421–426 (2001).
43. Benner, C. et al. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. *Am. J. Hum. Genet.* **101**, 539–551 (2017).
44. Yang, J. et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
45. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
46. Sieber, K. B. et al. Integrated functional genomic analysis enables annotation of kidney genome-wide association study loci. *J. Am. Soc. Nephrol.* **30**, 421–441 (2019).
47. Ko, Y. A. et al. Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease. *Am. J. Hum. Genet.* **100**, 940–953 (2017).
48. Gillies, C. E. et al. An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* **103**, 232–244 (2018).
49. Kirby, A. et al. Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat. Genet.* **45**, 299–303 (2013).
50. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
51. Helmstädter, M., Huber, T. B. & Hermle, T. Using the Drosophila Nephrocyte to model podocyte function and disease. *Front. Pediatr.* **5**, 262 (2017).
52. Hermle, T., Braun, D. A., Helmstädter, M., Huber, T. B. & Hildebrandt, F. Modeling monogenic human nephrotic syndrome in the Drosophila garland cell nephrocyte. *J. Am. Soc. Nephrol.* **28**, 1521–1533 (2017).
53. Huber, T. B. et al. Loss of podocyte aPKClambda/iota causes polarity defects and nephrotic syndrome. *J. Am. Soc. Nephrol.* **20**, 798–806 (2009).
54. Amsellem, S. et al. Cubilin is essential for albumin reabsorption in the renal proximal tubule. *J. Am. Soc. Nephrol.* **21**, 1859–1867 (2010).
55. Meng, X.-M., Nikolic-Paterson, D. J. & Lan, H. Y. TGF-β: the master regulator of fibrosis. *Nat. Rev. Nephrol.* **12**, 325–338 (2016).
56. Blobe, G. C., Schiemann, W. P. & Lodish, H. F. Role of transforming growth factor beta in human disease. *N. Engl. J. Med.* **342**, 1350–1358 (2000).
57. Fuchsberger, C., Taliun, D., Pramstaller, P. P. & Pattaro, C., CKDGen consortium. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. *Bioinformatics* **28**, 444–445 (2012).
58. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
59. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
60. Higgins, J. P. T., Thompson, S. G., Deeks, J. J. & Altman, D. G. Measuring inconsistency in meta-analyses. *BMJ* **327**, 557–560 (2003).
61. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).
62. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
63. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
64. Zeller, T. et al. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5**, e10693 (2010).
65. Fehrmann, R. S. N. et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7**, e1002197 (2011).
66. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
67. Joehanes, R. et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol.* **18**, 16 (2017).

68. Kirsten, H. et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Hum. Mol. Genet.* **24**, 4746–4763 (2015).

69. Wodarz, A., Ramrath, A., Grimm, A. & Knust, E. Drosophila atypical protein kinase C associates with Bazooka and controls polarity of epithelia and neuroblasts. *J. Cell Biol.* **150**, 1361–1374 (2000).

Acknowledgements

We thank Daniele Di Domizio (Eurac Research) and Jochen Knaus (University of Freiburg) for IT assistance, Toby Johnson (GSK) for sharing his code and helpful discussions related to the fine-mapping and co-localization analyses, and Roland Nitschke, Life Imaging Centre, University of Freiburg, for help with confocal microscopy. We thank Lina L. Kampf for technical assistance with the *Drosophila* experiments and Jan Halbritter for work related to the recruitment of the Sorbs cohort. This research has been conducted using the UK Biobank Resource under application numbers 20272, 10019, and 31852. Study-specific acknowledgements, consortium members, and funding sources are listed in the Supplementary Note 1.

Author contributions

Manuscript writing group: Alexander Teumer, Yong Li, Sahar Ghasemi, Bram P. Prins, Matthias Wuttke, Tobias Hermle, Nisha Bansal, Harold Snieder, Adam S. Butterworth, Adriana M. Hung, Cristian Pattaro, Anna Köttgen. Design of this study: Alexander Teumer, Matthias Wuttke, Tobias Hermle, Karsten B. Sieber, Adrienne Tin, Mathias Gorski, Christian Fuchsberger, Carsten A. Böger, Andrew P. Morris, Markus Scholz, Adam S. Butterworth, Cristian Pattaro, Anna Köttgen. Bioinformatics: Shreeram Akilesh, Daniela Baptista, Carsten A. Böger, Robert J. Carroll, Audrey Y. Chu, Massimiliano Cocca, Tanguy Corre, Frauke Degenhardt, Jasmin Divers, Georg Ehret, Andre Franke, Sahar Ghasemi, Ayush Giri, Mathias Gorski, Franziska Grundner-Culemann, Pavel Hamet, Iris M. Heid, Anselm Hoppmann, Katrin Horn, Johanna Jakobsdottir, Navya Shilpa Josyula, Chiea-Chuen Khor, Holger Kirsten, Anna Köttgen, Carl D. Langefeld, Man Li, Yong Li, Jianjun Liu, Leo-Pekka Lyytikäinen, Jonathan Marten, Dennis O. Mook-Kanamori, Peter J. van der Most, Raymond Noordam, Teresa Nutile, Sarah A. Pendergrass, Anna I. Podgornaia, Chengxiang Qiu, Markus Scholz, Sanaz Sedaghat, Christian M. Shaffer, Karsten B. Sieber, Albert V. Smith, Silke Szymczak, Alexander Teumer, Hauke Thomsen, Johanne Tremblay, Chaolong Wang, Matthias Wuttke, Yizhe Xu, Zhi Yu. Functional analysis of candidate genes in *Drosophila*: Tobias Hermle, Mengmeng Chen, Lea Gerstner. Genotyping: Daniela Baptista, Ralph Burkhardt, Carsten A. Böger, Ching-Yu Cheng, Georg Ehret, Mary F. Feitosa, Andre Franke, Christian Fuchsberger, Ron T. Gansevoort, Pavel Hamet, Pim van der Harst, Erik Ingelsson, Chiea-Chuen Khor, Wolfgang Koenig, Peter Kovacs, Florian Kronenberg, Mika Kähönen, Antje Körner, Leslie A. Lange, Terho Lehtimäki, Leo-Pekka Lyytikäinen, Thomas Meitinger, Dennis O. Mook-Kanamori, Andrew P. Morris, Josyf C. Mychaleckyj, Martina Müller-Nurasyid, Nicholette D. Palmer, Dermot F. Reilly, Fernando Rivadeneira, Jerome I. Rotter, Kent D. Taylor, Alexander Teumer, Hauke Thomsen, Johanne Tremblay, André G. Uitterlinden, Uwe Völker, Melanie Waldenberger, Chaolong Wang, Lihua Wang, James G. Wilson, Johan Ärnlöv. Interpretation of Results: Adam S. Butterworth, Carsten A. Böger, Ching-Yu Cheng, Katalin Dittrich, Jasmin Divers, Karlhans Endlich, Mary F. Feitosa, Janine F. Felix, Barry I. Freedman, Sahar Ghasemi, Ayush Giri, Mathias Gorski, Pavel Hamet, Pim van der Harst, Iris M. Heid, Kevin Ho, Katrin Horn, Shih-Jen Hwang, Bettina Jung, Holger Kirsten, Wolfgang Koenig, Anna Köttgen, Carl D. Langefeld, Man Li, Yong Li, Jonathan Marten, Kozeta Miliku, Andrew P. Morris, Nicholette D. Palmer, Cristian Pattaro, Sarah A. Pendergrass, Bram P. Prins, Dermot F. Reilly, Myriam Rheinberger, Markus Scholz, Sanaz Sedaghat, Karsten B. Sieber, Bamidele O. Tayo, Alexander Teumer, Hauke Thomsen, Adrienne Tin, Johanne Tremblay, André G. Uitterlinden, Niek Verweij, Suzanne Vogelezang, Matthias Wuttke, Yizhe Xu, Masayuki Yasuda. Management of an individual contributing study: Shreeram Akilesh, Stephan J.L. Bakker, Murielle Bochud, Eric Boerwinkle, Martin H. de Borst, Hermann Brenner, Adam S. Butterworth, Carsten A. Böger, Robert J. Carroll, Ching-Yu Cheng, Josef Coresh, John Danesh, Olivier Devuyst, Katalin Dittrich, Kai-Uwe Eckardt, Georg Ehret, Janine F. Felix, Oscar H. Franco, Barry I. Freedman, Ron T. Gansevoort, Vilmantas Giedraitis, Alessandro De Grandi, Vilmundur Gudnason, Tamara B. Harris, Pim van der Harst, Andrew A. Hicks, Kevin Ho, Adriana M. Hung, M. Arfan Ikram, Erik Ingelsson, Vincent W.V. Jaddoe, Bettina Jung, Chiea-Chuen Khor, Wieland Kiess, Wolfgang Koenig, Holly Kramer, Florian Kronenberg, Bernhard K. Krämer, Mika Kähönen, Antje Körner, Anna Köttgen, Terho Lehtimäki, Yong Li, Wolfgang Lieb, Su-Chi Lim, Markus Loeffler, Deborah Mascalzoni, Barbara McMullen, Andrew P. Morris, Renée de Mutsert, Jeffrey O'Connell, Afshin Parsa, Cristian Pattaro, Sarah A. Pendergrass, Annette Peters, Belen Ponte, Peter P. Pramstaller, Bruce M. Psaty, Ton J. Rabelink, Dermot F. Reilly, Rainer Rettig, Myriam Rheinberger, Heiko Runz, Charumathi Sabanayagam, Kai-Uwe Saum, Markus Scholz, Ben Schöttker, Harold Snieder, Kari Stefansson, Konstantin Strauch, Michael Stumvoll, Gardar Sveinbjornsson, E-Shyong Tai, Bamidele O. Tayo, Yih-Chung Tham, Joachim Thiery, Adrienne Tin, Johanne Tremblay, Anke Tönjes, Aiko P.J. de Vries, Uwe Völker, James G. Wilson, Otis D. Wilson, Charlene Wong, Tien-Yin Wong, Matthias Wuttke, Qiong Yang, Masayuki Yasuda. Subject Recruitment: Shreeram Akilesh, Hermann Brenner, Carsten A. Böger, Miao-Ling Chee, Katalin Dittrich, Valencia Hui Xian Foo, Barry I. Freedman, Ron T. Gansevoort, Vilmundur Gudnason, Vincent W.V. Jaddoe, Bettina Jung, Florian Kronenberg, Mika Kähönen, Anna Köttgen, Jeannette Jen-Mai Lee, Terho Lehtimäki, Wolfgang Lieb, Lars Lind, Christa Meisinger, Renée de Mutsert, Kjell Nikus, Isleifur Olafsson, Cristian Pattaro, Sarah A. Pendergrass, Belen Ponte, Tanja Poulain, Ton J. Rabelink, Rainer Rettig, Myriam Rheinberger, Nicholas Y. Q. Tan, Andrej Teren, Yih-Chung Tham, Johanne Tremblay, Anke Tönjes, Suzanne Vogelezang, Aiko P.J. de Vries, James G. Wilson, Johan Ärnlöv. Statistical Methods and Analysis: Mary L. Biggs, Carsten A. Böger, Robert J. Carroll, Jin-Fang Chai, Miao-Li Chee, Audrey Y. Chu, Massimiliano Cocca, James P. Cook, Tanguy Corre, Jasmin Divers, Todd L. Edwards, Mary F. Feitosa, Janine F. Felix, Barry I. Freedman, Sandra Freitag-Wolf, Christian Fuchsberger, Sahar Ghasemi, Ayush Giri, Mathias Gorski, Daniel F. Gudbjartsson, Martin Gögele, Pavel Hamet, Pim van der Harst, Iris M. Heid, Anselm Hoppmann, Katrin Horn, Shih-Jen Hwang, Johanna Jakobsdottir, Navya Shilpa Josyula, Bettina Jung, Chiea-Chuen Khor, Holger Kirsten, Holly Kramer, Anna Köttgen, Leslie A. Lange, Carl D. Langefeld, Man Li, Yong Li, Jianjun Liu, Leo-Pekka Lyytikäinen, Anubha Mahajan, Joseph C. Maranville, Jonathan Marten, Kozeta Miliku, Andrew P. Morris, Peter J. van der Most, Matthias Nauck, Boting Ning, Damia Noce, Raymond Noordam, Teresa Nutile, Cristian Pattaro, Sarah A. Pendergrass, Bram P. Prins, Laura M. Raffield, Myriam Rheinberger, Kenneth M. Rice, Fernando Rivadeneira, Kathleen A. Ryan, Markus Scholz, Sanaz Sedaghat, Yuan Shi, Karsten B. Sieber, Albert V. Smith, Benjamin B. Sun, Katalin Susztak, Gardar Sveinbjornsson, Silke Szymczak, Bamidele O. Tayo, Alexander Teumer, Chris H.L. Thio, Hauke Thomsen, Johanne Tremblay, Niek Verweij, Suzanne Vogelezang, Chaolong Wang, Lihua Wang, Matthias Wuttke, Yizhe Xu, Qiong Yang. Critical review of manuscript: all authors.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-11576-0.

Competing interests: Karsten B. Sieber is full-time employee of GlaxoSmithKline. Gardar Sveinbjornsson, Daniel F. Gudbjartsson, Hilma Holm, Unnur Thorsteinsdottir and Kari Stefansson are full-time employees of deCODE genetics, Amgen Inc. John Danesh is member of the Novartis Cardiovascular and Metabolic Advisory Board, received grant support from Novartis. Oscar H. Franco works in ErasmusAGE, a center for aging research across the life course funded by Nestlé Nutrition (Nestec Ltd.), Metagenics Inc., and AXA. Wolfgang Koenig received modest consultation fees for advisory board meetings from Amgen, DalCor, Kowa, Novartis, Pfizer and Sanofi, and modest personal fees for lectures from Amgen, AstraZeneca, Novartis, Pfizer and Sanofi. Anna I. Podgornaia and Dermot F. Reilly are employees of Merck Sharp Dohme Corp., Whitehouse Station, NJ, USA. Kevin Ho disclosed a research and financial relationship with Sanofi-Genzyme. Bruce M. Psaty serves on the DSMB of a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. Markus Scholz: Consultancy of and grant support from Merck Serono not related to this project. Adam S. Butterworth received grants from MSD, Pfizer, Novartis, Biogen and Bioverativ and personal fees from Novartis. Anna Köttgen received grant support from Gruenenthal not related to this project. The other authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Peer review information *Nature Communications* thanks Julian Dow and the other anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

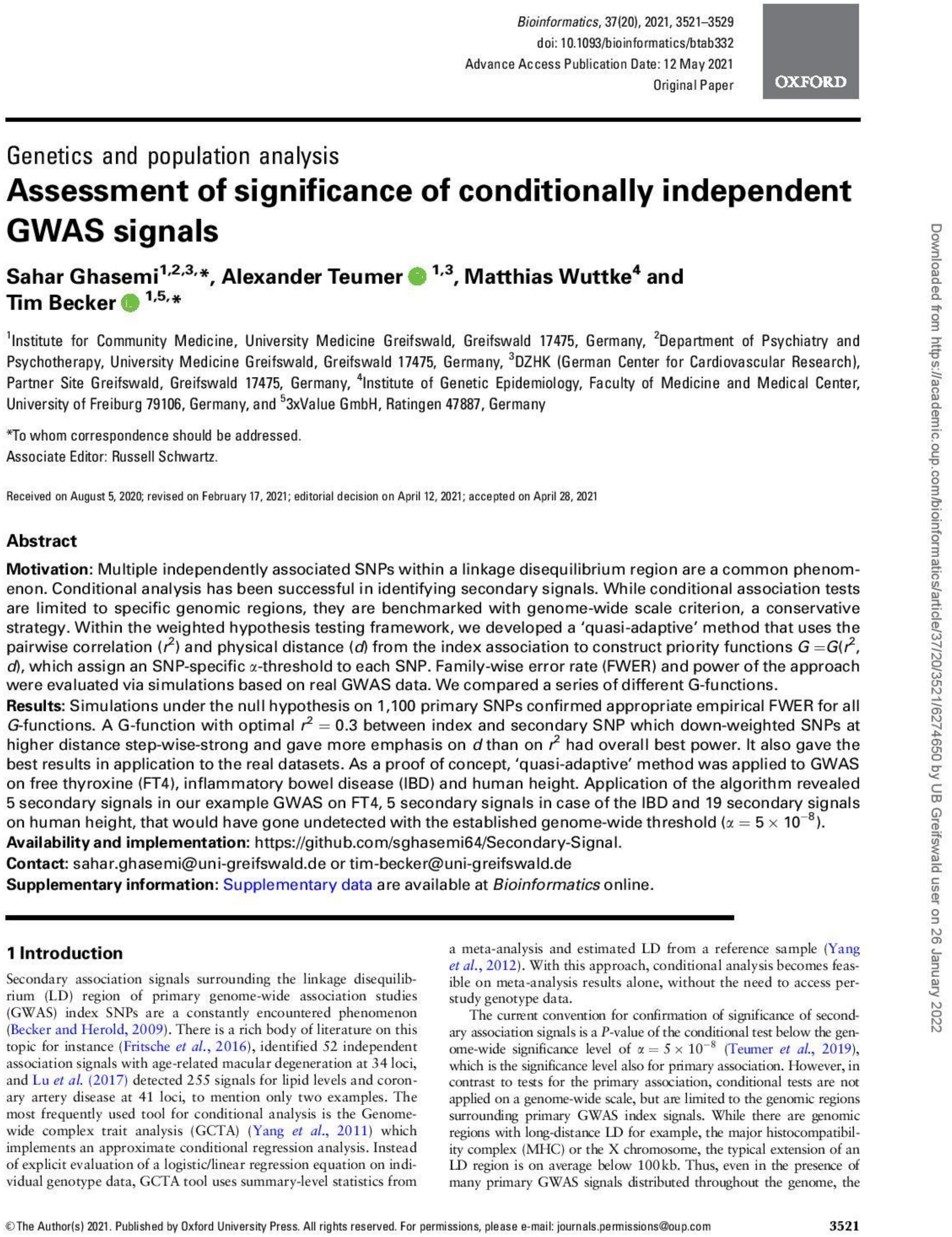
Alexander Teumer[1,2,160], Yong Li[3,160], Sahar Ghasemi[1,2,160], Bram P. Prins[4,160], Matthias Wuttke[3,5,160],
Tobias Hermle[5,160], Ayush Giri[6,7], Karsten B. Sieber[8], Chengxiang Qiu[9], Holger Kirsten[10,11],
Adrienne Tin[12,13], Audrey Y. Chu[14], Nisha Bansal[15,16], Mary F. Feitosa[17], Lihua Wang[17], Jin-Fang Chai[18],
Massimiliano Cocca[19], Christian Fuchsberger[20], Mathias Gorski[21,22], Anselm Hoppmann[3],
Katrin Horn[10,11], Man Li[23], Jonathan Marten[24], Damia Noce[20], Teresa Nutile[25], Sanaz Sedaghat[26],
Gardar Sveinbjornsson[27], Bamidele O. Tayo[28], Peter J. van der Most[29], Yizhe Xu[23], Zhi Yu[12,30], Lea Gerstner[5],
Johan Ärnlöv[31,32], Stephan J.L. Bakker[33], Daniela Baptista[34], Mary L. Biggs[35,36], Eric Boerwinkle[37],
Hermann Brenner[38,39], Ralph Burkhardt[11,40,41], Robert J. Carroll[42], Miao-Li Chee[43], Miao-Ling Chee[43],
Mengmeng Chen[5], Ching-Yu Cheng[43,44,45], James P. Cook[46], Josef Coresh[12], Tanguy Corre[47,48,49],
John Danesh[50], Martin H. de Borst[33], Alessandro De Grandi[20], Renée de Mutsert[51], Aiko P.J. de Vries[52],
Frauke Degenhardt[53], Katalin Dittrich[54,55], Jasmin Divers[56], Kai-Uwe Eckardt[57,58], Georg Ehret[34],
Karlhans Endlich[2,59], Janine F. Felix[26,60,61], Oscar H. Franco[26,62], Andre Franke[53], Barry I. Freedman[63],
Sandra Freitag-Wolf[64], Ron T. Gansevoort[33], Vilmantas Giedraitis[65], Martin Gögele[20],
Franziska Grundner-Culemann[3], Daniel F. Gudbjartsson[27], Vilmundur Gudnason[66,67], Pavel Hamet[68,69],
Tamara B. Harris[70], Andrew A. Hicks[20], Hilma Holm[27], Valencia Hui Xian Foo[43], Shih-Jen Hwang[71,72],
M. Arfan Ikram[26], Erik Ingelsson[73,74,75,76], Vincent W.V. Jaddoe[26,60,61], Johanna Jakobsdottir[77,78],
Navya Shilpa Josyula[79], Bettina Jung[21], Mika Kähönen[80,81], Chiea-Chuen Khor[43,82], Wieland Kiess[11,54,55],
Wolfgang Koenig[83,84,85], Antje Körner[11,54,55], Peter Kovacs[86], Holly Kramer[28,87], Bernhard K. Krämer[88],
Florian Kronenberg[89], Leslie A. Lange[90], Carl D. Langefeld[56], Jeannette Jen-Mai Lee[18], Terho Lehtimäki[91,92],
Wolfgang Lieb[93], Su-Chi Lim[18,94], Lars Lind[95], Cecilia M. Lindgren[96,97], Jianjun Liu[82,98],
Markus Loeffler[10,11], Leo-Pekka Lyytikäinen[91,92], Anubha Mahajan[99,100], Joseph C. Maranville[101,158],
Deborah Mascalzoni[20], Barbara McMullen[102], Christa Meisinger[103,104], Thomas Meitinger[84,105,106],
Kozeta Miliku[26,60,61], Dennis O. Mook-Kanamori[51,107], Martina Müller-Nurasyid[108,109,110],
Josyf C. Mychaleckyj[111], Matthias Nauck[2,112], Kjell Nikus[113,114], Boting Ning[115], Raymond Noordam[116],
Jeffrey O' Connell[117], Isleifur Olafsson[118], Nicholette D. Palmer[119], Annette Peters[84,120,121],
Anna I. Podgornaia[14], Belen Ponte[122], Tanja Poulain[11], Peter P. Pramstaller[20], Ton J. Rabelink[52,123],
Laura M. Raffield[124], Dermot F. Reilly[14], Rainer Rettig[125], Myriam Rheinberger[21], Kenneth M. Rice[36],
Fernando Rivadeneira[26,126], Heiko Runz[101,159], Kathleen A. Ryan[127], Charumathi Sabanayagam[43,44],
Kai-Uwe Saum[38], Ben Schöttker[38,39], Christian M. Shaffer[42], Yuan Shi[43,44], Albert V. Smith[67],
Konstantin Strauch[108,109], Michael Stumvoll[128], Benjamin B. Sun[4], Silke Szymczak[64], E-Shyong Tai[18,98,129],
Nicholas Y.Q. Tan[43], Kent D. Taylor[130], Andrej Teren[11,131], Yih-Chung Tham[43], Joachim Thiery[11,40],
Chris H.L. Thio[29], Hauke Thomsen[132], Unnur Thorsteinsdottir[27], Anke Tönjes[128], Johanne Tremblay[68,133],
André G. Uitterlinden[126], Pim van der Harst[134,135,136], Niek Verweij[134], Suzanne Vogelezang[26,60,61],
Uwe Völker[2,137], Melanie Waldenberger[84,120,138], Chaolong Wang[82,139], Otis D. Wilson[140],
Charlene Wong[45], Tien-Yin Wong[43,44,45], Qiong Yang[115], Masayuki Yasuda[43,141], Shreeram Akilesh[16,142],
Murielle Bochud[47], Carsten A. Böger[21,143], Olivier Devuyst[144], Todd L. Edwards[145,146], Kevin Ho[147,148],
Andrew P. Morris[46,99], Afshin Parsa[149,150], Sarah A. Pendergrass[151], Bruce M. Psaty[152,153],
Jerome I. Rotter[130,154,155], Kari Stefansson[27], James G. Wilson[156], Katalin Susztak[9], Harold Snieder[29],

Iris M. Heid[22], Markus Scholz[10,11], Adam S. Butterworth[4,157,161], Adriana M. Hung[140,146,161], Cristian Pattaro[20,161] & Anna Köttgen[3,12,161]

[1]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. [2]DZHK (German Center for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany. [3]Institute of Genetic Epidemiology, Department of Biometry, Epidemiology and Medical Bioinformatics, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany. [4]MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [5]Renal Division, Department of Medicine, Faculty of Medicine and Medical Center - University of Freiburg, Freiburg, Germany. [6]Division of Quantitative Sciences, Department of Obstetrics & Gynecology, Vanderbilt Genetics Institute, Vanderbilt Epidemiology Center, Institute for Medicine and Public Health, Vanderbilt University Medical Center, Nashville, TN, USA. [7]Biomedical Laboratory Research and Development, Tennessee Valley Healthcare System (626)/Vanderbilt University, Nashville, TN, USA. [8]Target Sciences - Genetics, GlaxoSmithKline, Collegeville, PA, USA. [9]Renal Electrolyte and Hypertension Division, Department of Medicine, Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Pennsylvania, PA, USA. [10]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [11]LIFE Research Centre for Civilization Diseases, University of Leipzig, Leipzig, Germany. [12]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [13]Epidemiology and Clinical Research, Welch Centre for Prevention, Baltimore, MD, USA. [14]Genetics, Merck & Co., Inc., Kenilworth, NJ, USA. [15]Division of Nephrology, University of Washington, Seattle, WA, USA. [16]Kidney Research Institute, University of Washington, Seattle, WA, USA. [17]Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA. [18]Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore, Singapore. [19]Institute for Maternal and Child Health - IRCCS "Burlo Garofolo", Trieste, Italy. [20]Eurac Research, Institute for Biomedicine (affiliated to the University of Lübeck), Bolzano, Italy. [21]Department of Nephrology, University Hospital Regensburg, Regensburg, Germany. [22]Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany. [23]Department of Medicine, Division of Nephrology and Hypertension, University of Utah, Salt Lake City, UT, USA. [24]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [25]Institute of Genetics and Biophysics "Adriano Buzzati-Traverso" - CNR, Naples, Italy. [26]Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [27]deCODE Genetics, Amgen Inc., Reykjavik, Iceland. [28]Department of Public Health Sciences, Loyola University Chicago, Maywood, IL, USA. [29]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [30]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [31]Department of Neurobiology, Care Sciences and Society, Division of Family Medicine and Primary Care, Karolinska Institutet, Stockholm, Sweden. [32]School of Health and Social Studies, Dalarna University, Falun, Sweden. [33]Department of Internal Medicine, Division of Nephrology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [34]Cardiology, Geneva University Hospitals, Geneva, Switzerland. [35]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [36]Department of Biostatistics, University of Washington, Seattle, WA, USA. [37]Human Genetics Centre, University of Texas Health Science Centre, Houston, TX, USA. [38]Division of Clinical Epidemiology and Aging Research, German Cancer Research Centre (DKFZ), Heidelberg, Germany. [39]Network Aging Research, University of Heidelberg, Heidelberg, Germany. [40]Institute of Laboratory Medicine, Clinical Chemistry and Molecular Diagnostics, University of Leipzig, Leipzig, Germany. [41]Institute of Clinical Chemistry and Laboratory Medicine, University Hospital Regensburg, Regensburg, Germany. [42]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. [43]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [44]Ophthalmology & Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore, Singapore. [45]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [46]Department of Biostatistics, University of Liverpool, Liverpool, UK. [47]Center for Primary Care and Public Health (Unisanté), University of Lausanne, Lausanne, Switzerland. [48]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [49]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [50]Department of Public Health and Primary Care, School of Clinical Medicine, University of Cambridge, Cambridge, UK. [51]Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands. [52]Section of Nephrology, Department of Internal Medicine, Leiden University Medical Centre, Leiden, The Netherlands. [53]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany. [54]Department of Women and Child Health, Hospital for Children and Adolescents, University of Leipzig, Leipzig, Germany. [55]Centre for Pediatric Research, University of Leipzig, Leipzig, Germany. [56]Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, NC, USA. [57]Intensive Care Medicine, Charité, Berlin, Germany. [58]Department of Nephrology and Hypertension, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Erlangen, Germany. [59]Department of Anatomy and Cell Biology, University Medicine Greifswald, Greifswald, Germany. [60]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [61]Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [62]Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland. [63]Internal Medicine - Section on Nephrology, Wake Forest School of Medicine, Winston-Salem, NC, USA. [64]Institute of Medical Informatics and Statistics, Kiel University, University Hospital Schleswig-Holstein, Kiel, USA. [65]Department of Public Health and Caring Sciences, Molecular Geriatrics, Uppsala University, Uppsala, Sweden. [66]Icelandic Heart Association, Kopavogur, Iceland. [67]Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. [68]Montreal University Hospital Research Centre, CHUM, Montreal, QC, Canada. [69]Medpharmgene, Montreal, QC, Canada. [70]Laboratory of Epidemiology and Population Sciences, National Institute on Aging, Intramural Research Program, National Institutes of Health, Bethesda, MD, USA. [71]NHLBI's Framingham Heart Study, Framingham, MA, USA. [72]The Centre for Population Studies, NHLBI, Framingham, MA, USA. [73]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, USA. [74]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA. [75]Molecular Epidemiology and Science for Life Laboratory, Department of Medical Sciences, Uppsala University, Uppsala, Sweden. [76]Stanford Diabetes Research Center, Stanford University, Stanford, CA, USA. [77]Icelandic Heart Association, Holtasmari 1, Kopavogur IS-201, Iceland. [78]The Centre of Public Health Sciences, University of Iceland, Sturlugata 8, Reykjavík IS-101, Iceland. [79]Geisinger Research, Biomedical and Translational Informatics Institute, Rockville, MD, USA. [80]Department of Clinical Physiology, Tampere University Hospital, Tampere, Finland. [81]Department of Clinical Physiology, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [82]Genome Institute of Singapore, Agency for Science Technology and Research, Singapore, Singapore. [83]Deutsches Herzzentrum München, Technische Universität München, Munich, Germany. [84]DZHK (German Centre for Cardiovascular Research), Partner Site Munich Heart Alliance, Munich, Germany. [85]Institute of Epidemiology and Biostatistics, University of Ulm, Ulm, Germany. [86]Integrated Research and Treatment Centre Adiposity Diseases, University of Leipzig, Leipzig, Germany. [87]Division of Nephrology and Hypertension, Loyola University Chicago, Chicago, IL, USA. [88]5th Department of Medicine (Nephrology, Hypertensiology, Rheumatology, Endocrinology, Diabetology), Medical Faculty Mannheim, University of Heidelberg, Mannheim, Germany. [89]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria.

[90]Division of Biomedical Informatics and Personalized Medicine, School of Medicine, University of Colorado Denver - Anschutz Medical Campus, Aurora, CO, USA. [91]Department of Clinical Chemistry, Fimlab Laboratories, Tampere, Finland. [92]Department of Clinical Chemistry, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [93]Institute of Epidemiology and Biobank Popgen, Kiel University, Kiel, Germany. [94]Diabetes Centre, Khoo Teck Puat Hospital, Singapore, Singapore. [95]Cardiovascular Epidemiology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden. [96]Nuffield Department of Medicine, University of Oxford, Oxford, UK. [97]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [98]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore, Singapore. [99]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [100]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK. [101]MRL, Merck & Co., Inc., Kenilworth, NJ, USA. [102]Vanderbilt University School of Medicine, Nashville, TN, USA. [103]Independent Research Group Clinical Epidemiology, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany. [104]Chair of Epidemiology Ludwig- Maximilians-Universität München at UNIKA-T Augsburg, Augsburg, Germany. [105]Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany. [106]Institute of Human Genetics, Technische Universität München, Munich, Germany. [107]Department of Public Health and Primary Care, Leiden University Medical Centre, Leiden, The Netherlands. [108]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Centre for Environmental Health, Neuherberg, Germany. [109]Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU, Munich, Germany. [110]Department of Internal Medicine I (Cardiology), Hospital of the Ludwig-Maximilians-University (LMU) Munich, Munich, Germany. [111]Centre for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [112]Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany. [113]Department of Cardiology, Heart Center, Tampere University Hospital, Tampere, Finland. [114]Department of Cardiology, Finnish Cardiovascular Research Center - Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. [115]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [116]Section of Gerontology and Geriatrics, Department of Internal Medicine, Leiden University Medical Centre, Leiden, The Netherlands. [117]University of Maryland School of Medicine, Baltimore, MD, USA. [118]Department of Clinical Biochemistry, Landspitali University Hospital, Reykjavik, Iceland. [119]Biochemistry, Wake Forest School of Medicine, Winston-Salem, NC, USA. [120]Institute of Epidemiology, Helmholtz Zentrum München - German Research Centre for Environmental Health, Neuherberg, Germany. [121]German Center for Diabetes Research (DZD), Neuherberg, Germany. [122]Service de Néphrologie, Geneva University Hospitals, Geneva, Switzerland. [123]Einthoven Laboratory of Experimental Vascular Research, Leiden University Medical Centre, Leiden, The Netherlands. [124]Department of Genetics, University of North Carolina, Chapel Hill, NC, USA. [125]Institute of Physiology, University Medicine Greifswald, Karlsburg, Germany. [126]Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [127]Division of Endocrinology, Diabetes and Nutrition, University of Maryland School of Medicine, Baltimore, MD, USA. [128]Department of Endocrinology and Nephrology, University of Leipzig, Leipzig, Germany. [129]Duke-NUS Medical School, Singapore, Singapore. [130]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, USA. [131]Heart Centre Leipzig, Leipzig, Germany. [132]Division of Molecular Genetic Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany. [133]CRCHUM, Montreal, QC, Canada. [134]Department of Cardiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [135]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [136]Durrer Centre for Cardiovascular Research, The Netherlands Heart Institute, Utrecht, The Netherlands. [137]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [138]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Centre for Environmental Health, Neuherberg, Germany. [139]School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China. [140]Vanderbilt University Medical Centre, Division of Nephrology & Hypertension, Nashville, TN, USA. [141]Department of Ophthalmology, Tohoku University Graduate School of Medicine, Sendai, Japan. [142]Anatomic Pathology, University of Washington Medical Center, Seattle, WA, USA. [143]Department of Nephrology, Diabetology and Rheumatology, Kliniken Südostbayern, Traunstein, Germany. [144]Institute of Physiology, University of Zurich, Zurich, Switzerland. [145]Division of Epidemiology, Department of Medicine, Vanderbilt Genetics Institute, Vanderbilt University Medical Centre, Nashville, TN, USA. [146]Department of Veteran's Affairs, Tennessee Valley Healthcare System (626)/Vanderbilt University, Nashville, TN, USA. [147]Kidney Health Research Institute (KHRI), Geisinger, Danville, PA, USA. [148]Department of Nephrology, Geisinger, Danville, PA, USA. [149]Division of Kidney, Urologic and Hematologic Diseases, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA. [150]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [151]Geisinger Research, Biomedical and Translational Informatics Institute, Danville, PA, USA. [152]Cardiovascular Health Research Unit, Department of Medicine, Department of Epidemiology, Department of Health Service, University of Washington, Seattle, WA, USA. [153]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA. [154]Department of Pediatrics, Harbor-UCLA Medical Centre, Torrance, CA, USA. [155]Department of Medicine, Harbor-UCLA Medical Centre, Torrance, CA, USA. [156]Department of Physiology and Biophysics, University of Mississippi Medical Centre, Jackson, MS, USA. [157]National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK. [158]Present address: Celgene Inc., Cambridge, MA, USA. [159]Present address: Biogen Inc., Cambridge, MA, USA. [160]These authors contributed equally: Alexander Teumer, Yong Li, Sahar Ghasemi, Bram P. Prins, Matthias Wuttke, Tobias Hermle. [161]These authors jointly supervised: Adam S. Butterworth, Adriana M. Hung, Cristian Pattaro, Anna Köttgen.

Assessment of significance of conditionally independent GWAS signals (Article II)



Bioinformatics, 37(20), 2021, 3521–3529
doi: 10.1093/bioinformatics/btab332
Advance Access Publication Date: 12 May 2021
Original Paper

OXFORD

Genetics and population analysis

Assessment of significance of conditionally independent GWAS signals

Sahar Ghasemi[1,2,3,*], **Alexander Teumer**[1,3], **Matthias Wuttke**[4] and **Tim Becker**[1,5,*]

[1]Institute for Community Medicine, University Medicine Greifswald, Greifswald 17475, Germany, [2]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald 17475, Germany, [3]DZHK (German Center for Cardiovascular Research), Partner Site Greifswald, Greifswald 17475, Germany, [4]Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center, University of Freiburg 79106, Germany, and [5]3xValue GmbH, Ratingen 47887, Germany

*To whom correspondence should be addressed.
Associate Editor: Russell Schwartz.

Received on August 5, 2020; revised on February 17, 2021; editorial decision on April 12, 2021; accepted on April 28, 2021

Abstract

Motivation: Multiple independently associated SNPs within a linkage disequilibrium region are a common phenomenon. Conditional analysis has been successful in identifying secondary signals. While conditional association tests are limited to specific genomic regions, they are benchmarked with genome-wide scale criterion, a conservative strategy. Within the weighted hypothesis testing framework, we developed a 'quasi-adaptive' method that uses the pairwise correlation (r^2) and physical distance (d) from the index association to construct priority functions $G = G(r^2, d)$, which assign an SNP-specific α-threshold to each SNP. Family-wise error rate (FWER) and power of the approach were evaluated via simulations based on real GWAS data. We compared a series of different G-functions.

Results: Simulations under the null hypothesis on 1,100 primary SNPs confirmed appropriate empirical FWER for all G-functions. A G-function with optimal $r^2 = 0.3$ between index and secondary SNP which down-weighted SNPs at higher distance step-wise-strong and gave more emphasis on d than on r^2 had overall best power. It also gave the best results in application to the real datasets. As a proof of concept, 'quasi-adaptive' method was applied to GWAS on free thyroxine (FT4), inflammatory bowel disease (IBD) and human height. Application of the algorithm revealed 5 secondary signals in our example GWAS on FT4, 5 secondary signals in case of the IBD and 19 secondary signals on human height, that would have gone undetected with the established genome-wide threshold ($\alpha = 5 \times 10^{-8}$).

Availability and implementation: https://github.com/sghasemi64/Secondary-Signal.

Contact: sahar.ghasemi@uni-greifswald.de or tim-becker@uni-greifswald.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Secondary association signals surrounding the linkage disequilibrium (LD) region of primary genome-wide association studies (GWAS) index SNPs are a constantly encountered phenomenon (Becker and Herold, 2009). There is a rich body of literature on this topic for instance (Fritsche *et al.*, 2016), identified 52 independent association signals with age-related macular degeneration at 34 loci, and Lu *et al.* (2017) detected 255 signals for lipid levels and coronary artery disease at 41 loci, to mention only two examples. The most frequently used tool for conditional analysis is the Genome-wide complex trait analysis (GCTA) (Yang *et al.*, 2011) which implements an approximate conditional regression analysis. Instead of explicit evaluation of a logistic/linear regression equation on individual genotype data, GCTA tool uses summary-level statistics from a meta-analysis and estimated LD from a reference sample (Yang *et al.*, 2012). With this approach, conditional analysis becomes feasible on meta-analysis results alone, without the need to access per-study genotype data.

The current convention for confirmation of significance of secondary association signals is a P-value of the conditional test below the genome-wide significance level of $\alpha = 5 \times 10^{-8}$ (Teumer *et al.*, 2019), which is the significance level also for primary association. However, in contrast to tests for the primary association, conditional tests are not applied on a genome-wide scale, but are limited to the genomic regions surrounding primary GWAS index signals. While there are genomic regions with long-distance LD for example, the major histocompatibility complex (MHC) or the X chromosome, the typical extension of an LD region is on average below 100 kb. Thus, even in the presence of many primary GWAS signals distributed throughout the genome, the

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

surrounding LD regions will cover only a small portion of the entire genome and contain only a reduced number of SNPs. It is obvious to hypothesize that, application of the established genome-wide significance level in conditional analyses is too strict and implies an unnecessary loss of power. Therefore, we developed a 'quasi-adaptive' method that exploits the type *I* error by providing SNP-specific α-thresholds to prioritize candidate SNPs for analysis. In addition, we investigate if the LD structure, pairwise correlation (r^2) and physical distance (d) between the index SNP and each SNP from LD surrounding region, can be used to improve signal detection. We explored different strategies of prioritization and α-spending (Demets and Lan, 1994) and applied the proposed method on different GWAS meta-analysis results.

2 Materials and methods

We prioritize SNPs for conditional analysis and derived SNP-specific α-thresholds by spending type *I* error over the SNPs from surrounding LD regions of index SNPs with a priority function. In essence, SNPs with higher priority have to meet less stringent α-levels. Physical distance (d) and LD distance or pairwise correlation (r^2) are used to assign priorities. Of note, d and r^2 operate to some extent in opposite direction: short distance and high r^2 are correlated and short distance increases the biological priority since SNPs are from the same functional unit. However, while high r^2 might also increases the biological priority (via haplotype effects), it reduces statistical power since the amount of independent information is reduced. In order to explore these counter running effects, we assign priorities via *G*-functions. Those *G*-functions are functions of d and r^2 and evaluate the impact of giving different relative weight on the d-component or the r^2-component. The entire algorithm can be described as follows:

1. Identify primary GWAS index SNPs. Let I_n be an index SNP located within a 1-Mb window size surrounding LD region (S_n), $n = 1, 2, \ldots, N$.
2. Prune the LD reference panel prior to analysis to remove SNPs with strong pair-wise LD ($r^2 \geq 0.99$). This step is optional to reduce the number of computations without loss of information. The following steps 3 to 8 are done for each I_n separately.
3. Retrieve r^2 and d between I_n and each SNP from S_n by using INTERSNP tool (Herold *et al.*, 2009).
4. Assign pre-weight w_r based on r^2 to each SNP from S_n. $r^2 = 0$ and $r^2 = 0.3$ were considered as the optimal values to built w_r as follows:

$$w_r = \begin{cases} \dfrac{1 - |r^2 - 0.3| - 0.3}{1 - 0.3} & \text{, if optimal} \quad r^2 = 0.3 \\ 1 - r^2 & \text{, if optimal} \quad r^2 = 0 \end{cases}$$

Note that, a secondary signal whose correlation is not 'too high' and not 'too low' with the index SNP is likely to provide much extra information for association test. Howey and Cordell (2014) showed that, r^2 of 0.3 is an optimal correlation to detect secondary signals.

5 Assign pre-weight w_d based on d (step-wise-strong or step-wise-moderate) to each SNP from S_n as follows:

Distance (Kb)	(0,1]	(1,10]	(10,50]	(50,100]	(100,500]
w_d (step-wise-strong)	1	0.5	0.25	0.125	0.0625
w_d (step-wise-moderate)	1	0.8	0.6	0.4	0.2

6 Combine w_d and w_r and assign final pre-weight w to each SNP from S_n. Three different relative weightings of d and r^2 with $k = 5$ were considered to construct w as follows:

$$\mathrm{w} = \begin{cases} (\mathrm{w_d} \times \mathrm{w_r})^{\frac{1}{2}} & \text{, for equal weight on d and } \mathrm{r}^2 \\ (\mathrm{w_d^k} \times \mathrm{w_r})^{\frac{1}{k+1}} & \text{, for more weight on d than } \mathrm{r}^2 \\ (\mathrm{w_d} \times \mathrm{w_r^k})^{\frac{1}{k+1}} & \text{, for more weight on } \mathrm{r}^2 \text{ than d} \end{cases}$$

7 Run conditional GCTA analysis (–cojo-cond), condition SNPs from S_n on I_n.

8 Joint the results from step 6 and step 7 and get the tabulated result with w and conditional *P*-value for each SNPs from S_n.

9 Merge the results from step 8 for N index SNPs and get the combine result for N loci. Let S be a set of all candidate SNPs, i.e. across N loci (LD regions), and m be the number of candidate SNPs within S.

10 Achieve the SNP-specific α-threshold by spending $\alpha = 0.05$over m candidate SNPs from S according to the priority *G*-function.

$$G_i = 1 - (1 - \alpha)^{W_i/m} \text{where,} \quad W_i = \frac{w_i \times m}{\sum_{i=1}^{m} w_i}, \quad i = 1, 2, \ldots, m.$$

Note that, the *G*-function is a realization of the weighted Sidak procedure for multiple hypothesis testing as described in Kang *et al.* (2009).

The combinations of pre-weights described in steps 4, 5, and 6 lead to 12 different *G*-functions which are summarized in Table 1. In addition, we considered the case of equal prior weight for all SNPs as further G-function (*G*13).

11 Fix one *G*-function and apply it to every SNP in S to retrieve the SNP-specific α-threshold.

12 Compared the GCTA *P*-value with the SNP-specific α-threshold to assess the significance of each candidate SNP. The ith SNP from S, $i = 1, 2, \ldots, m$ is genome wide conditionally independent significant if the conditional *P*-value (p_i) is smaller than G_i.

3 Simulation

Imputed genotypes [1000 Genomes data (Clarke *et al.*, 2012)] from the SHIP study (Völzke *et al.*, 2011) served as a basis of our simulation study. The dataset comprises 4070 individuals, for which we simulated a binary disease phenotype. For the simulation under the null hypothesis of no secondary signals, we randomly picked a set of 22 independent SNPs, one from each autosome, under the restriction that their minor allele frequency (MAF) > 0.10. We conducted different simulation series, in which we worked with different SNP sets, to overcome phenomena specific to the particular choice of the SNP set.

For each SNP, one of the alleles was assigned randomly as the risk allele. Each risk allele increases the probability of getting the (simulated) disease. We assumed that individuals with zero risk allele have a risk of 0.01 to be affected due to the environmental factors. For each individual, the disease risk was set to be $f = 0.01 \times \prod_{i=1}^{22} OR_i^{n_i}$ where, n_i is the number of risk allele each pre-defined SNP carries and OR_i is the odd ratio of corresponding SNP. $ln(OR_i)$ was randomly drawn from normal distribution $N(ln(1.15), sd = 0.03)$. The choice of the value 1.15 is motivated by effect sizes typically observed in GWAS studies. By construction, probands with 0 risk alleles have a probability of being a case of 0.01, and with every risk allele, the probability of being a case increases by OR_i.

We simulated a GWAS meta-analysis of three virtual studies. All three studies have as a basis the SHIP genotype data, but with different (simulated) phenotype data. The phenotypes were assigned according to the number of risk alleles at the pre-defined SNPs. For

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

Table 1. Summary of 13*G*-functions

G-function	Distance (*d*)	Optimal r^2	Relative weighting of *d* and r^2
G1	Step-wise-strong	0.0	Equal weight
G2	Step-wise-strong	0.0	More weight on *d*
G3	Step-wise-strong	0.0	More weight on r^2
G4	Step-wise-strong	0.3	Equal weight
G5	Step-wise-strong	0.3	More weight on *d*
G6	Step-wise-strong	0.3	More weight on r^2
G7	Step-wise-moderate	0.0	Equal weight
G8	Step-wise-moderate	0.0	More weight on *d*
G9	Step-wise-moderate	0.0	More weight on r^2
G10	Step-wise-moderate	0.3	Equal weight
G11	Step-wise-moderate	0.3	More weight on *d*
G12	Step-wise-moderate	0.3	More weight on r^2
G13	Weight $W_i = 1$ for all SNPs (Sidak-correction by the number of SNPs)		

each individual, a random number *x* from the uniform distribution *U*(0, 1) was drawn. In case $x \leq f$, the individual was set to be affected (status=2), otherwise, we set the affection status of the individual to 1. By construction, we obtain three GWAS, all of which have the same 'true' 22 primary SNP associations. It should be noted that, due to the inherent randomness of the simulation study approach, not each primary SNP necessary is genome-wide significant in each replicate. However, in the majority of cases, the primary SNPs are genome-wide significant within the replicated dataset and could be evaluated for type *I* error assessment.

Since in our set-up we simulate only primary associations, all other significant signals are either due to LD with an index SNP (and hence are conditional signals and not true unconditional signals) or are false positives. The latter case should not occur more than 5 percent of the time in our simulation datasets to control the family-wise error rate. The simulation scheme can be summarized for 50 replicates as follows:

1. Randomly pick a set of 22 primary SNPs.
2. Determine the disease risk for each individual and simulate three virtual GWAS phenotypes.
3. Perform association analysis for three GWAS studies. Since we had no population structure simulated, we did simple logistic regression with PLINK (Purcell *et al.*, 2007).
4. Perform meta-analysis on three GWAS using METAL (Willer *et al.*, 2010).
5. Identify all SNPs between pre-defined 22 SNPs that reach genome-wide significance threshold at $\alpha = 5 \times 10^{-8}$. These are the index SNPs for the current replicate. Let M_j be the number of significant SNPs between 22 pre-defined SNPs for replicate $j = 1, 2, \ldots, 50$.
6. Determine the SNP-specific α-threshold according to the algorithm in method section for *G*-functions (Table 1). SHIP imputed individual-level genotype dataset was served as an LD reference panel to estimate LD structure in steps 3 and 7 in the method section. SHIP reference panel has been checked for cryptic relatedness and population stratification.
7. Identify index SNPs with at least one secondary signal. Let N_j be the number of index SNPs with secondary signal for $j = 1, 2, \ldots, 50$.
8. Calculate $\text{EFWER} = \sum_{j=1}^{50} N_j / \sum_{j=1}^{50} M_j$ for 50 replicates.

For power analysis under the alternative hypothesis (pre-defined secondary signals), an analogous scheme was used, with the following modifications: we randomly picked 13 SNPs as primary index SNPs p_i and, for each of these, an additional SNP q_i from its LD region as secondary signal. We required p_i and q_i with MAF > 0.10 form 1-Mb LD region. Furthermore, we selected q_i with $0.2 < r^2 < 0.8$ with corresponding p_i. For p_i and q_i we randomly assigned their risk allele. The risk allele effect for p_i was chosen from normal distribution $N(ln(1.15), 0.05)$ and for q_i from $N(ln(1.10), 0.05)$. We counted the total number of risk alleles per individual, summing both over primary and secondary alleles, and randomly assigned disease status for three virtual GWAS phenotype as before. To conduct power analysis, step 3 to step 7 from simulation replicated 25 times, and the portion of index SNPs with secondary signal calculated as the power for each *G*-function.

We considered three different alternative scenarios which we implemented by imposing additional conditions on the random choice of the secondary SNPs: in scenario A, we selected q_i at random from the SNPs fulfilling the above condition to be a secondary SNP for corresponding p_i. In scenario B and C from the eligible SNPs, q_i was picked conditionally at random based on distance to p_i. SNPs with lower *d* got higher likelihood to be selected but with different impact of *d* between scenario B and C. Likelihood decreased step-wise moderate in scenario B and step-wise strong in scenario C from low *d* to high *d*.

4 Results

Different relative pre-weights depending on r^2 and pre-weights depending on *d* were defined to construct priority *G*-functions. Simulations under the null hypothesis of no secondary signal over 50 runs (1,100 primary SNPs) were set up to confirm the validity of the *G*-functions. In our simulation study, $\sum_{j=1}^{50} M_j = 982$ SNPs out of 1,100 pre-defined primary SNPs are significant index SNPs which allow evaluation of secondary effects. M_j has median = 19.5 [min = 17; max = 22] over 50 runs. Binomial exact test was used to assess the deviation of EFWAR from desired level 0.05. The results of simulation as well as the *P*-values of binomial exact test for *G*-functions are listed in Table 2. Deviations from the empirical level are overall small in size and not significant. *P*-values confirm that EFWERs are well controlled by *G*1 to *G*12. In addition, to compare our method against simple Sidak-correction (equal hypothesis weighting), EFWER was determined for the *G*13 function. *G*13 is also valid and controls type *I* error in our simulation study. Power simulations were set up under the alternative hypothesis over 25 runs (325 primary SNPs and 325 secondary signals) to evaluate the power of *G*-functions in detecting secondary signals. The results of power simulations for three scenarios are reported in Table 2 and Figure 1. *G*2 and *G*5 functions show the best median power over three scenarios. Both functions down-weighted SNPs at higher distance step-wise-strong and gave more weight on *d* than r^2, but with different optimal value of r^2 (Table 1). Note that, *G*2 and *G*5 show equal power in scenarios B and C where the secondary signals with lower *d* to the index SNP had more chance to be selected by the algorithm. In these cases, pre-weights on *d* are the same for both functions and the differences between pre-weights r^2 compensate since the algorithm gave more emphasis on *d* than r^2. On the other hand,

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

Table 2. Results of simulation study under the null hypothesis for 50 runs and the results of simulation power analysis for 25 runs

	G-function	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14
Simulation under the null hypothesis	EFWER	0.0519	0.0540	0.0468	0.0560	0.0570	0.0530	0.0509	0.0540	0.0468	0.0570	0.0550	0.0519	0.0499	—
	P-value	0.7694	0.5577	0.7143	0.3793	0.3051	0.6601	0.8835	0.5577	0.7143	0.3051	0.4637	0.7694	1.0000	—
	CI-low	0.3890	0.0407	0.0345	0.0425	0.0434	0.0398	0.0380	0.0407	0.0345	0.0435	0.0416	0.0389	0.0371	—
	CI-up	0.0677	0.0700	0.0620	0.0723	0.0734	0.0689	0.0666	0.0700	0.0620	0.0734	0.0711	0.0677	0.0654	—
Simulation power analysis	Power A	0.6937	0.7159	0.6716	0.7048	0.7232	0.7011	0.6900	0.7085	0.6642	0.7048	0.7085	0.6974	0.6753	0.4576
	Power B	0.7163	0.7305	0.6986	0.7199	0.7305	0.7163	0.7092	0.7199	0.6986	0.7199	0.7234	0.7128	0.7021	0.5071
	Power C	0.7112	0.7329	0.6968	0.7148	0.7329	0.7112	0.7112	0.7148	0.6895	0.7148	0.7184	0.7112	0.6931	0.5415
	Median	0.7112	0.7305	0.6968	0.7148	0.7305	0.7112	0.7092	0.7148	0.6895	0.7148	0.7184	0.7112	0.6931	0.5071

Note: EFWER, empirical family wise error rate; *P*-value, binomial exact test *P*-value; CI-low, binomial exact test-lower 95% confidence interval; CI-up, binomial exact test-upper 95% confidence interval; Power A, simulation power in scenario A; Power B, simulation power in scenario B; Power C, simulation power in scenario C; Median, median over three scenarios in power simulation; *G*13, simple Sidak-correction (equal hypothesis weighting); *G*14, established genome-wide threshold ($\alpha = 5 \times 10^{-8}$). The results of simulation under the null hypothesis were not shown for *G*14, since the EFWER depends on the number of regions and SNPs assumed in the simulation.

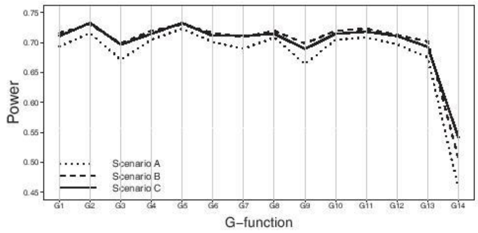


Fig. 1. Power simulation analysis results

*G*2 shows less power than *G5* in scenario A where the selection of secondary signals was random and not dependent on *d*. In this case, power simulation results confirm that the optimal value of $r^2 = 0.3$ provides extra power for the *G5* function to detect secondary signals with higher distance to the index SNP. According to the power analysis results, the weighting *G5* function with the overall best performance is selected as the final priority function. *G5* function outperforms the power of the Sidak-correction by 4 percentage points (median). In addition, power analysis for 3 scenarios was assessed for the established genome-wide threshold (5×10^{-8}) by *G*14 (Table 2 and Fig. 1). Our *G5* function shows improved power by 22 percentage points (median) compare to the *G*14. Note that, the results of simulation under the null hypothesis were not shown for *G*14 since the EFWER depends on the number of regions and SNPs assumed in the simulation. As a proof of concept, we applied our method on summary-level statistics (allele frequency, effect size, standard error and *P*-value) from three GWAS meta-analysis summary data. GWAMA for thyroid function including 7,971,759 genetic variants and up to 49,269 subjects of European ancestry revealed 21 known and novel loci robustly associated with free thyroxine (FT4) level ($P < 5 \times 10^{-8}$) (Teumer *et al.*, 2018) (Supplementary Table S1). To estimate the unbiased LD correlation the reference LD sample with individual genotype should be from either one of the participating studies of the meta-analysis, or from genotype data of the same general population with a size of at least 2,000 (Yang *et al.*, 2011). Since SHIP study is one of the participating cohorts in this GWAMA, we used SHIP imputed individual-level genotype data as an LD reference sample. We considered 21 associated top SNPs as index SNPs to find additional secondary signals. By applying the *G5* function, 10 secondary signals were discovered with GCTA *P*-value smaller than SNP-specific α-threshold (Table 3 and Supplementary Figs S1–S10). The method found five secondary signals rs182320282 *(AADAT)* on chromosome 4, rs1506085 *(SIM1)* on chromosome 6, rs7038480 *(TRMO)* on chromosome 9, rs56067456 *(NRXN3)* on chromosome 14 and rs1791197 *(TTR)* on chromosome 18 that would not have been found with the standard approach ($\alpha = 5 \times 10^{-8}$) (Fig. 2A–E or Supplementary Figs S2, S3, S5, S9 and S10). In order to investigate the effect of the LD reference sample on our method, we used a larger reference sample from the UK Biobank dataset (Bycroft *et al.*, 2018) with 13,558 European individuals and 16,969,363 SNPs. The UKBB reference sample was checked for cryptic relatedness and population stratification as described in Teumer *et al.* (2019) and Wuttke *et al.* (2019). Two index SNPs rs6854291 on chromosome 4 and rs11039355 on chromosome 11 (Supplementary Table S1) were missing in UKBB reference sample and the method was performed on 19 out of 21 index SNPs. *G5* function found 9 significant secondary signals (Table 3 and Supplementary Figs S11–19) of those 4 secondary signals rs13205255 *(SIM1)* on chromosome 6, rs7038480 *(TRMO)* on chromosome 9, rs10841679 *(SLCO1B3)* on chromosome 12 and rs150816132 *(NRXN3)* on chromosome 14 were found only by our method (Fig. 2F–G or Supplementary Figs S12, S14, S16 and S18). *G5* could find the secondary signals with the same nearest gene for 8 index SNPs rs2235544 *(DIO1)* on chromosome 1, rs17185536 *(LOC728012)* on chromosome 6,

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

Table 3. Summary of secondary signals identified by G5-function using SHIP and UKBB LD reference samples-FT4

Chr	Secondary signal	Locus-1	Index SNP	Locus-2	D[bp]	r^2-SHIP	r^2-UKBB	*P*-SHIP	*P*-UKBB	α-SHIP	α-UKBB
1	rs2294512	*DIO1*	rs2235544	*DIO1*	−5,264	0.044	0.049	4.48E-11	1.19E-11	7.49E-06	8.66E-06
4	rs182320282[a]	*AADAT*	rs6854291	*AADAT*	48,622	0.001	–	3.21E-06	–	4.13E-06	–
6	rs1506085[a]	*SIM1*	rs17185536	*LOC728012*	233,691	0.000	–	1.01E-07	–	1.30E-06	–
6	rs13205255[b]	*SIM1*	rs17185536	*LOC728012*	230,887	–	0.001	–	4.94E-07	–	1.50E-06
6	rs137964359	*TRIM38*	rs9356988	*SLC17A4*	224,261	0.004	0.004	2.58E-09	2.30E-09	1.30E-06	1.50E-06
9	rs7038480[c]	*TRMO*	rs10739496	*FOXE1*	109,492	0.082	0.049	1.09E-06	8.52E-08	1.34E-06	1.53E-06
9	rs2274115	*LHX3*	rs4842131	*LHX3*	2,094	0.549	0.412	2.70E-09	5.91E-09	7.51E-06	9.05E-06
12	rs7978253	*SLCO1B3*	rs4149056	*SLCO1B1*	−340,067	0.095	–	1.52E-09	–	1.35E-06	–
12	rs10841679[b]	*SLCO1B3*	rs4149056	*SLCO1B1*	−315,174	–	0.020	–	9.98E-08	–	1.51E-06
14	rs10220700	*LINC00524*	rs11626434	*DIO3OS*	−162,961	0.001	–	2.67E-09	–	1.30E-06	–
14	rs4274374	*LINC00524*	rs11626434	*DIO3OS*	−162,187	–	0.001	–	1.18E-08	–	1.50E-06
14	rs56067456[a]	*NRXN3*	rs225014	*DIO2*	−183,703	0.008	–	9.54E-07	–	1.31E-06	–
14	rs150816132[b]	*NRXN3*	rs225014	*DIO2*	−205,287	–	0.014	–	1.36E-06	–	1.51E-06
18	rs1791197[a]	*TTR*	rs113107469	*SLC25A52*	−117,280	0.021	–	1.75E-07	–	1.31E-06	–
18	rs145581407	*B4GALT6*	rs113107469	*SLC25A52*	−41,560	–	0.756	–	1.97E-14	–	4.39E-06

Note: The table contains the list of significant secondary signals with corresponding index SNPs. Locus-1, the closest gene to secondary signal; Locus-2, the closest gene to index SNP; D[bp] , the distance between index SNP and secondary signal; r^2-SHIP, LD correlation between index SNP and secondary signal using SHIP reference sample; r^2-UKBB, LD correlation between index SNP and secondary signal using UKBB reference sample; *P*-SHIP, GCTA *P*-value using SHIP reference sample; *P*-UKBB, GCTA *P*-value using UKBB reference sample; α-SHIP, SNP-specific α-threshold using SHIP reference sample; α-UKBB, SNP-specificα-threshold using UKBB reference sample.

[a] Secondary signal with GCTA *P*-value < SNP-specific α-threshold, was identified only by *G5* using SHIP reference sample.

[b] Secondary signal with GCTA *P*-value < SNP-specific α-threshold, was identified only by *G5* using UKBB reference sample.

[c] Secondary signal with GCTA *P*-value < SNP-specific α-threshold, was identified only by *G5* using SHIP and UKBB reference sample.

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

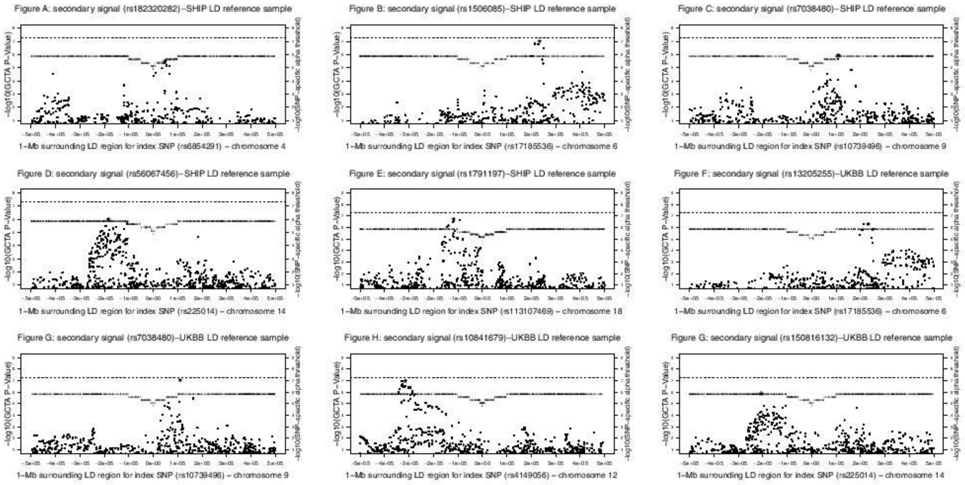


Fig. 2. (**A–G**) -$\log_{10}$ of the GCTA *P*-value (*y*-axis) for SNPs from 1-Mb surrounding LD region of the index SNP (*x*-axis). Lower curve indicates -$\log_{10}$ of the SNP-specific α-threshold for each SNP from surrounding LD region and upper dashed line is -$\log_{10}(5 \times 10^{-8})$. The secondary signal was discovered exclusively by *G5*-function is shown by asterisk. SHIP is the LD reference sample for figures A–E and UKBB is the LD reference sample for figures F–G

rs9356988 *(SLC17A4)* on chromosome 6, rs10739496 *(FOXE1)* on chromosome 9, rs4842131 *(LHX3)* on chromosome 9, rs4149056 *(SLCO1B1)* on chromosome 12, rs11626434 *(DIO3OS)* on chromosome 14 and rs225014 *(DIO2)* on chromosome 14 using two reference samples. More precisely for 4 index SNPs rs2235544 *(DIO1)* on chromosome 1, rs9356988 *(SLC17A4)* on chromosome 6, rs10739496 *(FOXE1)* on chromosome 9 and rs4842131 *(LHX3)* on chromosome 9 exactly the same secondary signals were discovered.

As the second test example, we used meta-analysis summary data of a GWAS on inflammatory bowel disease (IBD) (Liu *et al.*, 2015) with 11,555,662 genetic variants and up to 34,652 European individuals. We defined a locus as a chromosomal region at which two adjacent significant SNPs are less than 1-Mb distant. By our definition we identified 42 genome-wide significance (*P*-value $< \alpha = 5 \times 10^{-8}$) index SNPs from IBD meta-analysis summary data (Supplementary Table S2). Our method implemented on 39 out of 42 index SNPs since the index SNPs rs10800314 on chromosome 1, rs10175585 on chromosome 2 and rs131657 on chromosome 22 were missing in the UKBB reference sample. We discovered 12 index SNPs with significance secondary signals by *G5* function (Table 4 and Supplementary Figs S20–S31) of those 4 secondary signals rs12128452 *(RNF186)* on chromosome 1, rs7797798 *(SLC26A3)* on chromosome 7, rs4880099 *(NOTCH1)* on chromosome 9 and rs2129944 *(CDC37)* on chromosome 19 were found exclusively by our method (Supplementary Figs S21, S25, S26 and S30). In the following, the method was implemented for current example test using SHIP reference sample. Correspondingly 13 secondary signals were found using *G5* function (Table 4 and Supplementary Figs S32–S44) of those 5 secondary signals rs4233371 *(FCGR2A)* on chromosome 1, rs10185424 *(IL1R2)* on chromosome 2, rs7797798 *(SLC26A3)* on chromosome 7, rs28668598 *(JAK2)* on chromosome 9 and rs56380902 *(GSDMB)* on chromosome 17 were identified only by our method (Supplementary Figs S34, S35, S38, S39 and S42). *G5* function identified 8 significant secondary signals for index SNPs rs11209026 *(IL23R)* on chromosome 1, rs10737481 *(RNF186)* on chromosome 1, rs10045431 *(LOC285626)* on chromosome 5, rs4730272 *(SLC26A3)* on chromosome 7, rs4077515 *(CARD9)* on chromosome 9, rs2076756 *(NOD2)* on chromosome 16, rs12936409 *(ZPBP2)* on chromosome 17 and rs6062496 *(RTEL1-TNFRSF6B)* on chromosome 20 with the same nearest genes using both reference samples. Particularly for 6 index SNPs rs10737481 *(RNF186)*, rs4730272 *(SLC26A3)*, rs4077515 *(CARD9)*, rs2076756 *(NOD2)*, rs12936409 *(ZPBP2)* and rs6062496 *(RTEL1-TNFRSF6B)* exactly the same secondary signals were identified by using SHIP and UKBB reference samples.

As the third example, we considered the GWAS meta-analyses summary data by Wood *et al.* (2014) with 253,288 European individuals and 2,550,859 variants on adult human height. 697 secondary signals clustered in 423 loci were identified with genome-wide significance ($\alpha = 5 \times 10^{-8}$) using conditional GCTA (COJO) analysis in this investigation. We applied our method on 386 genome-wide significance ($\alpha = 5 \times 10^{-8}$) index regions identified by our definition on GWAS meta-analyses results (Supplementary Table S3). Using UKBB reference sample, 134 significant secondary signals with conditional *P*-value < SNP-specific α-threshold were identified by *G5* function (Supplementary Table S4). Exclusively, our method was able to detect 19 out of 134 secondary signals that would have gone undetected with the established genome-wide threshold (Supplementary Table S4 and Supplementary Figs S45–S66).

5 Conclusion

We presented a new method to assess the significance of secondary SNPs using SNP location and LD to define SNP-specific significance thresholds in a weighted hypothesis testing framework. With this procedure, we remove the current over-counting of the number of hypotheses considered and in addition, prioritize the hypotheses by plausible criteria. Via a simulation study, we confirmed the validity of the approach by evaluating EFWER. A series of different weighting schemes defined by respective *G*-functions showed improved power as compared to established criterion as well as in comparison to an equal weighting scheme (Sidak-correction). A *G*-function with optimal $r^2 = 0.3$ between index and secondary SNP which down-weighted SNPs at higher distance step-wise-strong and gave more emphasis on *d* than r^2 had overall best power and is our recommended default. In addition to the demonstrated power gain, our method is easy to use and can directly be applied to typically already existing GCTA results. Via re-analysis of existing GWAMAs, we found secondary signals that otherwise would have been overlooked: 5 signals in the case of FT4 levels, 5 in the case of IBD and

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

Table 4. Summary of secondary signals identified by G5-function using SHIP and UKBB LD reference samples-IBD

Chr	Secondary signal	Locus-1	Index SNP	Locus-2	D[bp]	r^2-SHIP	r^2-UKBB	*P*-SHIP	*P*-UKBB	α-SHIP	α-UKBB
1	rs10889669	*IL23R*	rs11209026	*IL23R*	−30,230	0.019	–	3.29E-30	–	2.01E-06	–
1	rs11805303	*IL23R*	rs11209026	*IL23R*	−30,442	–	0.029	–	9.70E-27	–	2.25E-06
1	rs12128452[a]	*RNF186*	rs10737481	*RNF186*	−27,808	0.000	0.001	1.08E-08	1.37E-07	1.99E-06	2.23E-06
1	rs4233371[b]	*FCGR2A*	rs10800314	*FCGR2A*	336	0.212	–	1.71E-06	–	6.79E-06	–
2	rs10185424[b]	*IL1R2*	rs10175585	*IL18RAP*	−413,219	0.006	–	5.31E-08	–	6.29E-07	–
5	rs6897022	*LINC00603*	rs6880778	*PTGER4*	−62,550	0.522	–	4.77E-17	–	1.15E-06	–
5	rs116437514	*PTGER4*	rs6880778	*PTGER4*	−26,376	–	0.008	–	1.30E-19	–	2.23E-06
5	rs56167332	*LOC285626*	rs10045431	*LOC285626*	13,236	0.042	–	6.06E-12	–	2.03E-06	–
5	rs6871626	*LOC285626*	rs10045431	*LOC285626*	12,259	–	0.039	–	4.58E-12	–	2.26E-06
6	rs117292830	*HLA-C*	rs148844907	*C6orf47*	−410,129	–	0	–	3.80E-14	–	7.01E-07
7	rs7797798[c]	*SLC26A3*	rs4730272	*SLC26A3*	−31,176	0.006	0.002	2.90E-07	1.68E-07	2.00E-06	2.23E-06
9	rs28668598[b]	*JAK2*	rs1887428	*JAK2*	−15,424	0.009	–	1.83E-06	–	2.00E-06	–
9	rs4880099[a]	*NOTCH1*	rs4077515	*CARD9*	144,093	0.085	0.094	1.36E-08	3.08E-07	6.48E-07	7.26E-07
12	rs1388585	*LINC02471*	rs140892874	*MUC19*	−293,107	–	0.066	–	3.07E-10	–	7.19E-07
16	rs146528649	*NKD1*	rs2076756	*NOD2*	−95,917	0.033	0.035	6.66E-12	1.52E-11	1.13E-06	1.27E-06
17	rs56380902[b]	*GSDMB*	rs12936409	*ZPBP2*	22,723	0.784	0.808	5.33E-08	1.04E-09	1.80E-06	1.97E-06
19	rs60542850	*TYK2*	rs142770866	*PDE4A*	−37,012	0.006	–	2.19E-08	–	2.00E-06	–
19	rs2129944[a]	*CDC37*	rs142770866	*PDE4A*	−9,174	–	0.011	–	5.68E-08	–	3.99E-06
20	rs6089763	*RTEL1*	rs6062496	*RTEL1-TNFRSF6B*	−20,745	0.158	0.116	4.50E-08	4.88E-08	2.11E-06	2.32E-06

Note: The table contains the list of significant secondary signals with corresponding index SNPs. Locus-1: the closest gene to secondary signal; Locus-2: the closest gene to index SNP; D[bp]: the distance between index SNP and secondary signal; r^2-SHIP: LD correlation between index SNP and secondary signal using SHIP reference sample; r^2-UKBB: LD correlation between index SNP and secondary signal using UKBB reference sample; *P*-SHIP: GCTA *P*-value using SHIP reference sample; *P*-UKBB: GCTA *P*-value using UKBB reference sample; α-SHIP: SNP-specificα-threshold using SHIP reference sample; α-UKBB: SNP-specificα-threshold using UKBB reference sample.

[a] Secondary signal with GCTA *P*-value < SNP-specific α-threshold, was identified only by *G5* using UKBB reference sample.

[b] Secondary signal with GCTA *P*-value < SNP-specific α-threshold, was identified only by *G5* using SHIP reference sample.

[c] Secondary signal with GCTA *P*-value < SNP-specific α-threshold, was identified only by *G5* using SHIP and UKBB reference sample.

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

19 signals for human height. For FT4, we found exclusively, for instance, the secondary signals rs1791197 (*TTR*) and rs7038480 (*TRMO*) (Table 3). Transthyretin (*TTR*) belongs to a group of proteins, which includes thyroxine-binding globulin and albumin, that bind to and transport thyroid hormones in the blood (Power *et al.*, 2000). *TRMO* is ubiquitously expressed in thyroid (Fagerberg *et al.*, 2014). In case of IBD, The 'quasi adaptive' method uniquely found secondary signals rs7797798 (*SLC26A3*) and rs56380902 *(GSDMB)* (Table 4). *SLC26A3* express in colon and the protein encoded by this gene is essential for intestinal chloride absorption, and mutations in this gene have been associated with congenital chloride diarrhea (Haggie *et al.*, 2018). Söderman *et al.* (2015) inspected the biological foundation of IBD and showed that *GSDMB* affects IBD susceptibility via effects on apoptosis and cell proliferation. These examples demonstrate that identification of secondary signals is not only relevant since it can increase the portion of explained variance of a genetic trait, but also since it can point to the functional mechanisms underlying the primary and secondary association signals.

In analysis of two GWAS examples (FT4 and IBD), we used two different LD reference panels both from the European population to evaluate the influence of the LD reference sample on the proposed method. Secondary signals obtained with the SHIP reference sample, showed overall good agreement with those obtained using UKBB. The 'quasi adaptive' method could find a reasonable number of secondary signals with the same nearest gene for both reference LD panels such that, for some index SNPs exactly with the same secondary signals. In some cases, we obtained different results depending on the choice of the LD reference panel. While tendencies were always similar, it happened that SNPs lay slightly above the significance threshold with one LD panel, but slightly below with another one. For instance, In case of FT4 GWAS analysis (Table 3) secondary signals rs1791197 (*TTR*) and rs145581407 (*B4GALT6*) on chromosome 18 were found for index SNP rs113107469 *(SLC25A52)* by SHIP and UKBB respectively. Furthermore, the secondary signal for index SNP rs6854291 *(AADAT)* on chromosome 4 was found only with SHIP, since rs6854291 was missing in UKBB. In IBD analysis results (Table 4) the secondary signal rs28668598 *(JAK2)* on chromosome 9 was found only by SHIP and secondary signals rs117292830 *(HLA-C)* on chromosome 6 and rs1388585 *(LINC02471)* on chromosome 12 were found only with UKBB reference sample. Index SNPs rs6880778 *(PTGER4)* on chromosome 5 and rs142770866 *(PDE4A)* on chromosome 19 have different secondary signals with different closest genes to the secondary signals for two reference samples. In addition, secondary signals for index SNPs rs10800314 *(FCGR2A)* on chromosome 1 and rs10175585 *(IL18RAP)* on chromosome 2 were found exclusively by SHIP since two index SNPs were missing in UKBB reference sample. Taken together, results are to some extent dependent on the reference sample which determines SNP availability and the LD structure estimated from it. In this context, Yang *et al.* (2011) recommended reference LD panel from the same population as the study data itself comes from and beyond a sample size of 5,000 for additional accuracy. Finally, we considered the GWAS on human height and note that, the authors pursued a modified approach of identifying primary and secondary signals, the conditioned SNPs on index signals from the whole genome rather than within LD regions. In our re-evaluation of the summary statistics, we applied our LD region-based approach, as used above and in the majority of GWAS publications. We found 386 primary signals and 134 secondary signals. Of the secondary signals, 19 were found exclusively with our approach (Supplementary Table S4).

In all of three data applications, as well as in a power simulation study, our method demonstrated improved performance when compared to current practice. In addition, our method has further potential for additional improvement, e.g. by not treating potential secondary signals as independent of each other in their multiplicity assessment. In summary, our method has the potential to reveal previously undetected secondary signals in already available data, and to uncover plausible underlying gene mechanisms. The method is easy to use, operates directly with typically already existing GWAMA results and makes use of existing analysis software (Yang *et al.*, 2011). The specific method presented here is implemented in the R and Shell scripts which can be found at [https://github.com/sghasemi64/Secondary-Signal].

Acknowledgements

SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research (grants no. 01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania. Genome-wide data have been supported by the Federal Ministry of Education and Research (grant no. 03ZIK012) and a joint grant from Siemens Healthineers, Erlangen, Germany and the Federal State of Mecklenburg- West Pomerania. The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH. This research has been conducted using the UK Biobank Resource under Application Number 20272.

Funding

The work was funded by project BE 3828/3-3 from the Deutsche Forschungsgemeinschaft (DFG).

Data availability: FT4 summary association results are available for full download on the 836 CHARGE dbGaP website under accession phs000930 [https://www.ncbi.nlm.nih.gov/gap] 837 and in the locuszoom web page: http://locuszoom.sph.umich.edu/genform.php. IBD summary association results are available for full download at https://www.ibdgenetics.org/downloads.html. Human Height summary association results are available for full download at http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files

Conflict of Interest: The authors declare that there is no conflict of interest.

References

Becker,T. and Herold,C. (2009) Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. *Eur. J. Hum. Genet.*, **17**, 1043–1049.

Bycroft,C. *et al.* (2018) The uk biobank resource with deep phenotyping and genomic data. *Nature*, **562**, 203–209.

Clarke,L. *et al.*; 1000 Genomes Project Consortium. (2012) The 1000 genomes project: data management and community access. *Nat. Methods*, **9**, 459–462.

Demets,D.L. and Lan,K.G. (1994) Interim analysis: the alpha spending function approach. *Stat. Med.*, **13**, 1341–1352.

Fagerberg,L. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics*, **13**, 397–406.

Fritsche,L.G. *et al.* (2016) A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.*, **48**, 134–143.

Haggie,P.M. *et al.* (2018) Slc26a3 inhibitor identified in small molecule screen blocks colonic fluid absorption and reduces constipation. *JCI Insight*, **3**, e121370.

Herold,C. *et al.* (2009) Intersnp: genome-wide interaction analysis guided by a priori information. *Bioinformatics*, **25**, 3275–3281.

Howey,R. and Cordell,H.J. (2014) Imputation without doing imputation: a new method for the detection of non-genotyped causal variants. *Genet. Epidemiol.*, **38**, 173–190.

Kang,G. *et al.* (2009) Weighted multiple hypothesis testing procedures. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–22.

Liu,J.Z. *et al.*; International IBD Genetics Consortium. (2015) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.*, **47**, 979–986.

Lu,X., GLGC Consortium. *et al.* (2017) Exome chip meta-analysis identifies novel loci and east Asian-specific coding variants that contribute to lipid levels and coronary artery disease. *Nat. Genet.*, **49**, 1722–1730.

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

Power,D. *et al.* (2000) Evolution of the thyroid hormone-binding protein, transthyretin. *Gen. Compar. Endocrinol.*, **119**, 241–255.

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Söderman,J. *et al.* (2015) Gene expression-genotype analysis implicates GSDMA, GSDMB, and LRRC3C as contributors to inflammatory bowel disease susceptibility. *BioMed Res. Int.*, **2015**, 834805.

Teumer,A. *et al.* (2018) Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. *Nat. Commun.*, **9**, 4455.

Teumer,A. *et al.* (2019) Genome-wide association meta-analyses and fine--mapping elucidate novel pathways influencing albuminuria. Nat. Commun., **10**, 1–19.

Völzke,H. *et al.* (2011) Cohort profile: the study of health in pomerania. *Int. J. Epidemiol.*, **40**, 294–307.

Willer,C.J. *et al.* (2010) Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, **26**, 2190–2191.

Wood,A.R. *et al.*; LifeLines Cohort Study. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.*, **46**, 1173–1186.

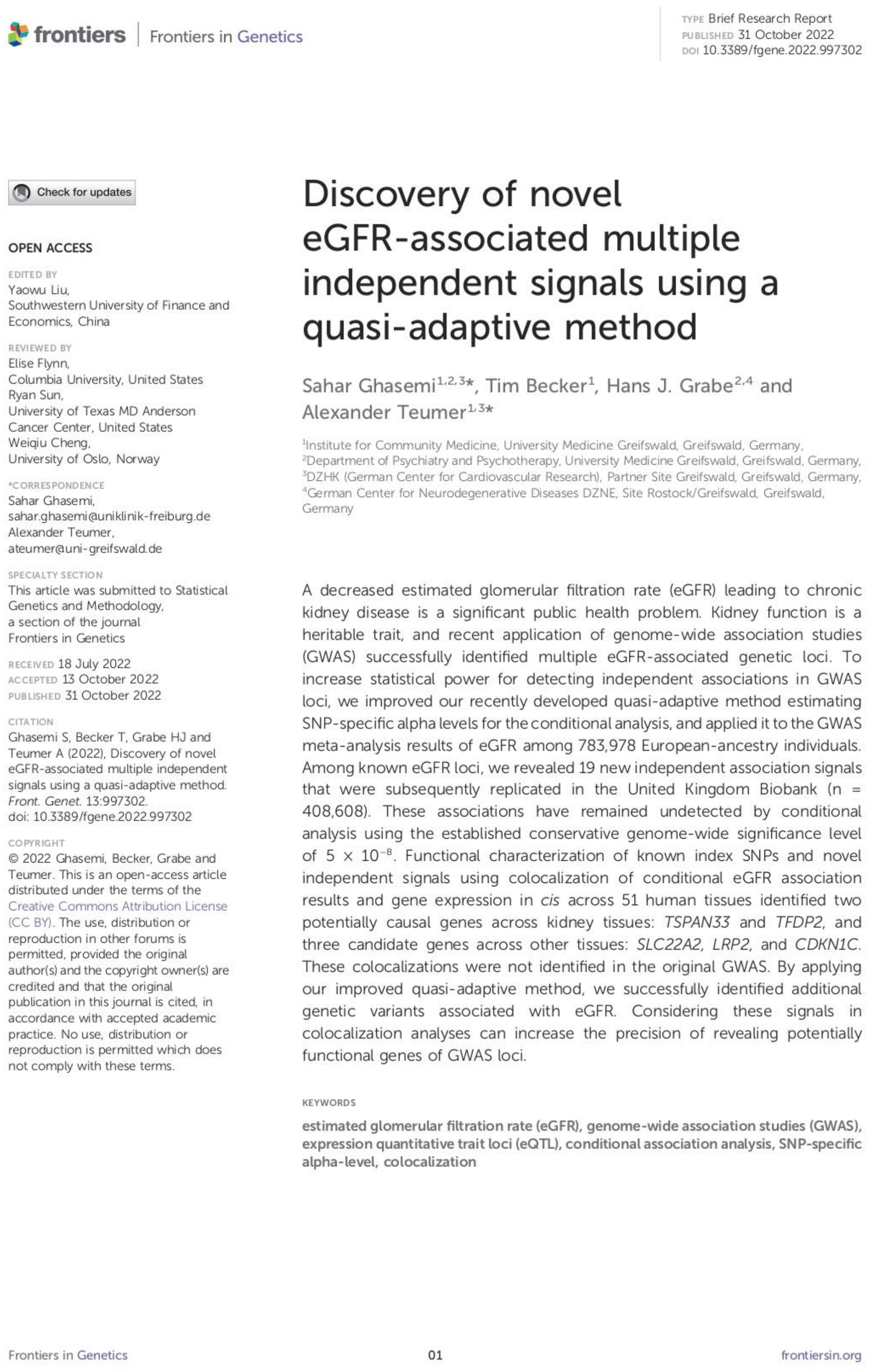
Wuttke,M. *et al.*; V. A. Million Veteran Program. (2019) A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.*, **51**, 957–972.

Yang,J. *et al.* (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.

Yang,J. *et al.*; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **44**, 369.

Downloaded from https://academic.oup.com/bioinformatics/article/37/20/3521/6274650 by UB Greifswald user on 26 January 2022

Discovery of novel eGFR-associated multiple independent signals using a quasi-adaptive method (Article III)



frontiers | Frontiers in Genetics

TYPE Brief Research Report
PUBLISHED 31 October 2022
DOI 10.3389/fgene.2022.997302

Check for updates

Discovery of novel eGFR-associated multiple independent signals using a quasi-adaptive method

Sahar Ghasemi[1,2,3]*, Tim Becker[1], Hans J. Grabe[2,4] and Alexander Teumer[1,3]*

[1]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany,
[2]Department of Psychiatry and Psychotherapy, University Medicine Greifswald, Greifswald, Germany,
[3]DZHK (German Center for Cardiovascular Research), Partner Site Greifswald, Greifswald, Germany,
[4]German Center for Neurodegenerative Diseases DZNE, Site Rostock/Greifswald, Greifswald, Germany

A decreased estimated glomerular filtration rate (eGFR) leading to chronic kidney disease is a significant public health problem. Kidney function is a heritable trait, and recent application of genome-wide association studies (GWAS) successfully identified multiple eGFR-associated genetic loci. To increase statistical power for detecting independent associations in GWAS loci, we improved our recently developed quasi-adaptive method estimating SNP-specific alpha levels for the conditional analysis, and applied it to the GWAS meta-analysis results of eGFR among 783,978 European-ancestry individuals. Among known eGFR loci, we revealed 19 new independent association signals that were subsequently replicated in the United Kingdom Biobank (n = 408,608). These associations have remained undetected by conditional analysis using the established conservative genome-wide significance level of 5×10^{-8}. Functional characterization of known index SNPs and novel independent signals using colocalization of conditional eGFR association results and gene expression in *cis* across 51 human tissues identified two potentially causal genes across kidney tissues: *TSPAN33* and *TFDP2*, and three candidate genes across other tissues: *SLC22A2*, *LRP2*, and *CDKN1C*. These colocalizations were not identified in the original GWAS. By applying our improved quasi-adaptive method, we successfully identified additional genetic variants associated with eGFR. Considering these signals in colocalization analyses can increase the precision of revealing potentially functional genes of GWAS loci.

KEYWORDS

estimated glomerular filtration rate (eGFR), genome-wide association studies (GWAS), expression quantitative trait loci (eQTL), conditional association analysis, SNP-specific alpha-level, colocalization

OPEN ACCESS

EDITED BY
Yaowu Liu,
Southwestern University of Finance and Economics, China

REVIEWED BY
Elise Flynn,
Columbia University, United States
Ryan Sun,
University of Texas MD Anderson Cancer Center, United States
Weiqiu Cheng,
University of Oslo, Norway

*CORRESPONDENCE
Sahar Ghasemi,
sahar.ghasemi@uniklinik-freiburg.de
Alexander Teumer,
ateumer@uni-greifswald.de

SPECIALTY SECTION
This article was submitted to Statistical Genetics and Methodology,
a section of the journal
Frontiers in Genetics

RECEIVED 18 July 2022
ACCEPTED 13 October 2022
PUBLISHED 31 October 2022

CITATION
Ghasemi S, Becker T, Grabe HJ and Teumer A (2022), Discovery of novel eGFR-associated multiple independent signals using a quasi-adaptive method.
Front. Genet. 13:997302.
doi: 10.3389/fgene.2022.997302

COPYRIGHT
© 2022 Ghasemi, Becker, Grabe and Teumer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Introduction

Glomerular filtration rate estimated from serum creatinine (eGFR) is used to quantify kidney function and define chronic kidney disease (CKD). CKD defined by low eGFR <60 ml/min/1.73 m^2 is strongly associated with an increased risk of major adverse clinical outcomes such as end-stage kidney disease (ESKD), cardiovascular (CV) outcomes, and mortality (Go et al., 2004; Chronic Kidney Disease PrognosisMatsushita et al., 2010; Hemmelgarn et al., 2010; Astor et al., 2011; Bello et al., 2011; Gansevoort et al., 2011; Gansevoort et al., 2013; Weiner et al., 2014; Matsushita et al., 2015). A better understanding of the biological mechanisms underlying kidney function is a prerequisite for initiating targeted treatments and reducing patient mortality, comorbidity, and associated healthcare costs. eGFR is a heritable trait with estimated $h^2 = 39\%$, and recent application of genome-wide association studies (GWAS) successfully identified multiple eGFR-associated genetic loci (Okada et al., 2012; Pattaro et al., 2012; Mahajan et al., 2016; Pattaro et al., 2016; Hishida et al., 2018; Kanai et al., 2018; Lee et al., 2018; Wuttke et al., 2019). Allelic heterogeneity within a GWAS locus is a common characteristic of complex traits and conditional analyses successfully identified multiple independent associations with eGFR. For instance, Gorski et al. (2017) (Gorski et al., 2017) detected 57 independent signals among the 49 loci. Morris et al. (2019) (Morris et al., 2019) delineated 127 distinct signals across the 93 loci. Hellwege et al. (2019) (Hellwege et al., 2019) discovered 18 independent signals at 15 loci, and Wuttke et al. (2019) (Wuttke et al., 2019) identified 253 independent SNPs at 228 loci explaining 7.3% of the eGFR variation.

To identify an independent signal, the SNPs of a locus are conditioned by the known significant associations. In case individual genotypes of a sample are available, the genotypes of known signals are added as covariates to the association model. Alternatively, these conditional associations can be approximated by using summary statistics and an appropriate linkage disequilibrium (LD) panel. Usually, the established genome-wide significance level of 5×10^{-8} was applied as a significance threshold for the conditional analysis, which is also the significance level for the primary GWAS. Since the conditional analysis is applied on a specific genomic region and not on a genome-wide scale, 5×10^{-8} is too conservative and implies a loss of power. In Ghasemi et al. (2021) (Ghasemi et al., 2021), we developed a quasi-adaptive method to determine SNP-specific significance levels in conditional analysis.

Although GWAS have discovered multiple eGFR-associated loci, the underlying genes that influence genetic associations have often remained unknown. Integration of GWAS signals and expression quantitative trait loci (eQTL) studies (Nica and Dermitzakis, 2013) to estimate the relation between gene expression of nearby genes and eGFR, termed colocalization (Giambartolomei et al., 2014), allows the identification of candidate genes and improves the functional interpretation of GWAS results. For instance, *FGF5*, *CDKL5*, *TPSAN33*, and *METTL10* colocalized with the eGFR-associated loci in kidney-specific tissues (Graham et al., 2019), and Wuttke et al. (2019) (Wuttke et al., 2019) detected 17 underlying genes expressed in kidney tissues including *UMOD*, *KNG1*, and *FGF5*.

Here, we improved and applied our quasi-adaptive method to the publically available GWAS meta-analysis results of 783,978 European-ancestry individuals (Wuttke et al., 2019) of the CKDGen Consortium to uncover additional independent signals for eGFR. Replication of the identified novel independent signals was conducted using individual-level participant data of the United Kingdom Biobank (UKBB) (Bycroft et al., 2018). The UKBB was not included in the primary GWAS meta-analysis, and thus represents an independent dataset for replication. We run colocalization analyses based on associations with eGFR and with gene expression (eQTLs) in *cis* across 49 human tissues included in the Genotype-Tissue Expression (GTEx) project v8[27], as well as the microdissected human glomerular and tubulo-interstitial kidney portions from 187 individuals from the NEPTUNE study (Gillies et al., 2018). Since the presence of multiple independent signals within a GWAS locus reduces power of colocalization, we provided the colocalization analyses with conditional eGFR-association analysis and eQTL to detect potential causal genes and compared these results to the unconditional approach. Our colocalization analyses used the latest version of GTEx-v8 compared to the GTEx-v6 in the previous report of eGFR (Wuttke et al., 2019).

The emerging list of novel eGFR-associated variants and genes influencing kidney disease etiology facilitate CKD targeted treatment and prevention.

Methods

Additional independent eGFR-associated signals identification by quasi-adaptive method

We obtained the CKDGen Consortium 2019 eGFR-association GWAS meta-analysis results for European-ancestry (Wuttke et al., 2019) from https://ckdgen.imbi.uni-freiburg.de. The downloaded file included chromosome, position (b37), SNP rsid, effect allele, non-effect allele, effect allele frequency, beta, standard error, *p*-value, and sample size for each variant. Wuttke et al. (2019) (Wuttke et al., 2019) identified 253 independent genome-wide-significant eGFR-associated SNPs through approximate conditional analyses implemented in GCTA (Yang et al., 2011) (GCTA COJO Slct algorithm) across 228 European-ancestry-specific and replicated loci. To identify additional independent eGFR-associated secondary signals, we applied our quasi-adaptive method to the aforementioned

GWAS meta-analysis with 8,885,712 genetic variants and 783,978 individuals. The method incorporated LD structure from individual-level genotype data of 15,000 randomly selected European-ancestry participants of the UKBB (Bycroft et al., 2018). The selected UKBB LD reference sample underwent the same data preparation procedure as described in (Wuttke et al., 2019) and (Teumer et al., 2019), except for the minor allele frequency (MAF) cut-off. We excluded SNPs with a MAF <0. 0001. The final dataset for estimating the LD structure included 13,558 unrelated European-ancestry individuals and 36, 228, 692 genetic variants. We used the published 228 replicated index SNPs (i.e., variants with the smallest p-value of a locus) as the basis for applying our method (Wuttke et al., 2019). A one megabase window around the index SNPs was considered as primary loci. Overlapping loci at which two adjacent index SNPs were less than one megabase apart or with pairwise correlation $r^2 > 0.1$ were merged using the lower-bound and the upper-bound of the merged regions as new locus borders, and the SNP with the smallest p-value as the new index SNP. This resulted in a final list of 190 independent loci (Supplementary Table S1). All SNPs except the index SNP were considered candidate SNPs within each locus. We conducted conditional analyses on this dataset using GCTA (GCTA COJO-cond algorithm) by adjusting for the corresponding index SNP across the 190 loci. The number of tested SNPs equals to the number of candidate SNPs included in the conditional analyses across the 190 loci. As described in Ghasemi et al. (2021) (Ghasemi et al., 2021), our method prioritizes the candidate SNPs and assigns a SNP-specific α-threshold to the candidate SNPs in conditional analysis. The pairwise correlation (r^2) and chromosomal distance (d) between the candidate SNPs and respective index SNP needed as inputs for our method were retrieved by the INTERSNP tool (Herold et al., 2009). Let m_2 be the number of tested SNPs from N_2 loci (here, $N_2 = 190$ with the index reflecting the analysis of secondary signals). Of note, m_2 and N_2 were named as m and N in the original paper (Ghasemi et al., 2021). The pre-weight based on r^2 ($w_{r_i^2}$) with optimal $r^2 = 0.3$ and a pre-weight based on d (w_{d_i}) which down-weighted SNPs at higher distance step-wise-strong are assigned to a candidate SNP(i), $(1 \le i \le m_2)$ as:

$$w_{r_i^2} = \frac{1 - \left|r_i^2 - 0.3\right| - 0.3}{1 - 0.3},$$

$$w_{d_i} = \begin{cases} 1 & if\ 0 < d \le 1\text{Kb} \\ 0.5 & if\ 1\text{Kb} < d \le 10\text{Kb} \\ 0.25 & if\ 10\text{Kb} < d \le 50\text{Kb} \\ 0.125 & if\ 50\text{Kb} < d \le 100\text{Kb} \\ 0.0625 & if\ 100\text{Kb} < d \le 500\text{Kb} \end{cases}$$

The pre-weight $w_{r_i^2}$ and w_{d_i} are combined (with more emphasis on d than on r^2) by the geometric mean $w_i = (w_{d_i}^k \times w_{r_i^2})^{\frac{1}{k+1}}$, with $k = 5$, to assign an optimal weight $W_i = \frac{w_i \times m_2}{\sum_{i=1}^{m_2} w_i}$ to SNP(i).

The quasi-adaptive method is applied on N_2 loci, spends type I error rate (α) over m_2 candidate SNPs by incorporating W_i into the weighted Šidák correction (Kang et al., 2009), and assigns the SNP-specific α-thresholds to SNP(i) by $G_i(\alpha, r^2, d)$ as follows:

$$G_i(\alpha, r^2, d) = 1 - (1 - \alpha)^{\frac{W_i}{m_2}},\ i = 1, 2, \ldots, m_2 \tag{1}$$

SNP(i) is a secondary signal if the conditional p-value is smaller than $G_i(\alpha, r^2, d)$.

(Ghasemi et al., 2021) showed that Equation 1 has the overall best power in detecting secondary signals while controlling the family-wise error rate (FWER) at the α-level. In our study, α was set to 0.05.

Improved quasi-adaptive method to identify multiple independent eGFR-associated signals

The original quasi-adaptive method was developed to determine one independent signal (secondary signal) with the smallest conditional p-value smaller than the correspondingly assigned $G(\alpha, r^2, d)$ at each locus. We extended the idea from the main paper (Ghasemi et al., 2021) to identify multiple independent signals (a tertiary signal, a signal of fourth, a signal of fifth, and beyond). To detect independent tertiary signals, only loci with confirmed secondary signals (confirmed according to the quasi-adaptive method) were considered. We proceeded according to the idea of the paper (Ghasemi et al., 2021) but performed conditional analyses by adjusting for the primary index SNP and confirmed secondary signal for each locus. Let N_3 be the number of loci with confirmed secondary signals and m_3 be the number of tested SNPs from N_3 loci (i.e., excluding index SNPs and secondary signals). Of note, the number of tested SNPs is lower for tertiary signals detection than for secondary signals detection ($m_3 < m_2$). As described in 2.1, the LD structure was determined between the index SNP and corresponding candidate SNPs at each locus. Our method was applied on N_3 loci according to the schema described in 2.1 and the SNP-specific α-thresholds assigned to SNP(i) by equation (2)

$$G_i(\alpha, r^2, d) = 1 - (1 - \alpha)^{\frac{W_i}{m_3}},\ W_i = \frac{w_i \times m_3}{\sum_{i=1}^{m_3} w_i},\ i = 1, 2, \ldots, m_3 \tag{2}$$

The improved method is an iterative process that is subsequently performed to detect higher-order independent signals (applied to loci with confirmed independent signals from the previous steps) until no additional independent signals are found. Finding higher-order independent signals keeps the FWER at the α-level because only the number of tested SNPs and the LD structure have to be taken into account (as shown in Equations 1, 2, where the LD structure does not change by analyzing higher-order independent signals.

Due to the complexity of the LD structure of the major histocompatibility complex (MHC) region, this region was excluded from the search for independent signals as also in the main GWAS (Wuttke et al., 2019).

Replication of the results in the UK biobank dataset

The novel independent eGFR-associated signals were tested for replication by conditional association analyses using the individual-level data of the UKBB (Bycroft et al., 2018) cohort. This cohort was not included in the initial GWAS of eGFR, and thus represents an independent dataset for replication. The phenotype definition, quality control, and analyses were performed using the same methods and scripts of the main GWAS (Teumer et al., 2019; Wuttke et al., 2019). As independent signals were identified from samples of European ancestry, conditional analyses were restricted to 408,608 UKBB participants of European ancestry with approximately 19 million autosomal SNPs that met the inclusion criteria of MAF ≥ 0.001 and imputation quality score > 0.3. For replication of each category of independent signals (secondary, tertiary, and beyond) across loci, a conditional analysis was conducted by including sex- and age-adjusted residual of log (eGFR), the first 15 genetic principal components, and the allele dosages of all corresponding conditioned SNPs as covariates in a mixed-model association method as implemented in BOLT-LMM, v2.3.2 (Loh et al., 2005). Within each locus, conditional analysis was performed for replication of an identified independent signal by conditioning on a known index SNP and (if present) on other known or replicated independent signals identified before the corresponding independent signal. Of note, non-replicated signals identified before the independent signal under investigation were excluded from the conditional analysis. Supplementary Table S2 shows the list of known index SNPs and known and novel independent signals with the list of covariates (SNPs) used for replication. Bonferroni correction of 0.05/9, 0.05/8, 0.05/6, 0.05/3, and 0.05, correcting for the number of tested SNPs per conditional analysis, was applied to assess the significance of the replication of secondary signals, tertiary signals, signals of fourth, signals of fifth, and signal of sixth, respectively.

Colocalization of eGFR signals with gene expression in cis

In the first instance, colocalization analyses were run for known index SNPs and novel independent signals using unconditional eGFR association analyses in the UKBB and expression quantitative trait (eQTL) studies (Nica and Dermitzakis, 2013). eQTL were quantified from 49 human tissues included in the GTEx project v8 release (Aguet et al., 2019), and the microdissected human glomerular and tubulointerstitial kidney portions from 187 individuals from the NEPTUNE study (Gillies et al., 2018). For colocalization, the effect alleles for GWAS and eQTLs were harmonized, and tissue gene pairs with eQTL data were identified within ± 100 kilobases of the independent signals. We used the eQTL *cis* window (1-megabase window from each side of the transcriptional start site) as the region for each colocalization test. We applied colocalization by using the approximate Bayes factor computations with the default prior probability = 1×10^{-5} on the signals available in both GWAS and eQTL as implemented in the coloc. fast function from the R package "gtx" version 2.1.6 (https://github.com/tobyjohnson/gtx). This function provides an adaptation of Giambartolomei's colocalization method (Giambartolomei et al., 2014).

Secondly, we re-run the colocalization analyses using conditional eGFR association analyses and the eQTL studies. Conditional analysis was performed for a known index SNP by adjusting for all known and novel independent signals and for a novel independent signal by conditioning on a known index SNP and (if present) on other known or novel independent signals within the corresponding locus. Supplementary Table S2 shows the list of covariates (SNPs) used in the eGFR association. We defined a variant as a colocalized signal (same causal variant underlying both the GWAS and eQTL association) if the posterior probability (PP) of a variant was greater than 80%.

Results

Novel eGFR-associated multiple conditionally independent signals

To detect additional eGFR-associated independent signals, our method was applied on 190 loci derived from the GWAS meta-analysis (Wuttke et al., 2019) (Methods and Supplementary Table S1). Our method identified in total 87 independent signals, including 53 secondary signals (Supplementary Table S3), 20 tertiary signals (Supplementary Table S4), 10 signals of fourth (Supplementary Table S5), three signals of fifth (Supplementary Table S6), and one signal of sixth (Supplementary Table S7), of which 27 were novel (Table 1). Of note, all novel SNPs were secondary or higher-order signals. We have listed the differences between the previous analysis (Wuttke et al., 2019) and our analysis in Supplementary Tables S3-S7 in a column labeled "Known". At a locus, an SNP detected by our method was considered known (yes) if it was exactly the independent signal or in high LD ($r^2 > 0.8$) with a SNP detected by Wuttke et al. (2019) (Wuttke et al., 2019). We detected 60 known loci, of which 54 loci comprised the same independent signal identified in the previous GWAS, and six loci with independent signals in high LD with the identified independent signals from the aforementioned GWAS.

TABLE 1 Summary of novel independent eGFR-associated signals identified by quasi-adaptive method and replication results.

—	Chr	Signal	Index	Closest gene	D [bp]	r^2	Pos (b37)	EA	EAF	GWAS-MA			GCTA			α-	Replication-UKBB		
										Effect	Se	P	Effect	Se	P	threshold	Effect	Se	P
Secondary signal	3	**rs147877018**	rs1397764	*TFDP2*	62,539	0.031	141,813,349	A	0.081	−0.0047	0.0007	1.18E-12	−0.0035	0.0007	1.13E-07	1.33E-07	−0.0042	0.0006	1.80E-11
	4	rs59664098	rs7667050	*PPARGC1A*	50,300	0.001	23,863,409	A	0.070	−0.0037	0.0007	2.30E-07	−0.0038	0.0007	8.66E-08	1.31E-07	−0.0011	0.0007	8.70E-02
	6	**rs3904600**	rs6921580	*RREB1*	−94,049	0.107	7,109,665	C	0.370	0.0027	0.0004	4.91E-14	0.0018	0.0003	9.75E-08	1.37E-07	0.0030	0.0004	1.80E-16
	7	rs12111979	rs700753	*LOC730338*	59,613	0.170	46,813,297	T	0.420	0.0004	0.0003	2.38E-01	0.0017	0.0003	8.07E-08	1.39E-07	0.0007	0.0004	4.70E-02
	8	**rs4566**	rs10086569	*SLC7A13*	−886,127	0.002	86,361,082	T	0.610	0.0020	0.0004	1.03E-08	0.0019	0.0004	5.90E-08	7.37E-08	0.0011	0.0003	1.50E-03
	12	rs2300127	rs11062167	*SLC6A13*	−49,290	0.011	315,449	T	0.570	0.0023	0.0004	1.62E-10	0.0019	0.0004	1.99E-07	2.35E-07	0.0008	0.0003	2.20E-02
	12	**rs11056376**	rs10846157	*RERG*	−17,637	0.020	15,307,394	A	0.910	0.0040	0.0007	5.43E-10	0.0034	0.0006	2.03E-07	2.36E-07	0.0021	0.0006	6.00E-04
	12	**rs3730071**	rs2634675	*ZNF641*	427,943	0.001	49,168,798	A	0.029	−0.0058	0.0011	1.86E-07	−0.0060	0.0011	7.20E-08	7.37E-08	−0.0040	0.0010	5.40E-05
	16	rs438339	rs113956264	*RPL3L*	6,421	0.001	2,003,425	T	0.880	0.0035	0.0007	5.36E-08	0.0034	0.0007	1.60E-07	4.17E-07	0.0010	0.0006	8.60E-02
Tertiary signal	2	**rs807574**	rs807624	*DDX1*	24,768	0.055	15,807,239	A	0.600	0.0011	0.0004	1.45E-03	0.0019	0.0003	8.09E-08	7.62E-07	0.0019	0.0004	9.90E-08
	7	**rs13227214**	rs3757387	*IRF5*	164,269	0.057	128,740,355	C	0.460	−0.0024	0.0003	1.12E-12	−0.0018	0.0003	5.64E-08	2.40E-07	−0.0027	0.0003	6.60E-15
	9	rs7035892	rs2039424	*PIP5K1B*	107,868	0.089	71,540,042	A	0.840	0.0053	0.0006	1.27E-17	0.0037	0.0006	1.23E-09	2.43E-07	−0.0016	0.0008	6.20E-02
	11	**rs81205**	rs233438	*KCNQ1*	4,412	0.261	2,798,804	A	0.540	0.0034	0.0004	4.73E-20	0.0018	0.0003	1.28E-07	1.45E-06	0.0020	0.0004	3.50E-07
	11	rs294345	rs3925584	*DCDC1*	−93,675	0.012	30,666,660	T	0.067	−0.0057	0.0007	2.32E-14	−0.0035	0.0007	3.01E-07	4.21E-07	−0.0001	0.0008	8.50E-01
	11	**rs1193692**	rs11227260	*KAT5*	42,911	0.025	65,504,069	A	0.600	−0.0027	0.0005	4.33E-09	−0.0024	0.0005	2.86E-07	7.54E-07	−0.0020	0.0007	4.50E-03
	15	**rs4775830**	rs1153855	*GATM*	−127,414	0.167	45,533,344	A	0.430	−0.0017	0.0004	1.14E-06	0.0019	0.0003	4.65E-09	2.49E-07	0.0018	0.0004	2.10E-06
	20	rs75041355	rs6127099	*CYP24A1*	6,360	0.056	52,737,762	A	0.034	0.0083	0.0011	3.66E-14	0.0059	0.0011	4.87E-08	1.36E-06	0.0031	0.0012	8.90E-03

(Continued on following page)

TABLE 1 (*Continued*) Summary of novel independent eGFR-associated signals identified by quasi-adaptive method and replication results.

—	Chr	Signal	Index	Closest gene	D [bp]	r^2	Pos (b37)	EA	EAF	GWAS-MA			GCTA			α-	Replication-UKBB		
										Effect	Se	P	Effect	Se	P	threshold	Effect	Se	P
Signal of 4th	2	**rs2075251**	rs35472707	*LRP2*	15,877	0.016	170,011,458	A	0.750	−0.0028	0.0004	8.41E-12	−0.0021	0.0004	1.12E-07	1.76E-06	−0.0031	0.0004	2.80E-15
	6	**rs6912283**	rs881858	*LINC01512*	−442,115	0.000	43,364,494	A	0.560	−0.0009	0.0004	1.34E-02	−0.0018	0.0003	8.72E-08	5.50E-07	−0.0018	0.0004	4.20E-07
	9	**rs4745268**	rs2039424	*PIP5K1B*	−26,649	0.107	71,405,525	T	0.310	0.0035	0.0004	4.45E-19	0.0022	0.0004	1.00E-08	1.82E-06	0.0012	0.0004	6.00E-03
	11	**rs1056819**	rs233438	*KCNQ1*	155,469	0.002	2,949,861	T	0.200	−0.0021	0.0004	2.10E-06	−0.0023	0.0004	1.65E-07	5.51E-07	−0.0022	0.0004	9.50E-07
	20	**rs2585441**	rs6127099	*CYP24A1*	6,553	0.056	52,737,955	C	0.190	−0.0039	0.0005	6.57E-14	−0.0025	0.0005	1.45E-06	3.18E-06	−0.0026	0.0004	1.00E-08
	20	**rs6062357**	rs2261092	*ZGPAT*	538,806	0.001	62,892,739	T	0.450	0.0020	0.0004	2.19E-08	0.0019	0.0004	8.24E-08	5.50E-07	0.0013	0.0003	2.50E-04
Signal of 5th	6	**rs76426793**	rs12207180	*SLC22A2*	−66,715	0.005	160,566,392	A	0.120	−0.0022	0.0005	4.85E-05	−0.0023	0.0005	1.29E-06	1.73E-06	−0.0049	0.0006	2.50E-16
	7	**rs2695565**	rs2365286	*LINC01006*	139,133	0.000	156,397,312	A	0.200	0.0022	0.0004	6.54E-07	0.0023	0.0004	5.27E-07	9.67E-07	0.0027	0.0004	7.00E-10
	15	rs4886425	rs10851885	*NRG4*	−2,179,960	0.001	74,124,543	A	0.170	−0.0027	0.0005	4.26E-09	−0.0023	0.0005	4.58E-07	5.43E-07	−0.0003	0.0004	4.50E-01
Signal of 6th	7	**rs6951593**	rs2365286	*LINC01006*	12,546	0.110	156,270,725	A	0.047	0.0060	0.0010	4.65E-10	0.0042	0.0009	3.90E-06	7.69E-06	0.0038	0.0009	1.30E-05

This table contains the list of novel independent eGFR-associated signals identified by the quasi-adaptive method and replication results. Chr: chromosome; Signal: novel independent signal identified by quasi-adaptive method; Index: known index SNP in the corresponding locus has previously been reported in GWAS of eGFR; Closest gene: the closest gene to index SNP; D[bp]: the distance between index SNP and signal; r^2: pairwise LD correlation between index SNP and signal using UKBB reference sample; Pos: position of signal; EA: effect allele of signal; EAF: frequency of the effect allele of signal; GWAS-MA: European-ancestry-specific GWAS meta-analysis; GCTA: approximate conditional analyses implemented in GCTA; Effect: effect of signal; Se: standard error of signal; P: *p*-value of signal; α-threshold: SNP-specific α-threshold assigned by quasi-adaptive method to a signal; Replication-UKBB: replication analysis by BOLT-linear mixed model in United Kingdom Biobank data set; Bold font indicates replicated independent signals.

Replication of novel multiple independent signals in European-ancestry individuals

To assess the validity of our newly identified independent signals, we conducted conditional eGFR-association analyses using individual-level genotype data among 408,608 European-ancestry participants of the UKBB as independent replication (Methods). For 27 novel independent signals, we conducted 27 conditional analyses (Supplementary Table S2). In total, replication was achieved for 19 signals (Five secondary signals, five tertiary signals, six signals of fourth, two signals of fifth, and one signal of sixth) after applying multiple testing corrections (Methods, Table 1 and Figure 1A). Of note, seven of these signals achieved genome-wide significant conditional *p*-values, and additional four signals were nominally significant ($p < 0.05$) in the replication analysis. Effect estimates for the replicated signals showed a strong correlation ($r^2 = 0.937$) with the discovery results (Figure 1B).

For better comparison, the regional association plots were generated for the unconditional associations and the conditional associations with the highlighted known index and the novel independent signal separately (Supplementary Figures S1-S57). Of note, the new independent signals rs3904600, rs13227214, rs81205, rs2075251, rs2695565, and rs6951593 (identified by the quasi-adaptive method based on the meta-analysis of the previous GWAS of eGFR (Wuttke et al., 2019)) showed smaller *p*-values in their unconditional analysis within the UKBB compared to their corresponding index SNP (Supplementary Figures S4, S19, S22, S31, S52, S55).

Colocalization with gene expression

Colocalization analyses were performed with eQTLs in *cis* across 51 tissues, including kidney cortex, glomerular, and tubulointerstitial for the 17 known eGFR-associated index SNPs as well as for the 19 new independent signals using unconditional and conditional eGFR results (Methods and Supplementary Table S2).

Using unconditional eGFR associations, we identified 56 genes mapping to 13 out of 17 index SNPs for which *cis*-eQTL in at least one tissue colocalized with an eGFR-associated signal with a high PP (≥80%) (Supplementary Table S8 and Supplementary Figure S58). Results for the 19 new independent signals using unconditional GWAS associations revealed significant colocalization in at least one tissue for 42 genes mapping to 11 of the 19 independent signals (Supplementary Table S8 and Figure 2A).

To determine more robust evidence of colocalization, we re-run the colocalization for each known index SNP using the corresponding conditional eGFR association. We identified 53 genes mapping to 11 index SNPs for which *cis*-eQTL in at least one tissue colocalized with an eGFR-associated signal with a high PP (Supplementary Table S9 and Supplementary Figure S59). We identified 10 genes that colocalized with four index SNPs exclusively using conditional associations, which would have remained undetected if only colocalization of unconditional associations had been considered (Table 2). Comparing colocalization for index SNPs based on unconditional with conditional associations across all tissues revealed consistent results for 45 genes mapping to eight index SNPs (Supplementary Table S10), which means that multiple independent signals did not affect the colocalization analyses at these loci. On the other hand, 11 genes mapping to six index SNPs were detected only by colocalization using unconditional association, indicating that multiple independent signals at these loci affected the colocalization analyses for the corresponding index SNPs (Supplementary Table S11).

Colocalization for each new independent signal using conditional association analysis mapped 12 genes to eight of the 19 independent signals with colocalization PP ≥ 80% in at least one tissue (Supplementary Table S9 and Figure 2B). We identified eight genes mapping to 4 novel independent signals with consistent results between colocalization based on unconditional and conditional associations, indicating accurate colocalization results for novel independent signals at these loci (Supplementary Table S10). In addition, five genes mapping to 5 novel independent signals were identified exclusively by colocalization using conditional associations, which would have remained undetected if only colocalization using unconditional associations had been considered (Table 2 and Figure 2B). On the other hand, 34 genes mapping to 9 novel independent signals were detected only by colocalization using unconditional associations, indicating that colocalization using unconditional association has less power to detect accurate results at these loci (Supplementary Table S11).

The complete comparison of the colocalization results for known index SNPs and novel independent signals using conditional *versus* unconditional associations are provided in Supplementary Figures S60-S76.

Discussion

Application of our recently developed quasi-adaptive method to the publicly available GWAS meta-analysis results of eGFR among 783,978 European-ancestry individuals (Wuttke et al., 2019) and subsequent replication in additional 408,608 individuals from UKBB identified 19 novel independent eGFR association signals. These signals included five secondary signals, five tertiary signals, six signals of fourth, two signals of fifth, and one signal of sixth. These results would have gone undetected by conditional analysis applying the commonly used but too conservative genome-wide significance level of 5×10^{-8}. Of note, the individuals included in the LD reference sample were also part of the replication stage,

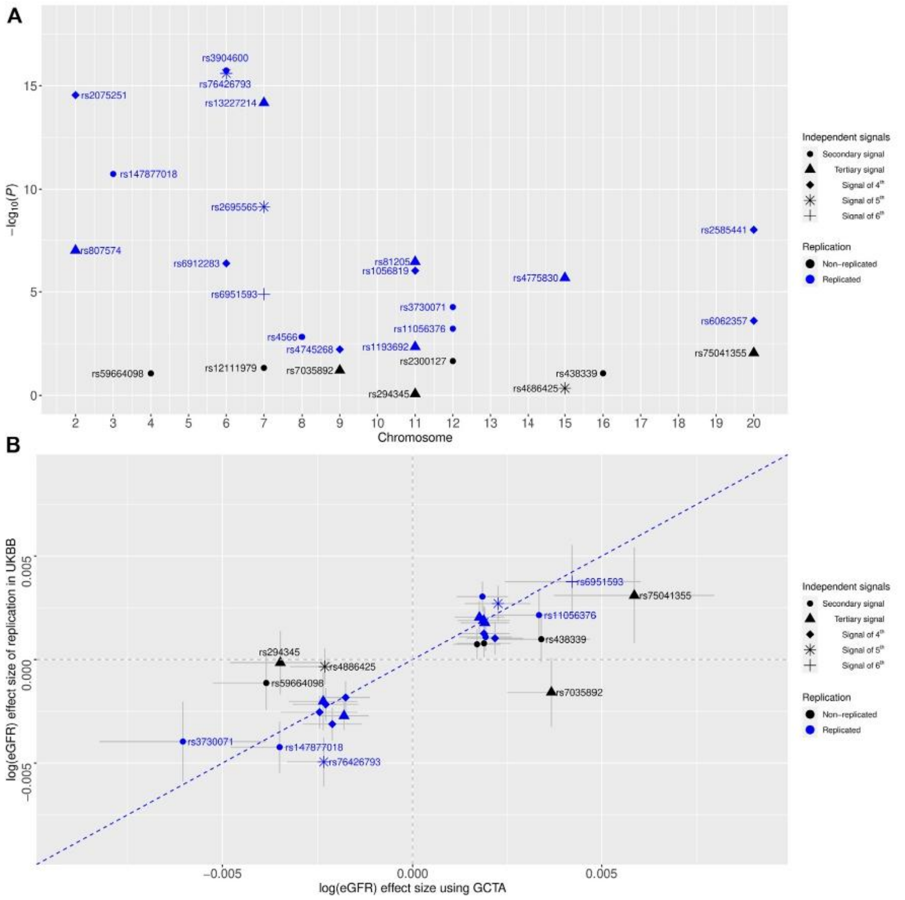


FIGURE 1
(A) Replication of eGFR-associated multiple independent signals identified by the quasi-adaptive method using the United Kingdom Biobank (UKBB) genotype data among European-ancestry individuals. The *x*-axis shows the chromosome number, and the *y*-axis is the $-log_{10}(P)$ of the conditional GWAS of eGFR. Color coding reflects evidence of replication, which is coded as replicated (blue) and non-replicated (black). Different shapes showed multiple independent signals. **(B)**, comparing genetic effect estimates between conditional analysis using GCTA on the GWAS meta-analysis of a previous GWAS of eGFR (*x*-axis) and by conditional GWAS of eGFR on UKBB (*y*-axis). Color coding reflects replication evidence, coded as significant (blue) and non-significant (black). Error bars correspond to 95% confidence intervals. Pearson's correlation coefficient $r^2 = 0.937$ (95% CI = 0.84, 0.98) for the replicated signals. The blue dashed line corresponds to the diagonal line.

but an influence of the results is very unlikely because of the substantially larger sample size in the replication analysis, and the different methods applied (summary statistics with LD reference vs individual level conditional analysis).

Some previous reports on eGFR support our findings. For instance, our secondary signal rs147877018 was previously discovered as an eGFR-associated signal through conditional analysis implemented in GCTA (at locus-wide significance, $p < 10^{-5}$)[20]. In addition, Wuttke et al. (2019) (Wuttke et al., 2019) reported *ADCY6* as a novel eGFR candidate gene in humans by performing a nested candidate gene analysis in mice. *ADCY6* has not been reported to contain genome-wide significant eGFR-associated SNPs or to be located near known loci. However, in our study, the secondary signal rs3730071 was discovered near *ADCY6* (Supplementary Figure S13).

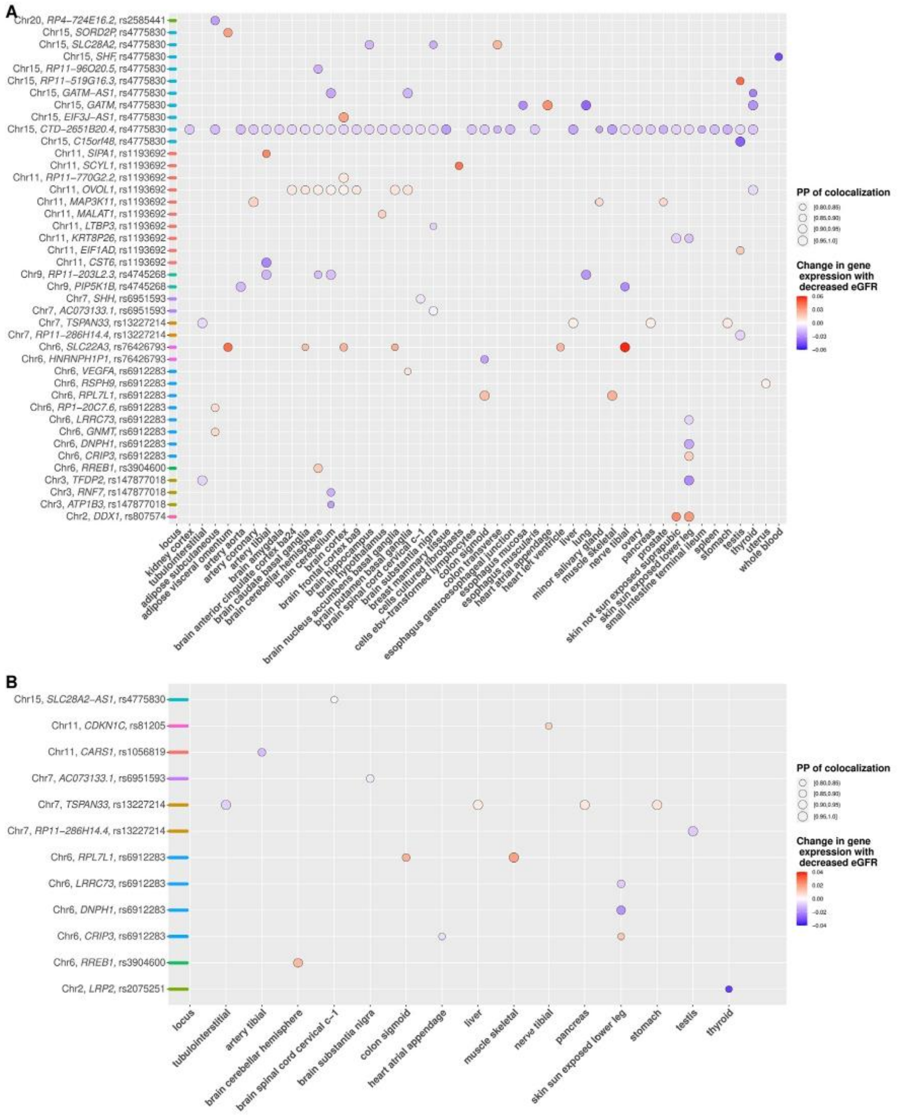


FIGURE 2

(A,B) Colocalization of eGFR association of novel independent signals with gene expression (cis eQTLs) across tissues. **(A** and **B)** depict colocalization results based on unconditional and conditional eGFR association analyses, respectively. Gene with at least one posterior probability of colocalization (PP ≥ 80%) across tissues (*x*-axis) is shown with the respective underlying variant and chromosome number (*y*-axis). Colocalizations are illustrated as dots, where dot size corresponds to the PP and are colored according to the predicted change in gene expression relative to the lower eGFR. Color coding on the *y*-axis reflects the locus.

Colocalization of eGFR-associated known index SNPs and novel independent signals and gene expression implicate specific potential functional genes for follow-up. We investigated the kidney by using *cis*-eQTL dataset from the publicly available GTEx project (Aguet et al., 2019). However, the human kidney tissues have been poorly covered by the GTEx study, and only the

TABLE 2 Summary of colocalization of eGFR association known index SNPs and novel independent signals with posterior probability (PP ≥ 80%). (A-B) contain summary of colocalization of eGFR association known index SNPs and novel independent signals with a high posterior probability of colocalization (PP) ≥ 80% in at least one tissue.

Rsid	Known	Chr	Gene	Tissue	Supplementary Figure
A: New colocalizations in kidney tissues with consistent results between conditional and unconditional association analyses					
rs1397764	Yes	3	TFDP2	tubulointerstitial	Supplementary Figure S62A
rs13227214	No	7	TSPAN33	tubulointerstitial	Supplementary Figure S76B
rs1153855	Yes	15	CTD-2651B20.4	kidney cortex	Supplementary Figure S74A
B: Summary of colocalization results identified exclusively by colocalization based on conditional association analyses, across all tissues					
rs35472707	Yes	2	KLHL41	brain spinal cord cervical c-1	Supplementary Figure S60A
rs2075251	No	2	LRP2	thyroid	Supplementary Figure S60B
rs12207180	Yes	6	RP11–288H12.3	small intestine terminal ileum	Supplementary Figure S63A
rs12207180	Yes	6	SLC22A2	esophagus gastroesophageal junction	Supplementary Figure S63A
rs12207180	Yes	6	SLC22A2	esophagus muscularis	Supplementary Figure S63A
rs12207180	Yes	6	SLC22A2	prostate	Supplementary Figure S63A
rs12207180	yes	6	SLC22A2	testis	Supplementary Figure S63A
rs12207180	yes	6	SLC22A3	artery tibial	Supplementary Figure S63A
rs6912283	no	6	CRIP3	heart atrial appendage	Supplementary Figure S65B
rs10086569	yes	8	RMDN1	adrenal gland	Supplementary Figure S68A
rs10086569	yes	8	WWP1	muscle skeletal	Supplementary Figure S68A
rs1056819	no	11	CARS1	artery tibial	Supplementary Figure S71B
rs81205	no	11	CDKN1C	nerve tibial	Supplementary Figure S72B
rs4775830	no	15	SLC28A2-AS1	brain spinal cord cervical c-1	Supplementary Figure S74B
rs2261092	yes	20	EEF1A2	whole blood	Supplementary Figure S75A
rs2261092	yes	20	MYT1	brain substantia nigra	Supplementary Figure S75A
rs2261092	yes	20	SLC17A9	brain substantia nigra	Supplementary Figure S75A
rs2261092	yes	20	ZGPAT	ovary	Supplementary Figure S75A

Rsid: SNP rsid; Known: SNP was reported as an index SNP in the previous report of eGFR from Wuttke et al. (2019) is labeled as "yes", and novel independent signals identified by quasi-adaptive method are labeled as "no"; Chr: chromosome; Supplementary Figure: comparison of the colocalization results for known index SNPs and novel independent signals using conditional *versus* unconditional associations.

kidney cortex with small sample size is included in this dataset. To overcome this limitation, we also investigated kidney tissue by using a *cis*-eQTL dataset from microdissected human glomerular and tubulointerstitial kidney portions from 187 individuals from the NEPTUNE study (Gillies et al., 2018).

The presence of multiple independent GWAS signals at a locus violates the assumption required by the applied colocalization method (one causal variant for each locus) and likely reduces the power to detect accurate colocalization results. In this context, Wu. et al. (2019) (Wu et al., 2019) showed that for a locus with multiple GWAS signals and/or multiple eQTL signals for the same gene, integration of conditional GWAS association and conditional eQTL led to more robust evidence of colocalization. Our project provides conditional eGFR association tests conducted in the UKBB individual-level genotype dataset. These tests were used to improve the colocalization analyses of the known index SNPs and novel independent signals to identify plausible effector genes related to eGFR. Our findings could be improved by adding the conditional eQTLs data, which may have affected our ability to colocalize signals. It is worth noting that the conditional eQTLs data are not available in our study.

The consistent results between colocalization using unconditional and conditional associations at a locus with multiple independent signals confirm that the colocalization based on unconditional association has enough power to detect accurate colocalization. On the other hand, inconsistent results indicate that colocalization based on unconditional association is affected by the presence of other independent signals at a locus and has less power to detect true colocalization. Therefore, we suggest more accurate results based on colocalization analyses using conditional association and eQTLs, revealing the plausible candidate genes after eliminating the potential effect of other multiple signals.

For instance, in tubulointerstitial and kidney cortex we revealed the known index SNPs rs1397764 and rs1153855 as the shared underlying variants for colocalization of lower eGFR with lower expression of *TFDP2* and *CTD–2651B20.4*, respectively. This was identified by colocalization based on both unconditional and conditional association analyses (Table 2A and Supplementary Figures S58, S59). Across other tissues, we suggest *SLC22A2* as a plausible candidate gene colocalized with index SNP rs12207180, which was detected only after eliminating the effect of other multiple signals at the locus (Table 2B and Supplementary Figure S59). *TFDP2*, *CTD−2651B20.4,* and *SLC22A2* were exclusively identified by our colocalization and have not been reported in the previous report of eGFR (Wuttke et al., 2019). *TFDP2* encodes E2F dimerization partner (DP)-2, which forms heterodimers with the E2F transcription factors resulting in transcriptional activation of cell cycle-regulated genes. Although the role of *TFDP2* in the context of renal disease has not been reported, several genetic associations in or near *TFDP2* have been reported in previous GWAS of eGFR and CKD (Kottgen et al., 2010; Pattaro et al., 2016; Hellwege et al., 2019; Morris et al., 2019; Wuttke et al., 2019). In addition, *TFDP2* was identified as a prioritized gene for eGFR by performing a transcriptome-wide association study (TWAS) and a summary Mendelian randomization test (Doke et al., 2021). Furthermore, the expression of *TFDP2* was associated with the eGFR index variant, specifically in kidney-specific eQTL associations (Graham et al., 2019). *CTD−2651B20.4* is a protein-kinase, interferon-inducible double-stranded RNA-dependent inhibitor, and repressor of (P58 repressor) (PRKRIR) pseudogene with Ensembl version identifier ENSG00000259433.2. There is no explicit function for *CTD−2651B20.4*, and it has not been reported to contain or be located near associated variants with phenotypes, diseases, and traits in humans or other species. *SLC22A2* is specifically expressed in the kidney and plays a critical role in the renal secretion of various cationic compounds (Aoki et al., 2008). *SLC22A2* encodes the polyspecific organic cation transporter (OCT2) and mediates tubular uptake of organic compounds including creatinine in the basolateral membrane of renal tubular epithelial cells (Urakami et al., 2004). *SLC22A2* has been reported to contain or to be located near genetic associations in multiple GWAS of eGFR and CKD (Kottgen et al., 2010; Mahajan et al., 2016; Morris et al., 2019; Wuttke et al., 2019).

Our colocalization of novel independent signals suggests rs13227214 as the shared underlying variant for colocalization of lower eGFR with lower expression of *TSPAN33*in tubulointerstitial tissue, which was robustly identified based on both unconditional and conditional association analyses (Table 2A and Figure 2). Furthermore, in thyroid and nerve tibial tissue, we suggest *LRP2* and *CDKN1C* as the plausible candidate genes colocalized with rs2075251 and rs81205, respectively, which were detected only by colocalization based on conditional associations (Table 2B and Figure 2B). *TSPAN33*, *LRP2*, and *CDKN1C* were identified exclusively by our colocalization of novel independent signals and would have remained undetected if only colocalization of the corresponding index SNPs rs3757387, rs35472707, and rs233438 were considered at these loci (Supplementary Figure S67, Supplementary Figure S60, and Supplementary Figure S72). *TSPAN33* is a member of the tetraspanin family and encodes a transmembrane protein. *TSPAN33* is highly expressed in the kidney and *TSPAN33* mRNA is detectable in the kidney by both microarray and qPCR (Luu et al., 2013). Furthermore, in colocalization analysis of kidney-specific eQTL association (kidney cortex (Ko et al., 2017), glomerulus, and tubule-interstitial compartments (Gillies et al., 2018), *TPSAN33* showed significant colocalization with the eGFR association (Graham et al., 2019). *LRP2* encodes the megalin receptor (Nielsen and Christensen, 2010) and connected to its seed gene *DAB2*, through protein–protein interaction (Hosaka et al., 2009). Chasman et al. (2012) identified *LRP2* related to the kidney function through connection with the previously known eGFR gene *DAB2* and prior biological knowledge about megalin system in kidney function (Chasman et al., 2012). *CDKN1C* expressed in the heart, brain, lung, skeletal muscle, kidney, pancreas and testis. Up-regulation of miR-199a-5p through suppressing *CDKN1C* might promote cell proliferation in autosomal dominant polycystic kidney disease tissues (Sun et al., 2015), which is a genetic disorder characterized by the growth of numerous cysts in the kidney often causes renal failure with many serious complications.

In summary, we have extended our quasi-adaptive method toward identifying multiple independent SNPs within a locus, applied this method to an eGFR meta-analysis result, and discovered and replicated novel eGFR-associated SNPs. Using these results, we revealed plausible candidate genes for eGFR by colocalization, partly undetected using standard approaches. These findings will help improve the understanding of biological mechanisms underlying kidney function and may subsequently help reducing the burden of CKD.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article and Supplementary Material.

Ethics statement

The studies involving human participants were reviewed and approved by the ethics committee of the respective studies provided the summary statistics included in this project. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

SG, TB, and AT contributed to conception and design of the study. SG performed the statistical analysis. AT supervised the project. AT and HG acquired funding for the analyses. SG wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Acknowledgments

We acknowledge support for the Article Processing Charge by the German Research Foundation and the Open Access Publication Fund of the University of Greifswald. This research has been conducted using the United Kingdom Biobank resource under application number 20272.

Conflict of interest

HG has received travel grants and speakers honoraria from Fresenius Medical Care, Neuraxpharm, Servier and Janssen Cilag as well as research funding from Fresenius Medical Care not related to the current project.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.997302/full#supplementary-material

References

Aguet, F., et al. (2019). *The GTEx Consortium atlas of genetic regulatory effects across human tissues*. bioRxiv, 787903. doi:10.1101/787903

Aoki, M., Terada, T., Kajiwara, M., Ogasawara, K., Ikai, I., Ogawa, O., et al. (2008). Kidney-specific expression of human organic cation transporter 2 (OCT2/SLC22A2) is regulated by DNA methylation. *Am. J. Physiol. Ren. Physiol.* 295 (1), F165–F170. doi:10.1152/ajprenal.90257.2008

Astor, B. C., Matsushita, K., Gansevoort, R. T., van der Velde, M., Woodward, M., Levey, A. S., et al. (2011). Lower estimated glomerular filtration rate and higher albuminuria are associated with mortality and end-stage renal disease. A collaborative meta-analysis of kidney disease population cohorts. *Kidney Int.* 79, 1331–1340. doi:10.1038/ki.2010.550

Bello, A. K., Hemmelgarn, B., Lloyd, A., James, M. T., Manns, B. J., Klarenbach, S., et al. (2011). Associations among estimated glomerular filtration rate, proteinuria, and adverse cardiovascular outcomes. *Clin. J. Am. Soc. Nephrol.* 6 (6), 1418–1426. doi:10.2215/CJN.09741110

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., et al. (2018). The UK biobank resource with deep phenotyping and genomic data. *Nature* 562 (7726), 203–209. doi:10.1038/s41586-018-0579-z

Chasman, D. I., Fuchsberger, C., Pattaro, C., Teumer, A., Boger, C. A., Endlich, K., et al. (2012). Integration of genome-wide association studies with biological knowledge identifies six novel genes related to kidney function. *Hum. Mol. Genet.* 21 (24), 5329–5343. doi:10.1093/hmg/dds369

Chronic Kidney Disease PrognosisMatsushitavan der Velde, M., Astor, B. C., Woodward, M., Levey, A. S., et al. (2010). Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: A collaborative meta-analysis. *Lancet* 375, 2073–2081. doi:10.1016/S0140-6736(10)60674-5

Doke, T., Huang, S., Qiu, C., Liu, H., Guan, Y., Hu, H., et al. (2021). Transcriptome-wide association analysis identifies DACH1 as a kidney disease risk gene that contributes to fibrosis. *J. Clin. Investig.* 131 (10), e141801. doi:10.1172/JCI141801

Gansevoort, R. T., Correa-Rotter, R., Hemmelgarn, B. R., Jafar, T. H., Heerspink, H. J. L., Mann, J. F., et al. (2013). Chronic kidney disease and cardiovascular risk: Epidemiology, mechanisms, and prevention. *Lancet* 382 (9889), 339–352. doi:10.1016/S0140-6736(13)60595-4

Gansevoort, R. T., Matsushita, K., van der Velde, M., Astor, B. C., Woodward, M., Levey, A. S., et al. (2011). Lower estimated GFR and higher albuminuria are associated with adverse kidney outcomes. A collaborative meta-analysis of general and high-risk population cohorts. *Kidney Int.* 80, 93–104. doi:10.1038/ki.2010.531

Ghasemi, S., Teumer, A., Wuttke, M., and Becker, T. (2021). Assessment of significance of conditionally independent GWAS signals. *Bioinformatics* 37 (20), 3521–3529. doi:10.1093/bioinformatics/btab332

Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., et al. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10, e1004383. doi:10.1371/journal.pgen.1004383

Gillies, C. E., Putler, R., Menon, R., Otto, E., Yasutake, K., Nair, V., et al. (2018). An eQTL landscape of kidney tissue in human nephrotic syndrome. *Am. J. Hum. Genet.* 103, 232–244. doi:10.1016/j.ajhg.2018.07.004

Go, A. S., Chertow, G. M., Fan, D., McCulloch, C. E., and Hsu, C. y. (2004). Chronic kidney disease and the risks of death, cardiovascular events, and hospitalization. *N. Engl. J. Med.* 351 (13), 1296–1305. doi:10.1056/NEJMoa041031

Gorski, M., Most, P. J. v. d., Teumer, A., Chu, A. Y., Li, M., Mijatovic, V., et al. (2017). Corrigendum: 1000 genomes-based meta-analysis identifies 10 novel loci for kidney function. *Sci. Rep.* 7, 46835. doi:10.1038/srep46835

Graham, S. E., Nielsen, J. B., Zawistowski, M., Zhou, W., Fritsche, L. G., Gabrielsen, M. E., et al. (2019). Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis. *Nat. Commun.* 10, 1847. doi:10.1038/s41467-019-09861-z

Hellwege, J. N., Velez Edwards, D. R., Giri, A., Qiu, C., Park, J., Torstenson, E. S., et al. (2019). Mapping eGFR loci to the renal transcriptome and phenome in the VA Million Veteran Program. *Nat. Commun.* 10, 3842. doi:10.1038/s41467-019-11704-w

Hemmelgarn, B. R., Manns, B. J., Lloyd, A., James, M. T., Klarenbach, S., Quinn, R. R., et al. (2010). Relation between kidney function, proteinuria, and adverse outcomes. *JAMA* 303, 423–429. doi:10.1001/jama.2010.39

Herold, C., Steffens, M., Brockschmidt, F. F., Baur, M. P., and Becker, T. (2009). Intersnp: Genome-wide interaction analysis guided by a priori information. *Bioinformatics* 25, 3275–3281. doi:10.1093/bioinformatics/btp596

Hishida, A., Nakatochi, M., Akiyama, M., Kamatani, Y., Nishiyama, T., Ito, H., et al. (2018). Genome-wide association study of renal function traits: Results from the Japan multi-institutional collaborative cohort study. *Am. J. Nephrol.* 47, 304–316. doi:10.1159/000488946

Hosaka, K., Takeda, T., Iino, N., Hosojima, M., Sato, H., Kaseda, R., et al. (2009). Megalin and nonmuscle myosin heavy chain IIA interact with the adaptor protein disabled-2 in proximal tubule cells. *Kidney Int.* 75, 1308–1315. doi:10.1038/ki.2009.85

Kanai, M., Akiyama, M., Takahashi, A., Matoba, N., Momozawa, Y., Ikeda, M., et al. (2018). Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet.* 50, 390–400. doi:10.1038/s41588-018-0047-6

Kang, G., Ye, K., Liu, N., Allison, D. B., and Gao, G. (2009). Weighted multiple hypothesis testing procedures. *Stat. Appl. Genet. Mol. Biol.* 8, Article23–22. doi:10.2202/1544-6115.1437

Ko, Y. A., Yi, H., Qiu, C., Huang, S., Park, J., Ledo, N., et al. (2017). Genetic-variation-driven gene-expression changes highlight genes with important functions for kidney disease. *Am. J. Hum. Genet.* 100 (6), 940–953. doi:10.1016/j.ajhg.2017.05.004

Kottgen, A., Pattaro, C., Boger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L., et al. (2010). New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* 42, 376–384. doi:10.1038/ng.568

Lee, J., Lee, Y., Park, B., Won, S., Han, J. S., and Heo, N. J. (2018). Genome-wide association analysis identifies multiple loci associated with kidney disease-related traits in Korean populations. *PLoS One* 13 (3), e0194044. doi:10.1371/journal.pone.0194044

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjalmsson, B. J., Finucane, H. K., Salem, R. M., et al. (2005). Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nat. Genet.* 47, 284–290. doi:10.1038/ng.3190

Luu, V. P., Hevezi, P., Vences-Catalan, F., Maravillas-Montero, J. L., White, C. A., Casali, P., et al. (2013). TSPAN33 is a novel marker of activated and malignant B cells. *Clin. Immunol.* 149 (3), 388–399. doi:10.1016/j.clim.2013.08.005

Mahajan, A., Rodan, A. R., Le, T. H., Gaulton, K. J., Haessler, J., Stilp, A. M., et al. (2016). Trans-ethnic fine mapping highlights kidney-function genes linked to salt sensitivity. *Am. J. Hum. Genet.* 99 (3), 636–646. doi:10.1016/j.ajhg.2016.07.012

Matsushita, K., Coresh, J., Sang, Y., Chalmers, J., Fox, C., Guallar, E., et al. (2015). Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: A collaborative meta-analysis of individual participant data. *Lancet. Diabetes Endocrinol.* 3, 514–525. doi:10.1016/S2213-8587(15)00040-6

Morris, A. P., Le, T. H., Wu, H., Akbarov, A., van der Most, P. J., Hemani, G., et al. (2019). Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nat. Commun.* 10, 29. doi:10.1038/s41467-018-07867-7

Nica, A. C., and Dermitzakis, E. T. (2013). Expression quantitative trait loci: Present and future. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368 (1620), 20120362. doi:10.1098/rstb.2012.0362

Nielsen, R., and Christensen, E. I. (2010). Proteinuria and events beyond the slit. *Pediatr. Nephrol.* 25, 813–822. doi:10.1007/s00467-009-1381-9

Okada, Y., Sim, X., Go, M. J., Wu, J. Y., Gu, D., Takeuchi, F., et al. (2012). Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* 44, 904–909. doi:10.1038/ng.2352

Pattaro, C., Kottgen, A., Teumer, A., Garnaas, M., Boger, C. A., Fuchsberger, C., et al. (2012). Genome-wide association and functional follow-up reveals new loci for kidney function. *PLoS Genet.* 8, e1002584. doi:10.1371/journal.pgen.1002584

Pattaro, C., Teumer, A., Gorski, M., Chu, A. Y., Li, M., Mijatovic, V., et al. (2016). Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* 7, 10023. doi:10.1038/ncomms10023

Sun, L., Zhu, J., Wu, M., Sun, H., Zhou, C., Fu, L., et al. (2015). Inhibition of MiR-199a-5p reduced cell proliferation in autosomal dominant polycystic kidney disease through targeting CDKN1C. *Med. Sci. Monit.* 21, 195–200. doi:10.12659/MSM.892141

Teumer, A., Li, Y., Ghasemi, S., Prins, B. P., Wuttke, M., Hermle, T., et al. (2019). Genome-wide association meta-analyses and fine-mapping elucidate pathways influencing albuminuria. *Nat. Commun.* 10, 4130–4219. doi:10.1038/s41467-019-11576-0

Urakami, Y., Kimura, N., Okuda, M., and Inui, K. i. (2004). Creatinine transport by basolateral organic cation transporter hOCT2 in the human kidney. *Pharm. Res.* 21, 976–981. doi:10.1023/b:pham.0000029286.45788.ad

Weiner, D. E., Tighiouart, H., Amin, M. G., Stark, P. C., MacLeod, B., Griffith, J. L., et al. (2014). Chronic kidney disease as a risk factor for cardiovascular disease and all-cause mortality: A pooled analysis of community-based studies. *J. Am. Soc. Nephrol.* 15 (5), 1307–1315. doi:10.1097/01.asn.0000123691.46138.e2

Wu, Y., Broadaway, K. A., Raulerson, C. K., Scott, L. J., Pan, C., Ko, A., et al. (2019). Colocalization of GWAS and eQTL signals at loci with multiple signals identifies additional candidate genes for body fat distribution. *Hum. Mol. Genet.* 28 (24), 4161–4172. doi:10.1093/hmg/ddz263

Wuttke, M., Li, Y., Sieber, K. B., Feitosa, M. F., Gorski, M., et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* 51 (6), 957–972. doi:10.1038/s41588-019-0407-x

Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). Gcta: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* 88 (1), 76–82. doi:10.1016/j.ajhg.2010.11.011

APPENDIX

Eigenständigkeitserklärung

Hiermit erkläre ich, dass diese Arbeit bisher von mir weder an der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Greifswald noch einer anderen wissenschaftlichen Einrichtung zum Zwecke der Promotion eingereicht wurde.
Ferner erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die darin angegebenen Hilfsmittel und Hilfen benutzt und keine Textabschnitte eines Dritten ohne Kennzeichnung übernommen habe.

Sahar Ghasemi

Curriculum vitae

PERSONAL INFORMATION

Name, family:	Sahar, Ghasemi
Birthday:	24.06.1985
Birthplace:	Zanjan, Iran
Nationality:	Iranian
Email:	s.ghasemi2013@gmail.com

EDUCATION

03.2015 – present:	**Ph.D. candidate** Institute for Mathematics and Computer Science University of Greifswald, Greifswald, Germany. Thesis: Statistical methods and applications for biomarker discover using large scale omics data set.
09.2008 – 09.2010:	**M.Sc. in Mathematical Statistics** Faculty of Economics, Allameh Tabatabaei University, Tehran, Iran. Thesis: Admissible and Inadmissible Estimators of Linear Combination of the Mean of Univariate and Multivariate Normal Distribution.
09.2004 – 08.2008:	**B.Sc. in Statistics,** Faculty of Mathematical Sciences, University of Mazandaran, Babolsar, Iran. Thesis: Principles of Total Quality Management (TQM)

LIST OF PUBLICATIONS

List of publications used in the respective thesis

1. Teumer, A., Li, Y., **Ghasemi, S**. *et al*. Genome-wide association meta-analyses and fine mapping elucidate novel pathways influencing albuminuria. *Nature communications*. **2019**;10 (1):1-19.
2. **Ghasemi, S**. *et al*. Assessment of genome-wide significance of conditionally independent signals. *Bioinformatics*. **2021**; 37(20), 3521–3529.
3. **Ghasemi, S**. *et al*. Discovery of novel eGFR-associated multiple independent signals using a quasi-adaptive method. Front Genet. **2022**; 13; 997302.

List of other publications

4. Schlosser, P. *et al* [including **Ghasemi, S**]. Meta-analyses identify DNA methylation associated with kidney function and damage. *Nature communications*. **2021**; 12(1), 7174.
5. Gorski, M. *et al*. [including **Ghasemi, S**]. Meta-analysis uncovers genome-wide significant variants for rapid kidney function decline. *Kidney international*. **2021**; 99(4), 926-939.
6. Jones, G. *et al*. [including **Ghasemi, S**]. Genome-wide meta-analysis of muscle weakness identifies 15 susceptibility loci in older men and women. *Nature communications*. **2021**; 12(1), 654.
7. Portilla-Fernández ,E. *et al*. [including **Ghasemi, S**]. Meta-analysis of epigenome-wide association studies of carotid intima-media thickness. *Eur J Epidemiol*. **2021**; 36, 1143–1155.
8. Tin, A. *et al*. [including **Ghasemi, S**]. Epigenome-wide association study of serum urate reveals insights into urate co-regulation and the SLC2A9 locus. *Nature communications*. **2021**; 12(1), 7173.
9. Shah, S. *et al*. [including **Ghasemi, S**]. Genome-wide association and Mendelian randomisation analysis provide insights into the pathogenesis of heart failure. *Nature communications*. **2020**; 11(1), 163.
10. Wuttke, M. *et al*. [including **Ghasemi, S**]. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics*. **2019**; 51(6),957.
11. Tin, A. *et al*. [including **Ghasemi, S**]. Target genes, variants, tissues and transcriptional pathways influencing human serum urate levels. *Nature genetics*. **2019**; 51(10):1459-1474.

Sahar Ghasemi

Funding

The work conducted in this thesis was funded by project BE 3828/3-3 from the Deutsche Forschungsgemeinschaft (DFG) and the Federal Ministry of Education and Research via the DZHK (grant 81X2400152).

Acknowledgment

I would like to begin by expressing my sincerest gratitude to my supervisor, Professor Volkmar Liebscher, for his unwavering support and guidance throughout my Ph.D. His insights, encouragement, and advice have been invaluable to me.

I am deeply grateful to my supervisor Alexander for providing me with the opportunity to join the group, work on various projects, providing me with dedicated guidance, support, and encouragement throughout my Ph.D. journey. His genius and enthusiasm for science have had a profound impact on me, and I have greatly benefited from his work ethic and generous support, particularly during challenging times. I am truly appreciative of his wonderful and supportive working environment.

I would also like to extend my heartfelt gratitude to my supervisor Tim, who provided me with guidance and support throughout my Ph.D. Despite working remotely, he always promptly responded to my questions and concerns, and his guidance was invaluable.

I would also like to extend my special gratitude to Professor Hans Grabe for his constant support. My sincere thanks to all of my colleagues at the Institute for Community Medicine and the department of psychiatry and psychotherapy, particularly Antoine and Luise, for their kindness, support, and collaboration.

I also want to express my heartfelt gratitude to the international community in Greifswald, particularly Goran, Kasia, and Lorna. We had so much fun together during movie nights, Christmas parties, laser tag, Chinese buffets, and barbecues. I also want to thank my Iranian friends for making me feel at home. The memories I have made in Greifswald, including the Christmas market, Fritz and Kontor, the harbor, Eldena, Rügen, and Szczecin, among others, will be treasured forever. These experiences have made my time in Greifswald an unforgettable one.

Lastly, I would like to express my deepest gratitude to my parents for their unconditional love, trust, and unwavering support throughout my Ph.D. journey. Their support has been my foundation and the driving force behind my success.