Genome-wide association studies and follow-up kernel approaches in the longitudinal PsyCourse Study

Dissertation
zur Erlangung des humanwissenschaftlichen Doktorgrades in der Medizin der Georg-August-Universität Göttingen

vorgelegt von
Bernadette Wendel
aus Kirn

Göttingen, 2022

Supervisor:	Heike Bickeböller Institut für Genetische Epidemiologie Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Second Thesis Committee Member:	Tim Beißbarth Institut für Medizinische Bioinformatik Universitätsmedizin Göttingen Georg-August-Universität Göttingen
Third Thesis Committee Member:	Thomas Kneib Professur für Statistik Wirtschaftswissenschaftliche Fakultät Georg-August-Universität Göttingen
Day of Disputation:	17 February, 2023

Affidavit

I hereby declare that my doctoral thesis entitled "Genome-wide association studies and follow-up kernel approaches in the longitudinal PsyCourse Study" has been written independently with no other sources and aids than quoted.

Bernadette Wendel

Göttingen, December 2022

Acknowledgements

At this point I would like to seize the opportunity to express my heartfelt gratitude to all people who supported me during this journey and whose contributions made this endeavour possible.

I would first like to thank my doctoral thesis supervisor Professor Heike Bickeböller for guidance, valuable feedback and continuous motivation. Prof. Bickeböller enabled opportunities for me to meet and collaborate with scientists from all over of the world gaining an inside into the fascinating world of science, I am very grateful. I would also like to express my gratitude to the members of thesis committee, Professor Thomas Kneib and Professor Tim Beißbarth for their valuable suggestions and feedback during my thesis reports and their kind words of encouragement.

To my former and current colleagues of the department of genetic epidemiology thank you for your help, support and contribution. I especially want to mention Albert Rosenberger whose words of wisdom helped and inspired me during my work. Special thanks to Andrew Entwistle who generously spared his time to proofread all of my drafts. He patiently helped me to improve my English skills. To Katharina Stahl - thank you for shouldering part of my teaching responsibilities granting me time to focus on my thesis. I would also like to thank my collaboration partners based in Munich from the PsyCourse Study. In particular, I wish to express my greatest gratitude to Dr. Urs Heilbronner for his constant support and guidance. To all of my co-authors thank you for your assistance.

To all of my friends as I cannot name you all, thank you for your patience and encouragement in the last years. To my friend Summaira Yasmeen - thank you for making me laugh when I was not feeling like laughing, I miss our brainstorming sessions. To Jenny Lübcke - thank you for welcoming me to Göttingen and for being my friend. To my dear friend Kristina Wicke - thank you for being you! Thank you for listening, and for always being there whenever needed during our time in Greifswald and from the other side of the world.

I want to express my greatest gratitude to my family and my extended family, who encouraged me at every step of the way. First and foremost, I want thank my parents, Heidrun and Werner, for their never-ending faith in me but also for always keeping me grounded and listening to all of my repetitive talks. A very special thank you goes to my elder sister, Caroline, who supported and protected me from day one. To my dear nephew, nieces and godson - thank you for making me smile during stressful times.

Finally, I would like to thank all probands of the PsyCourse Study who participated in the study and gave their valuable time, without their cooperation none of these analyses would have been possible.

Abstract

A genetic association study is a popular method to analyse the connection of genomic factors with disorders or disease-related phenotypes. There are various study types including genome-wide association studies (GWASs) and pathway analyses. In the simplest case, the studied phenotype is either a case-control status or a quantitative trait, e.g., a cognitive test score, with one measurement per individual. This thesis focuses on GWASs and pathway analyses for longitudinal phenotypes in which multiple correlated phenotype measurements per individual are available. The focal phenotypes of this thesis are a group of essential cognitive functions in the longitudinal PsyCourse Study, the executive functions (EFs).

Longitudinal GWAS requires special statistical methods, in which hundreds of thousands of single nucleotide polymorphisms (SNPs) are tested for association with a longitudinal phenotype. Linear mixed models (LMMs) are one popular option to model the correlation structure of the multiple assessments with random effects. LMMs can also handle missing measurements, a frequently occurring problem with longitudinal data. Moreover, LMMs are connected with kernel machine regression (KMR) analyses, which are based on kernel methods. We can apply KMR to perform a pathway analysis, in which a whole pathway (or gene set) is tested for association. These kernel methods can handle high-dimensional genetic data by transforming the data into a lower-dimensional similarity matrix. This similarity matrix or kernel matrix describes the genetic similarities of every pair of study subjects and can be modelled very flexibly. Thus, we are able to integrate additional biological aspects into the kernel, e.g. network information.

This thesis begins with conducting a longitudinal GWAS, in which we aim to identify SNPs influencing the short-term course of EFs. We apply LMMs to study the course over time of EFs. We use data from the PsyCourse Study, in which EFs are assessed at multiple measurement points with cognitive tests, e.g., the Trail Making Test, part B (TMT-B). Nine highly correlated genome-wide significant SNPs are identified as being associated with the change over time in TMT-B. This result is replicated in an independent sample.

The main objective of this thesis is the extension of KMR to long-KMR to enable the performance of a longitudinal pathway analysis. We include additional random effects to KMR to create long-KMR. Long-KMR is further able to integrate network information by utilising a network-based kernel and thus can be applied as a topology-based pathway analysis. Moreover, long-KMR is able to model a pathway as a main genetic and/or genetic-time-interaction effect, either of which can be tested for association. The genetic-time-interaction effect allows studying the association of a pathway with the time course of a phenotype. Overall, long-KMR demonstrates a higher power compared to another longitudinal KMR method previously developed. The power increases further when applying the network kernel to include biological information. Long-KMR is available as an R package *kalpra*.

Contents

1. Introduction

During the last two decades, a large number of genetic association studies have been performed analysing the genetic background of Mendelian disorders and complex diseases. These studies test whether a phenotype of interest, e.g. a case-control status or a normally distributed trait is correlated with a genetic variation [48]. There are various approaches to performing genetic association studies, including the popular genome-wide association study (GWAS) [10, 48, 69, 70] and the newer pathway analysis [30, 38, 72].
In this cumulative thesis, the focus lies on these two types of genetic association study in the context of longitudinal data on unrelated individuals. These comprise multiple measurements for each individual. Here, data from the longitudinal PsyCourse Study [8] are applied to perform longitudinal genetic association studies. The central phenotypes are a group of higher-level cognitive abilities [27], the executive functions (EFs). EFs are essential to accomplishing daily-life tasks by controlling and organising mental processes for both mentally ill and healthy individuals. We study the longitudinal course of these EFs and analyse their genetic background.

1.1. The area of genome-wide association studies

In 2005, the first GWAS was performed with only 96 individuals [32]. Since then thousands more GWASs with larger sample sizes have been and are still conducted. The GWAS Catalog [9] is a publicly available catalogue (`https://www.ebi.ac.uk/gwas/`) containing information and results from 6130 publications (date retrieved: 29 November 2022).

In a GWAS, hundreds of thousands of single genetic markers are tested for association with a phenotype of interest [5, 69], e.g. binary case-control status or a normally distributed phenotype. Hereby, a genetic marker is defined as statistically associated with a phenotype when a specific variant of a genetic marker occurs more often than expected by chance alone in connection with the phenotype [48, 69, 79]. Graphically, the GWAS results can be described as a Manhattan plot (see Fig. 1). Each individual genetic marker is represented as a dot and is displayed according to its genomic location (chromosomes on x-axis) against its $-log_{10}$(p-value). In Fig. 1, the red horizontal line represents the genome-wide significance level 5×10^{-8}. This level is derived from the Bonferroni correction [69]. This multiple testing correction method needs to be performed because of the large number of statistical tests conducted. In the Bonferroni correction, the global significance level α is divided by the number of independent association tests executed

[69]. Thus for a GWAS, we obtain $\frac{\alpha=5\%}{\text{one million}} = 5 \times 10^{-8}$ [79, p.374]. A SNP with a p-value$< 5 \times 10^{-8}$ is denoted as genome-wide significant. In particular, we aim to detect a peak consisting of a number of genome-wide significant genetic markers as demonstrated by Fig. 1 on chromosome 7.

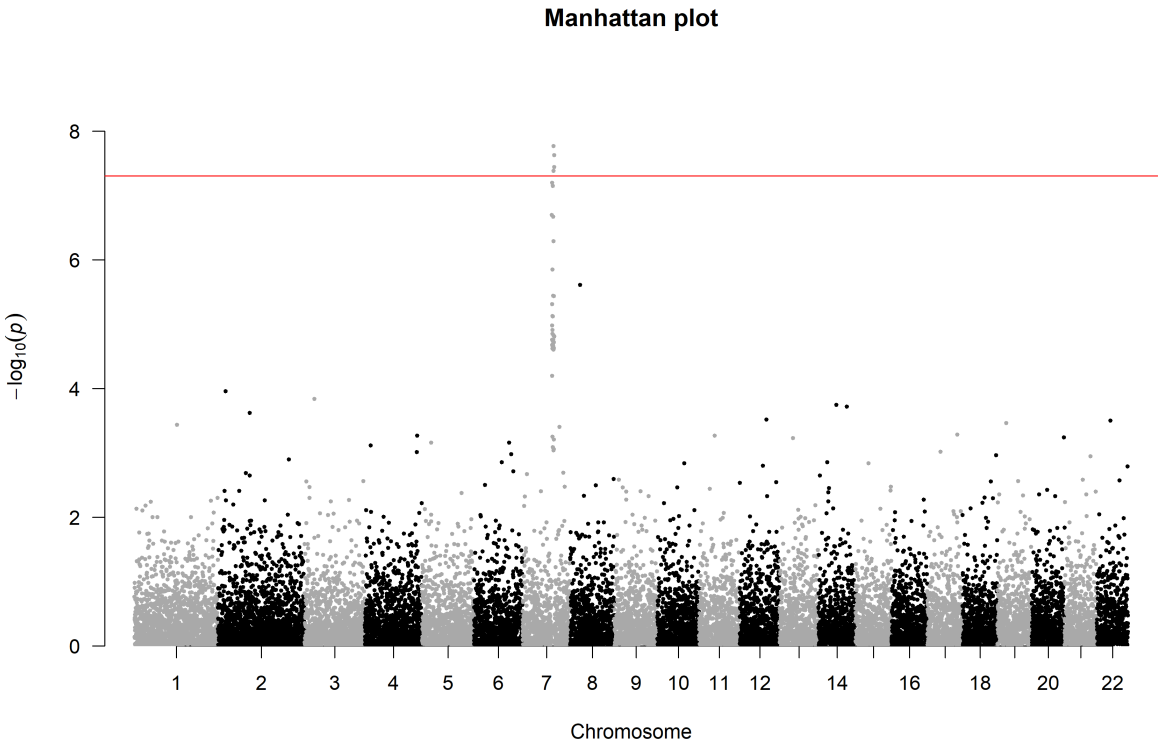


Figure 1. Manhattan plot of simulated results of a genome-wide association study. The red line represents the genome-wide significance level 5×10^{-8}.

As genetic marker type, we focus on single nucleotide polymorphisms (SNPs). A SNP is a single base pair with a known position in the DNA [7]. In humans, a SNP is a biallelic marker as it possesses, in the majority of the cases only two different variants [7]. These variants are termed alleles. The allele that is more frequent in the population is denoted as major allele and the allele with the lower frequency as minor allele. The frequency of a minor allele is referred to as minor allele frequency or MAF. By genotyping each individual, both alleles of the individual SNP are determined. Additionally, we often perform a genotype imputation in a GWAS in which the genotypes of SNPs not directly determined are estimated [10, 54, 69]. This leads to an increased number of available SNPs and to a power boost [54]. A SNP is most commonly coded as 0, 1, or 2, counting the number of minor alleles present. This coding corresponds to the additive genetic model [10, 13] for which it is assumed that the (disease) risk increases in a linear fashion [10]. The additive

genetic model is the standard model in a GWAS, as it possesses acceptable power to detect additive and dominant genetic effects [10, 13].

There are different statistical tests to study an individual SNP for association. In a case-control study with binary phenotype, a classical test of independence, the χ^2-test [10, 13, 79], or a more flexible logistic regression model [10, 13] can be applied. In this thesis, the focus lies on normally distributed phenotypes. Here, we can employ a linear regression model and its extensions, such as a linear mixed model (LMM). In a linear regression model, the SNP is included as independent variable to test the null hypothesis that the mean of the phenotype is equal for all genotype groups [10]. The regression model can also be extended by including additional covariates, for example, gender or age. This expansion allows adjustment for potential influences on the studied phenotype. One of the greater problems in a population-based association study is population stratification. Here, differences in the allele frequencies are based on systematic ancestry distinctions [10, 35, 59], such as population substructures. The most frequent approach to adjust for population stratification is principal component analysis (PCA) [59], in which the principal components (PCs) of the genotype data across the genome are determined. The PCs are the "axes of genetic variation" [79, p.297-298] and a number of top PCs explaining the most variability [59] are included as covariates in the regression model.
The fitted regression model in a GWAS can then be controlled by creating a quantile-quantile (Q-Q) plot. For a Q-Q-plot, the observed test statistics are sorted in ascending order and plotted against the values expected under the null hypothesis [79, p.296]. We can also compute a genomic inflation factor λ, which is the ratio of the empirical median to the median of χ^2_1 (=0.4549) [35, 79]. This so-called genomic control [17] helps identify any inflation of the test statistic, which can be caused by population stratification or confounders for example [35]. A $\lambda > 1$ (in particular $\lambda > 1.2$) [13, 35] indicates an inflation and thus would require model adjustment.

1.2. The world of pathway analysis

In spite of the various findings by GWASs [70], the detected SNPs only explain a small proportion of the genetic variation of complex diseases [51]. This phenomenon is denoted as the problem of missing heritability [51, 53]. One approach to explain the missing heritability is based on the view that complex diseases most commonly do not emerge because of an individual SNP but result from an interplay of various SNPs functionally related to specific biological systems. Thus, we consider pathways or gene sets now. The terms

pathway and gene set are often utilised interchangeably; there are however differences. A gene set is a group of genes defined in respect to common biological features [30], e.g. tissue expression or functionality. A pathway is a group of genes that additionally includes details on how single genes interact with each other [30].
A pathway or gene-set analysis tests a pathway or a gene set for association with a trait of interest [16, 38, 72, 73]. For example, a pathway analysis can be performed as a follow-up study after a GWAS to understand more of the complex underlying biological aspects [16, 43, 73]. An initial approach was GSEA, gene-set enrichment analysis [71]. GSEA utilises as input data GWAS summary statistics containing effect sizes, p-values, and test statistics for the tested SNPs. Since then, many additional pathway analysis tools have been developed: for example, MAGMA [15], VEGAS2Pathway [56], or SKAT [75], and SKAT-O [47]. SKAT and SKAT-O are part of a specific group of pathway analyses applying kernel methods. Kernel methods belong to a class of machine learning algorithms [31], which are often utilised when analysing high-dimensional data such as genomic data. Particularly for high-dimensional genomic data, traditional regression models reach their limits and more sophisticated methods are required [63]. This thesis sets the focus on a kernel-based pathway analysis, kernel machine regression (KMR) analysis.

Mathematically, a pathway can be regarded as a graph G with $G = (V, E)$ in which V is a set of nodes and E a set of edges [19, p.2]. The nodes of V represent the genes of the pathway and the edges outline the gene interactions in which an edge links two nodes when they interact with each other. These connections may be undirected, e.g. "gene A interacts with gene B" or directed, e.g. "gene A activates gene B". The type of information that can be integrated in an analysis depends on the chosen pathway database. A large number of different pathway databases are available: KEGG [41], MSigDB [49], Reactome [40], and Pathway Commons [60], to name but a few. Pathguide [2] provides a total overview of all available databases (`http://www.pathguide.org/`), which holds information on 702 biological pathway-related repositories (date retrieved: 11 December 2022). A number of databases solely contain information on gene sets, for example MSigDB; other databases, such as KEGG or Reactome, also entail information on interactions. For example, Reactome has a pathway browser displaying graphically the connections between the various pathways (`https://reactome.org/PathwayBrowser/`). The focus in this thesis lies on pathways, which allow including gene interaction information in the pathway analysis. Thus, a so-called topology-based pathway analysis is created.

There are several options to perform a pathway analysis, reflected by the huge number

of available pathway analysis tools (for an overview, see de Leeuw et al. [16]). Here, an example process is outlined inspired by the pathway analysis performed in this thesis. A pathway is tested under the self-contained null hypothesis [16, 38, 73], which states that the pathway tested is not associated with the trait of interest [38, 73].
In the performed pathway analysis, raw genotype data in the form of SNPs encoded as $0, 1, 2$ serve as input data. The individual SNPs are assigned to gene(s) of the tested pathway. Most frequently a SNP is assigned to a gene if it lies directly within a gene or in a specified region around the gene [73]. Then, a statistical association test can be performed. Different options are available, for example, the Kolmogorov-Smirnov test [71], linear regression models with an F-test [16], or a variance component test [47, 75] to name a number. In the KMR applied, a variance component test is conducted (refer to chapter 2 for details). In contrast to a GWAS, the multiple testing is largely reduced, but still required, as pathways instead of hundreds of thousands of SNPs are tested. Here, we can utilise either the Bonferroni approach or a less conservative approach, for example Benjamini-Hochberg [4]. The latter method corrects for the false discovery rate (FDR) instead of the family-wise error rate (FWER). But we can also apply a more complex approach [36, 45], which considers potential dependency structures between pathways based on overlapping SNPs or genes.

1.3. Longitudinal data, their properties, and analysis methods

In a longitudinal study, individuals are assessed repeatedly at multiple times, also denoted as measurement points, over a specific period of time [3, 11, 20]. This key characteristic allows direct analysis of the change over time [20] and/or the relationship between risk factors and the development of disorders [11]. We are also able to distinguish between the changes within and between the individuals of the study sample [20]. Various longitudinal studies analyse these aspects [12, 28, 42, 52, 66]. Here, it must be considered that the multiple measurements within the study subjects are correlated [3, 20], preventing the use of common analysis methods. For example, the often applied simple regression model assumes independence between the individual observations [3] and thus cannot be utilised without modifications. Furthermore, longitudinal studies have higher financial and temporal costs [11].

In the special case of only two measurement points per individual in a longitudinal study pre/post-analyses can be applied, including, for example, the change analysis and the ANCOVA model [3]. For the often applied change analysis, the difference between the

first and second measurement point is computed and regressed against covariates utilising a simple regression model. For the ANCOVA model, the second measurement is included as covariate in a simple regression model [3].
For longitudinal data, one can also apply simple approaches in which the measurements are summarised in one single summary measurement. This summary value can be tested with a cross-sectional approach [55], e.g. a linear regression model. Different summary measurements can be utilised, such as the mean or median of the measurements. However, all of these simple approaches do not use the full potential of longitudinal data. More sophisticated methods have also been developed, including the popular linear mixed model (LMM) [58] and generalized estimating equations (GEE) [3]. In GEE, two models are specified: a regression model for the main response and a second model for the within-subject correlation [3]. However, LMMs play a central role in this thesis, owing to their diverse applicability and their connection to other methods (please refer to chapter 2.2 for more details). In a LMM, so-called random effects are added to a simple linear regression model correcting for the dependence structure of the multiple measurements [23, 58]. A LMM contains the basic regression coefficients denoted as fixed effects, which model the population-average effect. The random effects included in the LMM describe the subject-specific effect. For example, a random effect can be a random intercept enabling each individual to have their own starting point. Further, we can integrate a random slope modelling the individual course for each study subject. We assume that these random effects follow approximately a normal distribution with a specific mean and variance. In the following, the notation of a LMM based on [23, 58] is introduced. Thus, let us assume that y_i is the vector containing a quantitative phenotype for an individual i with $i = 1, \ldots, n$ (n sample size) and m measurement points. The LMM is given as follows:

$$y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{1}$$

where β is the vector of the regression coefficients (fixed effects), X_i and Z_i are the design matrices for the fixed and random effects, respectively. The random effect vector is b_i with $b_i \sim \mathcal{N}(0, D_i)$, where the mean is zero and covariance-variance matrix of individual i is D_i. The random error ϵ_i follows a $\mathcal{N}(0, R_i)$ with R_i being the covariance-variance matrix. We assume that ϵ_i and b_i are independent.

A great caveat of longitudinal studies is missing data. Individuals can drop out or miss single measurement points leading to biased results when not handled appropriately [3, 11, 20]. One possible approach is imputation filling in the missing assessments. A

variety of imputation approaches are available [21, 39]; for example, multiple imputation in which we perform the imputation step multiple times obtaining multiple estimates for one missing value. However, some analysis methods, such as LMM can handle incomplete longitudinal data as long as a specific missing mechanism can be assumed [58].
Overall, three underlying missing mechanisms [3, 58] have been distinguished: the Missing Completely at Random (MCAR), the Missing at Random (MAR), and the Missing Not at Random (MNAR) mechanisms. For the first, the data are assumed to be randomly missing independent of any observed or unobserved measurements. Under a MAR mechanism, the data missing depend on other measurements observed in the study and not on the absent assessment itself [3, 58]. When the data are missing not at random, the measurements depend on non-observed assessments and potentially on observed data. For MNAR, the reason for the missing values is unknown and can lead to biased results. LMM assumes that the missing data are either MCAR or MAR [58].

1.4. Outline

This cumulative thesis aims to perform longitudinal genetic association studies on the single marker and pathway level. To correct for the dependence structure of longitudinal data, LMMs and their connections to KMR analysis are employed. The KMR analysis is extended to conduct a longitudinal pathway analysis, which can simultaneously integrate network information. Furthermore, this extended version is capable of modelling and studying both main genetic and genetic-time-interaction effects. The latter are often of interest when studying longitudinal data. The focal phenotypes in this thesis are executive functions. These specific cognitive functions are assessed in the longitudinal PsyCourse Study [8] and are investigated in the analyses. To perform the longitudinal pathway analysis, an R package *kalpra* was developed.

This thesis is structured as follows: Chapter 1 introduces the basis of genetic association studies and longitudinal data with the focus on GWAS and pathway analyses. Chapter 2 presents kernel methods as well as their application as pathway analysis in the form of KMR. Furthermore, the extension of KMR to long-KMR analysing longitudinal data and its variations are explained. Chapter 3 describes the PsyCourse Study, the data implemented in the analyses, and as real-world data example. Chapter 4 provides summaries of the peer-reviewed publications and represents the main body of this thesis given its cumulative nature. Finally, chapter 5 discusses the advantages and disadvantages of the presented method.

2. The diverse facets of kernel methods

This section introduces kernel methods and their application in regression models as kernel machine regression (KMR) analysis. We will explain the connection between KMR and linear mixed models and present the variance component test utilised to test a pathway for association. Next, long-KMR is described, which is the expansion of KMR to analysing longitudinal data. Lastly, variations of long-KMR are specified including the genetic-time-interaction model and the kernel principal component analysis (KPCA).

2.1. Kernel methods and kernel machine regression

Kernel methods are machine learning algorithms and are used in a variety of research areas and analysis problems, such as clustering, correlation or dimension reduction [31]. The basic idea of kernel methods states that relationships between data points are easier modelled if they are transformed from their original representation into a higher dimensional feature space [31]. The data points are transformed via a user-specified feature map [31] converting the data into a similarity or dissimilarity matrix. This matrix contains quantitative values describing the similarity/dissimilarity between individuals and is also denoted as kernel matrix.
When we consider a genetic association study then a large number of SNPs for each individual are specified. These SNPs are transformed into a similarity measure assessing the genetic resemblance between every pair of individuals in the study sample [31, 62, 63]. The similarity matrix or kernel matrix contains the similarity values. This kernel matrix can be analysed in a regression framework creating a kernel machine regression analysis.

A KMR is a semi-parametric regression analysis, in which we regress the phenotype of interest onto the kernel matrix [31]. Subsequently, we assume that the genomic data analysed are given in form of SNPs for which the genotypes for each individual in the study sample are determined. The SNPs are coded as 0,1, or 2, and we have no missing genotypes. This information is presented in form of an $n \times s$ genotype matrix G with n being the sample size and s the number of SNPs available. Let us assume that y_i is a cross-sectional normally distributed phenotype for individual i with $i = 1, \ldots, n$. Then, the basic KMR model is as follows:

$$y_i = x_i^T \beta + h(g_i) + \epsilon_i, \tag{2}$$

where β is a $p \times 1$ vector of regression coefficients, x_i^T is the transposed covariate vector and

g_i is an $s \times 1$ vector containing the genetic information (SNPs) of individual i. The random error is $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$. The covariates are modelled parametrically and g_i is modelled with the unknown non-parametric function $h \in \mathcal{H}_k$ with $\mathcal{H}_k$ being a reproducing kernel Hilbert space (RKHS). A RKHS is a vector space with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k} : \mathcal{H}_k \times \mathcal{H}_k \to \mathbb{R}$, which is complete with respect to the norm induced by the inner product [31]. The computation of h can be very complex and computationally intensive, thus we employ that $h \in \mathcal{H}_k$ and approximate h by a feature map φ. We utilise the basic idea of the kernel method and transform the genotype data to a higher dimensional feature space in which $\{\varphi_k\}$ are basis functions. Thus, we are able to describe $h(g_i)$ as a linear combination of the basis functions:

$$h(g_i) = \sum_{l=1}^{q} \varphi_l(g_i)\omega_l,$$

where $i = 1, \ldots, n$ and $\omega_l \in \mathbb{R}$ with $(l = 1, \ldots, q)$ [31]. This is also denoted as the primal representation. However, according to Mercer's theorem [31] we can define a kernel function k for any pair of individuals i and j as

$$k(g_i, g_j) = \langle \varphi(g_i), \varphi(g_j) \rangle_{\mathcal{H}_k},$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is the inner product of the RKHS. The RKHS is uniquely determined by this positive semi-definite kernel function k and the associated $n \times n$ kernel matrix (Gram matrix). The kernel matrix is defined as $K := \{k(g_i, g_j)\}_{i,j=1}^{n}$ in which the elements of K are the similarity measurements computed with k. By applying the kernel function, any $h(g) \in \mathcal{H}_k$ can be expressed as a linear combination:

$$h(g) = \sum_{i=1}^{n} \alpha_i k(g_i, g),$$

with α_i being constants [50]. This formulation is called the dual representation of function h in kernel methods. As $h \in \mathcal{H}_k$ with $\mathcal{H}_k$ being a RKHS the reproducing property of $\mathcal{H}_k$ ensures the existence of the primal and dual representation [31]. Any kernel function k can be applied as long as the function and the associated kernel matrix K is positive semi-definite, i.e. $c^T K c \geq 0, \forall c \neq 0$ with $c \in \mathbb{R}^n$.

Because of this flexibility of the kernel function a number of different kernels exist [63]. One of the most popular kernels is the linear kernel, which implies an additive independent effect between SNPs. For two individuals i and j whose genotypes are represented by g_i

and g_j, respectively, we obtain the linear kernel as follows:

$$k(g_i, g_j) = g_i^T \cdot g_j. \tag{3}$$

Another often utilised kernel is the quadratic kernel [76], which is a polynomial kernel putting a greater weight on already high entries in the genotype matrix. We compute K for two individuals i and j as follows:

$$k(g_i, g_j) = (g_i^T \cdot g_j + 1)^2. \tag{4}$$

Additionally, we can select a more complex kernel integrating additional information, e.g. network information. A basic way to define new kernel matrices is described by Schaid [63]. The network-based kernel defined by Freytag et al. [24] includes topology information of a pathway in form of an adjacency matrix. When a pathway is portrayed as a graph the adjacency matrix N is a quadratic matrix with $n_{uv} = 1$, if the gene u and v interact with each other or zero otherwise. The network kernel is defined as follows:

$$K = GANA^TG^T, \tag{5}$$

where G is the genotype matrix, N represents the adjacency matrix, and A describes as annotation matrix the assignment of each individual SNP $s \in G$ to a gene of the pathway. The elements of A are either $a_{su} = 1$ if SNP s is mapped to gene u or zero otherwise. The annotation matrix can also be size-adjusted with respect to the different gene sizes by dividing a_{su} by the square root of the total number of SNPs mapped to gene u [24]. The network kernel with the size-adjusted annotation matrix is denoted as size-adjusted network kernel [24].

2.2. Connection between kernel machine regression and linear mixed models

Next, the connection between a KMR and a LMM [31, 50] is described. We start with the estimation of the parameters β and h of model 2, the KMR. Based on Ge et al. [31], we minimize the scaled penalized likelihood function:

$$\mathcal{J}(\beta, h) = \underbrace{\frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta - h(g_i))^2}_{\text{loss function}} + \underbrace{\frac{\lambda}{2}||h||^2_{\mathcal{H}_k}}_{\text{penalty term}}, \tag{6}$$

where $||\cdot||_{\mathcal{H}_k}$ is the norm induced by the inner product on the RKHS with $h \in \mathcal{H}_k$ and λ is a tuning parameter. The tuning parameter controls the balance of the model complexity and the model fit [31, 50]. As we focus on quantitative data, the selected loss function is the squared error loss function and the squared norm is chosen as penalty. The Representer Theorem [31, 50] specifies that the general solution for h is:

$$h(\cdot) = \sum_{j=1}^{n} \alpha_j k(\cdot, g_j), \tag{7}$$

where $\alpha_j \in \mathbb{R}$ [50]. By substituting equation 7 into equation 6, we receive:

$$\begin{aligned}\mathcal{J}(\beta, \alpha) &= \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta - \sum_{j=1}^{n}\alpha_j k(g_i, g_j))^2 + ||\sum_{j=1}^{n}\alpha_j k(\cdot, g_j)||^2_{\mathcal{H}_k} \\ &= \frac{1}{2}(y - X\beta - K\alpha)^T(y - X\beta - K\alpha) + \frac{\lambda}{2}\alpha^T K\alpha\end{aligned}$$

with K being the associated kernel matrix, X the design matrix and y the phenotype vector of model 2 in matrix notation. In order to compute the penalty term, we apply the matrix notation and the reproducing property of the RKHS. We receive the estimates of β and h by setting the derivatives of $J(\beta, \alpha)$ with respect to α and β to zero. We obtain the following estimates according to Ge et al. [31]:

$$\begin{aligned}\hat{\beta} &= [X^T(K + \lambda I)^{-1}X]^{-1}X^T(K + \lambda I)^{-1}y \\ \hat{h} &= K(K + \lambda I)^{-1}(y - X\hat{\beta}),\end{aligned}$$

where I is the identity matrix.

Next, we consider a LMM for a normally distributed phenotype vector y for n individuals in matrix notation:

$$y = X\beta + h + \epsilon,$$

where X is an $n \times p$ covariate matrix, β is a $p \times 1$ fixed effect vector, h is an $n \times 1$ random vector that follows a $\mathcal{N}(0, \tau K)$ in which K is an $n \times n$ kernel matrix. We assume that the random error ϵ follows a $\mathcal{N}(0, \sigma_\epsilon^2 I)$ and that h is independent of ϵ. To connect KMR and LMM, τ can be expressed as a function of the tuning parameter λ and the variance of the residuals σ_ϵ^2 with $\tau = \lambda^{-1}\sigma_\epsilon^2$ [62]. When we look at the marginal distribution of y, which follows a $\mathcal{N}(X\beta, \Sigma = \tau K + \sigma_\epsilon^2 I)$ and apply $\tau = \lambda^{-1}\sigma_\epsilon^2$ to Σ, we get $\Sigma = \tau(\lambda I + K)$. By maximizing the mixed log likelihood function [23, 62] and taking the first derivative,

we receive the best linear unbiased estimator (BLUE) $\hat{\beta}_{LMM}$ and the best linear unbiased predictor (BLUP) $\hat{h}_{LMM}$ for the LMM:

$$\hat{\beta}_{LMM} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$$
$$\hat{h}_{LMM} = K(K+\lambda I)^{-1}(y - X\hat{\beta}).$$

When we consider that $\Sigma = \tau(\lambda I + K)$, we see that the BLUE $\hat{\beta}_{LMM}$ and BLUP $\hat{h}_{LMM}$ obtained for the LMM are equivalent to the estimates $\hat{\beta}$ and $\hat{h}$ of the KMR. This connection enables a common framework with respect to model fitting and statistical inference for LMM and KMR.

2.3. Pathway analysis with a variance component test

To perform an association study, especially to test a pathway for association with a phenotype of interest a statistical test is required. Thus Liu et al. [50] proposed a score test to perform a variance component test. In this pathway analysis, the pathway is given as genotype data G and modelled with the non-parametric function h. To test the pathway for association, we formulate the null hypothesis H_0 as $h(\cdot) = 0$, which is equal to $H_0 : \tau = 0$ owing to the connection of KMR and LMM. Subsequently, we consider model 2 as LMM in matrix notation. Let us assume that y is a cross-sectional normally distributed phenotype vector for n individuals. Then, the LMM is as follows:

$$y = X\beta + h + \epsilon, \tag{8}$$

where β is the $p \times 1$ vector of the regression coefficients, X is the $n \times p$ design matrix, $h \sim \mathcal{N}(0, \tau K)$ with K being the kernel matrix and $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I)$. The marginal distribution of y is $\mathcal{N}(X\beta, \Sigma = \tau K + \sigma_\epsilon^2 I)$, which is used to receive the score test statistic Q_{cross}. To determine the score test statistic for the variance component test, we consider the restricted log likelihood function to avoid biased results. As we will explain the single deduction steps for the longitudinal model the details are skipped here (see Ge et al. [31] and Liu et al. [50] for details). The final score test statistic for the cross-sectional test is:

$$Q_{cross} = \frac{1}{2\sigma_\epsilon^2}(y - X\hat{\beta})^T K (y - X\hat{\beta}), \tag{9}$$

where $\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$ is the maximum likelihood estimator for the fixed regression coefficients [31]. The score statistic Q_{cross} is a quadratic form and follows a mixture of χ^2 distributions, particularly $Q_{cross} = \sum_{l=1}^{L} \lambda_l \chi_l^2$, where λ_l are eigenvalues of

$\frac{1}{2}KP_0$ with $P_0 = I - X(X^TX)^{-1}X^T$. We can choose between different methods to obtain the p-value for the test statistic. Most commonly, either Davies' method [14] or the Satterthwaite approximation [61] are applied. For the latter a moment-matching method is utilised (please refer to Ge et al. [31] and Liu et al. [50] for details). This approximation can lead to an inflated type I error rate, especially for small significance levels whereas Davies' method requires more time to compute.

2.4. Expansion to longitudinal kernel machine regression

After introducing the basic requirements, we describe now the extension of KMR to long-KMR. Long-KMR enables the analysis of longitudinal data. To correct for the correlation structure of the multiple measurements of each individual we include additional random effects in the KMR model on top of the non-parametric function h. As Yan et al. [77], we exploit the estimation equivalence between LMMs and KMR. In the following, we assume for individual i with $i = 1, \ldots, n$ (n sample size) that y_i is a normally distributed phenotype with m measurements (complete data). The long-KMR model for individual i is as follows:

$$y_i = X_i\beta + h(g_i) + Z_ib_i + \epsilon_i, \tag{10}$$

where β is a $p \times 1$ coefficients vector, b_i a $q \times 1$ random effect vector with $b_i \sim \mathcal{N}(0, D_i)$ in which D_i is the covariance-variance matrix. Further, X_i ($m \times p$ matrix) and Z_i ($m \times q$ matrix) are design matrices of the fixed and the random effects, respectively. The genetic vector g_i of individual i is modelled with a non-parametric function h for which we assume that $h \sim \mathcal{N}(0, \tau K)$ with K being the $n \times n$ kernel matrix. The random error ϵ_i follows a normal distribution with mean zero and covariance-variance matrix R_i ($\mathcal{N}(0, R_i)$). It is assumed that b_i and ϵ_i are uncorrelated to avoid model overfitting.

In order to determine the altered score test statistic of long-KMR, we use the LMM form of the long-KMR in matrix notation. For the phenotype vector y the model is:

$$y = X\beta + h + Zb + \epsilon, \tag{11}$$

where $b \sim \mathcal{N}(0, D)$ with $D = diag(D_1, \ldots, D_n)$, $\epsilon \sim \mathcal{N}(0, R)$ with $R = diag(R_1, \ldots, R_n)$ and the design matrices are $X = (X_1, \ldots, X_n)^T$ and $Z = diag(Z_1, \ldots, Z_n)$. We assume that b and ϵ are uncorrelated. The non-parametric function h follows a $\mathcal{N}(0, \tau\mathbf{K})$ for which the kernel matrix $\mathbf{K}$ is now depending on the sample size n and the number of

measurement points m. Under the assumption of an equal number of measurements for each individual (complete data) $\mathbf{K}$ can be obtained by $\mathbf{K} = K \otimes 1_{m\times m}$ in which $\otimes$ represents the Kroenecker product. The $n \times n$ kernel matrix K is only determined once as the SNPs are time invariant. The resulting matrix $\mathbf{K}$ is a $(n \cdot m) \times (n \cdot m)$ matrix.
We follow the steps proposed by Liu et al. [50] to obtain the altered test statistic Q_{long} under the null hypothesis $H_0 : \tau = 0$. As the primary interest is on population-based inference the inference is based on the marginal distribution of y as most commonly done for LMM [58]. The marginal distribution of y is $y \sim \mathcal{N}(X\beta, \Sigma)$ with $\Sigma = Cov(y) = Cov(\epsilon) + Cov(h) + Cov(Zb) = R + \tau\mathbf{K} + ZDZ^T$.

The general form of the score test [31], also denoted as Lagrange multiplier test is:

$$Q = \frac{U(\theta)}{I_E(\theta)^{1/2}}|_{\theta=\theta_0}, \tag{12}$$

where $U(\theta)$ is the first derivative of the log likelihood (= score) with respect to parameter θ and $I_E(\theta)$ is the Fisher information matrix or expected information matrix of θ [31]. We utilise the restricted log likelihood l_{REML} of the marginal model to obtain unbiased results [50] as the main focus lies on the variance parameter. The l_{REML} is as follows:

$$\begin{aligned} l_{REML} &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}(y - X\hat{\beta})^T\Sigma^{-1}(y - X\hat{\beta}) \\ &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}y^TPy, \end{aligned}$$

where $\hat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$ and $P := \Sigma^{-1} - \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$. We apply $P\Sigma P = P$ and $\Sigma^{-1}(y - X\hat{\beta}) = Py$ [31]. Next, we specify the numerator of Q for which l_{REML} is differentiated with respect to τ and receive:

$$\begin{aligned} U(\tau) &= \frac{\partial l_{REML}}{\partial\tau} \\ &= -\frac{1}{2}tr(\Sigma^{-1}\frac{\partial\Sigma}{\partial\tau}) + \frac{1}{2}tr((X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\frac{\partial\Sigma}{\partial\tau}\Sigma^{-1}X) + \frac{1}{2}y^TP\frac{\partial\Sigma}{\partial\tau}Py \\ &= -\frac{1}{2}tr(P\frac{\partial\Sigma}{\partial\tau}) + \frac{1}{2}y^TP\frac{\partial\Sigma}{\partial\tau}Py \\ &= -\frac{1}{2}tr(P\mathbf{K}) + \frac{1}{2}(y - X\hat{\beta})^T\Sigma^{-1}\mathbf{K}\Sigma^{-1}(y - X\hat{\beta}) \end{aligned}$$

where P is defined as above. Here, we first apply multiple differential rules: the chain rule and the differential rules for matrices $d|X| = tr(X^{-1}dX)$ and $dX^{-1} = -X^{-1}(dX)X^{-1}$. Next, we expand and utilise the definition of P. For the final step, we utilise that the first

derivative of Σ with respect to τ is $\frac{\partial \Sigma}{\partial \tau} = \mathbf{K}$ and $\Sigma^{-1}(y - X\hat{\beta}) = Py$. Under $H_0 : \tau = 0$, the final score is received:

$$U(\tau)|_{\tau=0} = -\frac{1}{2}tr(P_0\mathbf{K}) + \frac{1}{2}(y - X\hat{\beta}_0)^T\Sigma_0^{-1}\mathbf{K}\Sigma_0^{-1}(y - X\hat{\beta}_0), \tag{13}$$

where $P_0 = \Sigma_0^{-1} - \Sigma_0^{-1}X(X^T\Sigma_0^{-1}X)^{-1}X^T\Sigma_0^{-1}$ with $\Sigma_0 = R_0 + ZDZ^T$. To determine the Fisher information matrix I_E first the observed information matrix I_O is specified as $I_E = E(I_O)$ [31]. The I_O matrix with respect to τ [31] is:

$$\begin{aligned} I_O(\tau) &= -\frac{\partial l_{REML}}{\partial \tau^2} \\ &= -\frac{1}{2}tr(P\frac{\partial \Sigma}{\partial \tau}P\frac{\partial \Sigma}{\partial \tau}) + \frac{1}{2}tr(P\frac{\partial^2 \Sigma}{\partial \tau^2}) + y^TP\frac{\partial \Sigma}{\partial \tau}P\frac{\partial \Sigma}{\partial \tau}Py - \frac{1}{2}y^TP\frac{\partial^2 \Sigma}{\partial \tau^2}Py, \end{aligned}$$

where P is defined as before. We can now specify I_E and take the derivative with respect to τ:

$$\begin{aligned} I_E(\tau) &= E(-\frac{\partial l_{REML}}{\partial \tau^2}) \\ &= -\frac{1}{2}tr(P\frac{\partial \Sigma}{\partial \tau}P\frac{\partial \Sigma}{\partial \tau}) + \frac{1}{2}tr(P\frac{\partial^2 \Sigma}{\partial \tau^2}) + tr(P\frac{\partial \Sigma}{\partial \tau}P\frac{\partial \Sigma}{\partial \tau}) - \frac{1}{2}tr(P\frac{\partial^2 \Sigma}{\partial \tau^2}) \\ &= \frac{1}{2}tr(P\frac{\partial \Sigma}{\partial \tau}P\frac{\partial \Sigma}{\partial \tau}) = \frac{1}{2}tr(P\mathbf{K}P\mathbf{K}), \end{aligned}$$

with $\frac{\partial \Sigma}{\partial \tau} = \mathbf{K}$ and P as above. When we consider $H_0 : \tau = 0$, the final I_E is obtained:

$$I_E(\tau)|_{\tau=0} = \frac{1}{2}tr(P_0\mathbf{K}P_0\mathbf{K}), \tag{14}$$

where P_0 and Σ_0 are defined as for equation 13. Thus, by substituting equations 13 and 14 into the general form of the score test represented by equation 12, we are able to determine the altered score statistic:

$$Q_{long} = \frac{-\frac{1}{2}tr(P_0\mathbf{K}) + \frac{1}{2}(y - X\hat{\beta}_0)^T\Sigma_0^{-1}\mathbf{K}\Sigma_0^{-1}(y - X\hat{\beta}_0)}{\sqrt{\frac{1}{2}tr(P_0\mathbf{K}P_0\mathbf{K})}}.$$

By replacing Σ_0 with its estimate $\hat{\Sigma}_0 = \hat{R}_0 + Z\hat{D}Z^T$ and because the terms $tr(P_0\mathbf{K})$ and $tr(P_0\mathbf{K}P_0\mathbf{K})$ are independent of y, the final score statistic is:

$$Q_{long} = \frac{1}{2}(y - X\hat{\beta}_0)^T\hat{\Sigma}_0^{-1}\mathbf{K}\hat{\Sigma}_0^{-1}(y - X\hat{\beta}_0), \tag{15}$$

where $\hat{\beta}_0 = (X^T\hat{\Sigma}_0^{-1}X)^{-1}X^T\hat{\Sigma}_0^{-1}y$. The altered test statistic Q_{long} is a quadratic form as its cross-sectional counterpart. We use this characteristic to approximate the distribution applying the theorem in Yuan and Bentler [78]:

Theorem. *Let $x \sim \mathcal{N}(0, \Gamma)$ and $T = x^TWx$ be a quadratic form in x. Γ is typically of full rank while W is non-negative definite. Let the rank of W be L and the nonzero eigenvalues of $W\Gamma$ be $\lambda_1, \ldots, \lambda_L$. There exists $T = x^TWx = \sum_{l=1}^{L} \lambda_l z_l^2$, where $z_l \sim \mathcal{N}(0,1)$ and are independent.*

The theorem implies that $x = (y - X\hat{\beta}_0)$ is $\mathcal{N}(0, \Gamma)$ with $\Gamma = V_0 := \hat{\Sigma}_0 - X(X^T\hat{\Sigma}_0^{-1}X)^{-1}X^T$ and $W = \mathbf{K}$ is positive semi-definite (=non-negative definite). Thus, we obtain that $Q_{long} = \sum_{l=i}^{L} \lambda_l \chi_1^2$, where λ_l are eigenvalues of $\frac{1}{2}\hat{\Sigma}_0^{-1}K\hat{\Sigma}_0^{-1}V_0$. The p-values can be computed as described in the previous section.
When we perform a pathway analysis either for a cross-sectional or longitudinal phenotype with KMR, we have a big advantage. By testing $H_0 : \tau = 0$, our model under the null hypothesis does not contain the kernel matrix modelling the tested pathway. When testing multiple pathways simultaneously the null model thus needs to be fitted only once. This represents a large benefit, particularly for a longitudinal pathway analysis for which the kernel matrix can have large dimensions. Nevertheless, a multiple testing correction needs to be performed.

2.5. Variations of longitudinal kernel machine regression

In genetic longitudinal studies, there can also be an interest in testing the genetic-time interaction effect. Thus SNPs associated with the change over time, i.e. influencing the phenotype development are of interest. This interaction effect can be integrated by adding a kernel modelling $t \times G$ to model 11 in which t is the time vector for the m measurements and G the genotype matrix. When considering the model in LMM form with the notation as above we get:

$$y = X\beta + h_1 + h_2 + Zb + \epsilon, \tag{16}$$

where $\epsilon \sim \mathcal{N}(0, R)$ with $R = diag(R_1, \ldots, R_n)$, $b \sim \mathcal{N}(0, D)$ with $D = diag(D_1, \ldots, D_n)$ and $h_1 \sim \mathcal{N}(0, \tau_1\mathbf{K_1})$ and $h_2 \sim \mathcal{N}(0, \tau_2\mathbf{K_2})$. The non-parametric function h_2 models the genetic-time interaction $t \times G$ with the kernel matrix $\mathbf{K_2}$. The former $\mathbf{K}_1$ contains the main genetic effect G as for long-KMR. The marginal distribution for model 16 with the genetic-time-interaction effect is $y \sim \mathcal{N}(X\beta, \Sigma = R + \tau_1\mathbf{K_1} + \tau_2\mathbf{K_2} + ZDZ^T)$. By testing the genetic-time-interaction effect for association the null hypothesis transforms to

$H_0 : \tau_2 = 0$. To gain the correct score statistic, we perform the previous steps with slight alterations. First, we take the derivatives with respect to τ_2 with $\frac{\partial \Sigma}{\partial \tau_2} = \mathbf{K_2}$. Second, we need to take into account that $\Sigma = R + \tau_1 \mathbf{K_1} + \tau_2 \mathbf{K_2} + ZDZ^T$ and $\Sigma_0 = R_0 + \tau_1 \mathbf{K_1} + ZDZ^T$. Here, we consider only the score $U(\tau)$ (please refer to chapter 2.4, equation 13), which looks as follows under H_0:

$$U(\tau)|_{\tau_2=0} = -\frac{1}{2} tr(P_0 \mathbf{K_2}) + \frac{1}{2}(y - X\hat{\beta}_0)^T \Sigma_0^{-1} \mathbf{K_2} \Sigma_0^{-1} (y - X\hat{\beta}_0),$$

where P_0 is defined as in equation 13. However, Σ_0 is the covariance-variance matrix of model 16 and thus, has a different structure (described above). The final score test statistic is:

$$Q_{longI} = \frac{1}{2}(y - X\hat{\beta}_0)^T \hat{\Sigma}_0^{-1} \mathbf{K_2} \hat{\Sigma}_0^{-1} (y - X\hat{\beta}_0), \tag{17}$$

where $\hat{\beta}_0 = (X^T \hat{\Sigma}_0^{-1} X)^{-1} X^T \hat{\Sigma}_0^{-1} y$ and Σ_0 is replaced with its estimate $\hat{\Sigma}_0$. The test statistic Q_{longI} is also a quadratic form and thus, $Q_{longI} = \Sigma_{l=1}^{L} \lambda_l \chi_1^2$ with λ_l being the eigenvalues of $\frac{1}{2} \hat{\Sigma}_0^{-1} \mathbf{K_2} \hat{\Sigma}_0^{-1} V_0$ with $V_0 = \hat{\Sigma}_0 - X(X^T \hat{\Sigma}_0^{-1} X)^{-1} X^T$. The computation of p-values can be performed as for KMR and long-KMR. However, this model has some disadvantages. As this null model still contains the kernel modelling the main genetic effect, the null model needs to be fitted for every pathway to be tested. Secondly, we fit the KMR by integrating the main genetic kernel as covariance-variance matrix. Both aspects are computationally very expensive.

To reduce the computation time of this analysis of the genetic-time-interaction, we looked for another approach. This approach is similar to the correction of population stratification in a GWAS for which we perform a PCA on the genotype matrix including data from across the genome and receive a number of PCs. These PCs describe the variance in the genetic data. Here, we execute a PCA on the kernel matrix to obtain the PCs. This application of PCA is denoted as kernel principal component analysis, KPCA [65] and has already been conducted in different scenarios [29, 64].
We apply KPCA on K_1, which is the kernel modelling the main genetic effect and is only computed on one measurement point. We receive PCs, which are integrated in the long-KMR. Thus, we describe the similarities modelled in K_1 with a number of top PCs downgrading model 16 to model 11, a basic long-KMR with only one single kernel matrix. This hugely reduces the computational costs. Nevertheless, we can correct for the main genetic effect whilst modelling and testing for the genetic-time-interaction effect.

3. The PsyCourse Study, a longitudinal multi-centre study

The PsyCourse Study [8] is a longitudinal multi-centre study comprising individuals from the affective-to-psychotic continuum and mentally healthy individuals. The affective-to-psychotic continuum encompasses diseases, such as schizophrenia, and bipolar disorders I and II. One of the main goals in the development of the study was to analyse the phenotypic and genetic overlap of schizophrenia and bipolar disorder [8]. To this point only few longitudinal projects studied this connection [8].
The PsyCourse Study enables the analysis of various other research issues owing to the large number of available phenotypes. For example, the study comprises a variation of cognitive tests enabling the analysis of executive functions (EFs). EFs are responsible for controlling and coordinating mental processes essential to all human beings and are often impaired in various psychological disorders, for example, schizophrenia [18]. In this thesis, EFs are considered as the main phenotypes for which the aim is to study the genetic underpinnings of the longitudinal course.

3.1. The study details

We applied two different versions of the PsyCourse Study, Version 3.0, and Version 5.0. At the point of our first analysis (Version 3.0), the study was still on-going and probands had still been recruited. Version 5.0 [34] contains all available data presenting the PsyCourse Study after the recruitment has been finalised.
The PsyCourse Study comprises approximately 1800 individuals in which four-fifths are individuals with a disorder of the affective-to-psychotic spectrum and one-fifths are controls. The controls are males and females who have no history of mental disorders and are older than 18 years. A number of different diagnoses are presented in the mentally-ill patients. These diagnoses were assessed according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) and a subset according to the ICD-10 criteria. To classify the individuals suffering from a mental disorder, the individuals were broadly divided in two diagnostic groups, the affective and the psychotic group. The affective group comprises patients with predominantly affective symptoms including individuals with bipolar-I disorder (DSM-IV 296.X) or bipolar-II disorder (DSM-IV 296.89) and major depression (DSM-IV 296.3). Probands with predominantly psychotic symptoms are part of the psychotic group, which are patients with schizophrenia (DSM-IV 295.10/295.20/295.30/295.60/295.90 and ICD-10 schizophrenia),

schizo-affective disorder (DSM-IV 295.70), schizophreniform (DSM-IV 295.40) and brief psychotic disorder (DSM-IV 298.80). The affective and psychotic group represent with the control group the three broad diagnostic groups of the PsyCourse Study. The individuals were recruited in a number of different study centres in Germany, and in Austria. All individuals gave written consent and the study protocol was approved from the respective ethic committees at the various recruitment centres [8].

The longitudinal study comprises four measurement points, each approximately six months apart. For each individual the study lasted approximately 18 months in which the individuals participated at various questionnaires at each single measurement point. The test battery spans a number of aspects including basic demographic information, family and psychiatric history, medical treatments (e.g. medication), clinical symptomatology, and neuropsychological assessments [8]. Different clinical rating scales, such as the Global Assessment of Functioning (GAF) are part of the clinical symptomatology (please refer to [34] for more details). GAF evaluates the psychosocial functioning of an individual. The neuropsychological assessments comprises different neuropsychological tests, e.g. the Trail Making Test (TMT) or the Verbal Digit Span (VDS) Test. As these tests assess EFs, we will explain these tests in more details in the next chapter.
In addition to deep-phenotyping, biological materials were sampled at each measurement point including DNA, RNA, plasma and serum samples [8]. In order to determine genotypes of the individuals the Illumina Infinium PsychArray and the Illumina Infinium GSAChip (Global Screening Array) were applied, followed by thorough quality control (QC) [1, 67].

3.2. Executive functions

Executive functions are a specific group of higher-level cognitive abilities [27] influencing different mental processes. EFs comprise functions that organize, plan, and complete daily tasks and thus, are essential in the daily life even affecting success in life of every individual [18]. These cognitive abilities develop over time and decline with increasing age. Besides the decline in age EFs are also often impaired in individuals suffering from a mental disorder, for example, schizophrenia patients [18]. EFs are often rated under the first symptoms of a mental disorder, such as for Alzheimer's disease. Owing to their control of essential tasks, impaired EFs compromise largely the quality of life of individuals with mental disorders.
Neurobiologically, it is well-documented that EFs are connected with the pre-frontal cor-

tex [25, 57]. Twins studies also lend support to EFs being one of the most heritable traits in psychology [26]. Various studies including twin studies and molecular genetic studies [74] investigated the genetic background. However, most of the molecular genetic studies did not detect any specific genetic marker [74]. Thus, the understanding of the genetic background of EFs is of great interest, in particular the genetic influence on the development of EFs over time.

Most commonly, EFs are distinguished in three core areas, which have according to the "unity and diversity" concept a common latent factor [26]. The three core skills are (i) set-shifting (cognitive flexibility), (ii) updating (working memory), and (iii) inhibition [18, 57]. The set-shifting ability allows an individual to flexibly deal and adjust to new tasks and challenges [18, 26]. The updating skill enables an individual to keep information in mind and to process and modify this information [18, 26]. The last core skill, inhibition supports the self-control and impulse control of an individual [18]. A number of different neuropsychological tests are available to assess the EF abilities, such as the Trail Making Test for set-shifting and the Verbal Digit Span Test for working memory. The measurement of inhibition is often performed with go/no-go and stop-signal tasks [18].

The focus of our analyses lies on the set-shifting and updating abilities and thus we explain shortly the two cognitive tests applied to assess these skills. The Trail Making Test, part B (TMT-B) [6] is utilised to measure the set-shifting ability. The TMT-B is the second part of the two-part Trail Making Test. In the TMT-B, the time in seconds is measured in which an individual is able to alternately connect numbers (numbers: 1-26) and letters of the alphabet in an ascending order. The time in seconds is used as TMT-B test score.

For the updating, the Verbal Digit Span Test Backwards (VDS-B) [37] is performed. Here, the study participant is presented verbally with up to seven pairs of number sequences of increasing length by a trained interviewer. Each sequence of numbers needs to be repeated backwards by the study participant. For each correctly named sequence the participant gains a point with a maximum score of two for each sequence pair. The test is terminated when both of the sequence are falsely repeated. The sum of all correctly repeated sequence pairs represents the VDS-B test score.

4. Summaries

4.1. A genome-wide association study of longitudinal course of executive functions

Executive functions are a group of essential cognitive abilities that highly influence the quality of life. The change over time of EFs is of great importance, especially for mentally-ill individuals. Thus, it is also of great interest to understand the genetic basis of the course over time to improve treatments for patients suffering from mental disorders. We conducted two GWASs to investigate this genetic basis. In particular, we aimed to identify SNPs influencing the short-term course (comprising 18 months) of EFs. The focus lay on two EF core abilities: set-shifting and updating (see chapter 3.2), which were assessed with two cognitive tests, the Trail Making Test, part B (TMT-B), and the Verbal Digit Span Test Backwards (VDS-B), respectively. In each GWAS, a LMM was applied in which the interaction effect of SNP and time were tested for association.

Data from the longitudinal PsyCourse Study [8] Version 3.0, were implemented to study the short-term course. A total of 1338 genotyped individuals were included in the analysis, comprising 550 affective patients, 530 psychotic individuals, and 258 controls. The Illumina Infinium PsychArray was utilised to determine the genotypes of the individuals. After an imputation with SHAPEIT2/IMPUTE2 and standard quality control (QC) [1], approximately 8.2 million SNPs in each GWAS were tested.

For each phenotype, the missing measurement points across the different diagnostic groups and measurement points were studied, and a MAR mechanism was assumed based on this analysis. The TMT-B was log transformed (lgTMT-B) to yield a normally distributed phenotype. For a phenotype Y (lgTMT-B or VDS-B), the LMM for individual i at measurement point t_{ij} with $j = 1, 2, 3, 4$ was as follows:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 age_i + \beta_3 gender_i + \beta_4 diagnosis_i + \sum_{k=1}^{5} \beta_{4+k} PC_{ik} + b_{0i} + b_{1i} t_{ij} + c_i center_i + \beta_{10} SNP_i + \beta_{11} SNP_i \cdot t_{ij} + \epsilon_{ij},$$

where age_i, $gender_i$, and $diagnosis_i$ represent the age at first measurement, gender, and diagnostic group of individual i, respectively. These fixed effects were added to correct for confounders. To adjust for population stratification, the model contains the top five ancestry components (PC_{ik}), which were determined with a PCA. The random effects of the LMM encompassed a random intercept, random slope, and a random centre effect.

The latter was integrated because of the multiple recruitment centres in the PsyCourse Study. The former random effects model the individual course and correct for the correlated measurements. The genetic data were included as main genetic effect with SNP_i and as time-interaction term, $SNP_i \cdot t_{ij}$. The genetic-time-interaction effect was the main focus in the GWASs, as the interest lay in SNPs influencing the change over time of the phenotypes. Thus, we tested the interaction effect for significance.

For lgTMT-B, nine genome-wide significant SNPs were identified. The nine SNPs were all positioned in one single linkage disequilibrium block ($r^2 \geq 0.85$), i.e. a block of strongly correlated SNPs on chromosome 5. The top SNP was rs150547358 with a p-value of 7.2×10^{-10} and an effect estimate $\hat{\beta}$ of 1.16 seconds per measurement on the original TMT-B scale. Four of the nine SNPs were directly located in an intron region of ring finger protein 180 (RNF180), the product of which is involved in protein modification. For VDS-B, no genome-wide significant hits were determined.

A replication study with data of the research consortium FOR2107 [44] was performed to replicate the TMT-B findings. FOR2107 is a longitudinal cohort study comprising two measurement points approximately two years apart. The sample implemented in the replication study comprised 1795 individuals (851 affective, 112 psychotic patients, and 832 controls). As only two measurements were available, we applied a simple difference model instead of the LMM of the GWAS to avoid a high number of parameters owing to the random effects. Thus, we considered the difference of the lgTMT-B measurements as phenotype and added the main genetic effect (SNP_i), age, gender, diagnostic group, and PCs as covariates to the model. SNP_i was tested for significance and the effect interpreted as the difference between the average change of the TMT-B between the genotypes. Aiming to replicate the linkage disequilibrium block, the significance level remained at 5%. The previously identified top SNP rs150547358 had a p-value of 0.015 and was thus successfully replicated.

4.2. Kalpra: a kernel approach for longitudinal pathway regression analysis integrating network information with an application to the longitudinal PsyCourse Study

Longitudinal pathway analyses are still uncommon and only a limited number of approaches are available. Here, the main goal was to expand the existing KMR pathway analysis to study longitudinal data. To create a longitudinal KMR or long-KMR, the esti-

mation equivalence of the KMR and LMM was exploited. The basic KMR model was expanded by adding random effects modelling the correlation of the multiple measurements within each individual. The details are presented in chapter 2.4. Moreover, long-KMR is able to integrate pathway topology information by applying the network-based kernel [24]. This network kernel models gene interaction by integrating the adjacency matrix of the pathway, which describes the connections between the single genes (please refer to chapter 2.1). Thus, long-KMR can be considered as topology-based pathway analysis. Long-KMR has been implemented as an R package *kalpra* (see next section).

A similar approach was developed by Yan et al. [77], denoted as KMgene. KMgene tests single genes for association with a longitudinal phenotype. In contrast, long-KMR studies a whole pathway for significance enabling the modelling of a pathway with different kernels. Moreover, long-KMR allows the integration and testing of a genetic-time-interaction effect either with or without a simultaneous adjustment for the main genetic effect (see chapter 2.5). By testing this interaction effect for association, we study the genetic influence on the course over time of a phenotype.

We evaluated the performance of long-KMR in a simulation study exploring different aspects in various simulation settings. There the following questions were addressed:

- How does the performance of long-KMR change when the number of measurement points in the longitudinal study alters?
  Here, longitudinal studies with two and four measurements were considered, simulating complete phenotype data and data sets with 25% and 50% missing measurements, and assuming a MAR mechanism to contemplate the influence of missing data. For the two-measurements data, long-KMR was compared with the ANCOVA model. When studying four measurements, KMgene was applied as comparison model.

- How does the network kernel influence the performance of long-KMR?
  We compared the performance of long-KMR when using a linear kernel and a network kernel. The network kernel was also studied in regard to its behaviour when the pathway topology changed. Here, the pathway density was selected as distinctive characteristic. The density (d) is a graph-theoretical feature, which is defined as the ratio of the existing gene connections in the pathway divided by the number of maximal available links ($d \in [0, 1]$). Three pathways were compared with a density of 0.2, 0.5 and 0.8.

In each scenario, the simulated study sample embraced 1000 unrelated individuals and 950 simulated SNPs. A pathway serving as foundation pathway was required to apply the network kernel. Thus, a real pathway from the Reactome database [40], the "signaling by ERBB4" pathway [68] was selected, which is related to our application example. This pathway contains 19 genes and has a density of 0.46 (rounding up to 0.5). Based on this foundation pathway, two additional artificial pathways with a density of 0.8 and 0.2 were created.

To assess the type I error rate, the phenotypes were simulated with a LMM containing no genetic effects. For the power simulation, we designed three genetic effect models for which each model contained three causal "pseudo-" genes comprising three causal SNPs (in total: nine causal SNPs). We created two single-effect models, the main genetic and time-interaction-effect model. The former contained only the following sum of additive SNP effects $\sum_{k=1}^{9} \beta_k \cdot SNP_{ik}$ for each individual i. The time-interaction model included $\sum_{k=1}^{9} \beta_k \cdot (SNP_{ik} \cdot t_{ij})$ for each individual i at time point t_{ij} $(j = 1, \ldots, m)$ to model a time-interaction effect. In addition, the joint model incorporated both sums $(\sum_{k=1}^{9} \beta_k \cdot SNP_{ik} + \sum_{k=1}^{9} \beta_k \cdot (SNP_{ik} \cdot t_{ij}))$, representing a more complex model. The main genetic kernel was tested for association in the single-effect model containing only the main genetic effect, whereas in the latter two, the time-interaction kernel was tested. For the joint model one effect size $\beta = 0.04$ was studied; for the single-effect models, there were three effect sizes β =0.04, 0.06, and 0.08.
In the analysis of the joint model, we performed a kernel principal component analysis (KPCA, please refer to chapter 2.5 for details) on the kernel modelling the main genetic effect. Adding the top two principal components to the analysis model enabled the adjustment for the main genetic effect. We applied this approach to reduce the computational costs.

Overall, long-KMR maintained the type I error rates for the different simulations settings and demonstrated slightly conservative rates. In contrast, the combined pathway p-values of KMgene displayed inflated type I error rates. For the single-effect models (both main genetic and time-interaction), the power simulation revealed that the power of long-KMR increased with increasing number of available measurement points. Furthermore, long-KMR demonstrated a higher power in comparison to both the ANCOVA model and KMgene. The network kernel displayed a superior power to the linear kernel. Here, the power gain for the network kernel increased with decreasing density (d0.2 > d0.5 > d0.8). These observations were reflected in the single-effect models for all effect

sizes. In the joint model, only one exception was identified.

For the real-world example, we applied data from the PsyCourse Study [8], Version 5.0. Here, the focus lay solely on the TMT-B assessing the set-shifting ability (see chapter 3.2 for details) because of the results of the first publication (please refer to chapter 4.1). In total, 17 pathways selected from the Reactome database [40] were tested for association with TMT-B. The SNPs of the PsyCourse Study genotyped with Illumina Infinium Global Screening Array-24 Kit were mapped according to their genomic location ($\pm$ 500 kbp).
The TMT-B was log-transformed for normality. Each analysis model included the following fixed effects: age at first measurement, gender, diagnostic group, measurement point, and the top five ancestry principal components. A random intercept and a random slope were integrated to correct for the correlation of the longitudinal data. For each pathway, the genetic data were modelled with a linear kernel, a network kernel, and a size-adjusted network kernel. We modelled and tested each pathway as main genetic and genetic-time-interaction effect. The multiple testing correction was conducted with the Gao approach [36], for which the number of effective pathways was determined.
None of the pathways was significantly associated with TMT-B according to α_{Gao}. However, seven pathways had a p-value < 0.05, e.g. the "signaling by ERBB4" pathway that was previously chosen as foundation pathway for the simulations.
All in all, long-KMR enables a longitudinal topology-based pathway analysis to be conducted, in which genetic effects can be flexibly modelled. In particular, long-KMR allows testing a genetic-time-interaction effect to investigate the longitudinal course of a phenotype.

4.3. R package *kalpra*: Kernel Approach for Longitudinal Pathway Regression Analysis

We developed the R package *kalpra* - Kernel Approach for Longitudinal Pathway Regression Analysis that enables a longitudinal pathway analysis to be performed with a kernel machine regression approach. *kalpra* has various features and enables the user to execute a longitudinal analysis for normally distributed phenotypes, as well as a cross-sectional kernel regression for binary and normally distributed phenotypes. The core functions of the package listed in the following conduct the variance component test and compute the p-value for each pathway. The function

- *KMR.Cross.bin()* performs a KMR for binary cross-sectional phenotypes,

- *KMR.Cross.quan()* studies a normally distributed cross-sectional phenotype and
- *KMR.Long()* conducts long-KMR, the analysis of a normally distributed longitudinal phenotype.
- *KMR.KernelTime()* allows modelling and testing of a genetic-time-interaction effect while adjusting for the main genetic effect.

For all analysis types, two p-value methods are available: Davies' method [14] and/or the Satterthwaite approximation [61]. The first three functions can handle missing phenotype values as long as the MAR can be assumed.

The user can model the pathway with three different kernels: the linear, the quadratic, and the network-based kernel developed by Freytag et al. [24]. Depending on the chosen kernel matrix different input data are required. However, for all kernels, the genotype data need to be presented as a matrix with no missing values, which only contains 0s,1s, or 2s. The determined kernel matrix then serves as input to the above described core functions. When a longitudinal study is conducted, data need to be transformed into long format. In long format, each individual has multiple rows, one for each measurement point. The transformation into long format can be performed with *makeLongGenotype()* or *reduceGenotype()* (see Fig.2), whereas the latter function attunes the dimensions of the genotype matrix to the phenotype data containing missing values. In these two functions, the user can also select to model a main genetic or genetic-time-interaction effect with the kernel.
As the network kernel integrates additional network information, *kalpra* is able to download information directly from the Pathway Commons databases [60]. Pathway Commons holds information from different pathway database, e.g. Reactome [22]. The databases can be searched via the function *searchPathway()* either for a pathway name or based on a keyword. The pathway is downloaded in SIF (standard interchange format) and can be directly transformed to an adjacency matrix. The adjacency matrix describes the gene interactions. Furthermore, an annotation matrix can be created by assigning SNPs to the genes of the pathway downloaded. The adjacency and annotation matrix are required as input data for the network kernel. The pathway can also be graphically illustrated and different pathway characteristics can be determined, e.g. number of nodes or average node degree.
Figure 2 displays the steps required to conduct a pathway analysis.

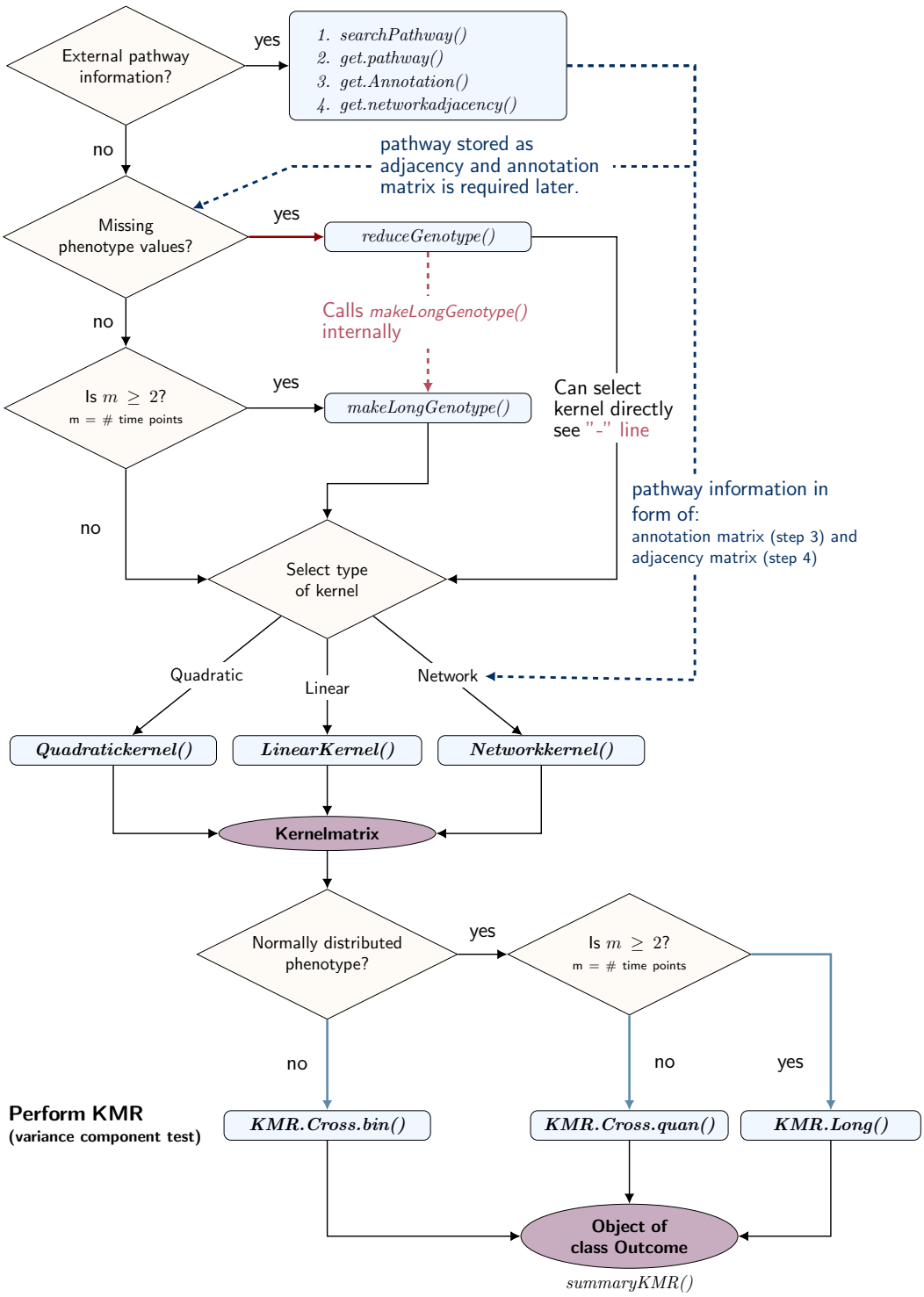


Figure 2. Example workflow of a pathway analysis performed with *kalpra*.
(source: `https://gitlab.gwdg.de/bernadette.wendel/kalpra`)

5. Discussion

Genetic association studies are performed in miscellaneous forms, including single genetic marker tests and whole pathway association analyses. This cumulative thesis focuses on GWASs and pathway analyses examining longitudinal data. In order to execute a longitudinal pathway analysis, the KMR analysis was extended to long-KMR. An overview of the diverse KMR modifications are given by Larson et al. [46]. The large number of available KMR analyses reflects the versatility of the approach and the underlying kernel methods.
Kernel methods are highly flexible, which has many advantages. The kernel itself needs only to be positive semi-definite allowing a number of kernel variations, for example, kernels including biological information or modelling genetic interaction effects. Moreover, the basic idea of kernel methods results in a dimensionality reduction of the analysed data. Here, the high-dimensional genotype matrix is transformed into a low-dimensional kernel matrix for which the dimensions are in general only dependent on the study sample size.

We developed long-KMR, exploiting these flexibilities. For example, long-KMR enables the researcher to apply the network kernel [24], creating a topology-based longitudinal pathway analysis that includes additional biological information. Furthermore, we can model a genetic-time-interaction effect with the kernel matrix enabling study of the longitudinal course of phenotypes. However, these benefits can also create obstacles. A kernel matrix in long-KMR does not only depend on the sample size but also on the number of measurement points in the longitudinal study. This can result in a high-dimensional kernel matrix for which the computational costs increase. Depending on the longitudinal study design, long-KMR can reach its computational limits. At this point, most longitudinal studies are limited by the high temporal and financial costs, leading to either a relatively small sample size or a small number of measurements. In this setting, the computation of long-KMR is nevertheless still feasible. Advancements are necessary when larger longitudinal studies are available. This progress is especially important when more complex KMR models are studied. For example, models can include more than one kernel, such as main genetic and genetic-time-interaction effects.

The flexibility of the kernel definition allows a great number of options to integrate biological pathway information. Schaid [63] suggested defining a kernel matrix K integrating genomic information as $K = GSG^T$, where G is the genotype matrix and the matrix S

contains similarity score of SNPs. However, many options can lead to many questions concerning methodological and practical issues. These questions include: "which information should be modelled?", "how should the information be modelled?" or "how can the results be interpreted?", to name a few examples. One possible approach is to model a specific biological hypothesis that can be applied to develop weighting functions modelling more of the pathway topology. This specific biological information can be obtained from a variety of available pathway databases.
In this context, we also want to mention a great potential of long-KMR, which was not discussed until now. Apart from the analysis of the longitudinal course of phenotypes, long-KMR is also capable to analyse molecular data varying over time. Until now solely SNPs are tested, which maintain the same variation at all measurement points. However, there are also molecular data available that change over time, such as methylation or gene expression data. This type of data enables a great number of opportunities. The kernel matrix can either directly model these genetic data or utilise it to create detailed weighting functions for pathways. Because of the flexibility of the kernel definition, this different kind of information can be utilised in various ways.

This thesis also focuses on the analysis of EFs. In particular, we studied two core EF abilities, the set-shifting and updating skill with respect to their genetic background. Of great interest is the short-term course of the single core abilities. These analyses cover only a small amount of the EF variability. According to the "unity and diversity" model developed by Friedman et al. [26], all EF core abilities have a latent common underlying EF factor. Thus, the analysis of single core EF abilities reflects only upon the individual skills and does not consider the common factor. Hatoum et al. [33] computed a latent common EF factor by performing a confirmatory factor analysis. Here, Hatoum et al. [33] also conducted a GWAS in which 129 genome-wide significant SNPs were identified. One can also conduct a confirmatory factor analysis to compute a common EF factor for each individual in the PsyCourse Study data. Subsequently, a GWAS may be performed, either on the cross-sectional or longitudinal level, for this common EF factor to analyse the genetic basis. In this context, it would also be interesting to perform a pathway analysis for the common EF factor. These results can be compared to the results of the single core EF skill (e.g. TMT-B) pathway analyses to identify common or different pathways. To this point, the molecular genetic basis of EFs still remain largely unknown. Thus, more analyses are required in the future to unravel this complex EF structure.
To decipher this underlying EF structure, long-KMR and/or other KMR variations may also be applied. However, EFs are not the only application. Because of the flexible mod-

elling of genomic data, KMR analyses can be utilised to study various innovative research questions. For example, one possible analysis direction can be to examine multiple phenotypes at once, i.e. a multivariate KMR with multiple outcome phenotypes. This allows the analysis of correlated phenotypes in a convenient setting. Other challenges accompanying the analysis of longitudinal data may also be considered to further expand long-KMR. One important aspect is how to handle missing phenotype measurements. One possibility here is to impute the missing phenotype values and integrate this in long-KMR, for example by including multiple imputation or other imputation approaches. Overall, the presented method still allows numerous options for further extensions, which will also open doors to new applications.

Bibliography

[1] T. F. M. Andlauer, D. Buck, G. Antony, A. Bayas, L. Bechmann, A. Berthele, A. Chan, C. Gasperi, R. Gold, C. Graetz, J. Haas, M. Hecker, C. Infante-Duarte, M. Knop, T. Kümpfel, V. Limmroth, R. A. Linker, V. Loleit, F. Luessi, S. G. Meuth, M. Mühlau, S. Nischwitz, F. Paul, M. Pütz, T. Ruck, A. Salmen, M. Stangel, J.-P. Stellmann, K. H. Stürner, B. Tackenberg, F. T. Bergh, H. Tumani, C. Warnke, F. Weber, H. Wiendl, B. Wildemann, U. K. Zettl, U. Ziemann, F. Zipp, J. Arloth, P. Weber, M. Radivojkov-Blagojevic, M. O. Scheinhardt, T. Dankowski, T. Bettecken, P. Lichtner, D. Czamara, T. Carrillo-Roa, E. B. Binder, K. Berger, L. Bertram, A. Franke, C. Gieger, S. Herms, G. Homuth, M. Ising, K.-H. Jöckel, T. Kacprowski, S. Kloiber, M. Laudes, W. Lieb, C. M. Lill, S. Lucae, T. Meitinger, S. Moebus, M. Müller-Nurasyid, M. M. Nöthen, A. Petersmann, R. Rawal, U. Schminke, K. Strauch, H. Völzke, M. Waldenberger, J. Wellmann, E. Porcu, A. Mulas, M. Pitzalis, C. Sidore, I. Zara, F. Cucca, M. Zoledziewska, A. Ziegler, B. Hemmer, and B. Müller-Myhsok. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. *Science Advances*, 2(6), jun 2016. doi:10.1126/sciadv.1501678.

[2] G. D. Bader. Pathguide: a pathway resource list. *Nucleic Acids Research*, 34(90001): D504–D506, jan 2006. doi:10.1093/nar/gkj126.

[3] G. V. Belle, L. D. Fisher, P. J. Heagerty, and T. Lumley. Longitudinal Data Analysis. In *Wiley Series in Probability and Statistics*, pages 728–765. John Wiley & Sons, Inc., Sept. 2004. doi:10.1002/0471602396.ch18.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, Jan. 1995. doi:10.1111/j.2517-6161.1995.tb02031.x.

[5] H. Bickeböller and C. Fischer. *Einführung in die Genetische Epidemiologie*. Springer Berlin Heidelberg, 2007. doi:10.1007/978-3-540-33568-9.

[6] C. R. Bowie and P. D. Harvey. Administration and interpretation of the trail making test. *Nature Protocols*, 1(5):2277–2281, Dec. 2006. doi:10.1038/nprot.2006.390.

[7] A. J. Brookes. The essence of SNPs. *Gene*, 234(2):177–186, July 1999. doi:10.1016/s0378-1119(99)00219-x.

[8] M. Budde, H. Anderson-Schmidt, K. Gade, D. Reich-Erkelenz, K. Adorjan, J. L. Kalman, F. Senner, S. Papiol, T. F. M. Andlauer, A. L. Comes, E. C. Schulte, F. Klöhn-Saghatolislam, A. Gryaznova, M. Hake, K. Bartholdi, L. Flatau, M. Reitt, S. Quast, S. Stegmaier, M. Meyers, B. Emons, I. S. Haußleiter, G. Juckel, V. Nieratschker, U. Dannlowski, S. K. Schaupp, M. Schmauß, J. Zimmermann, J. Reimer, S. Schulz, J. Wiltfang, E. Reininghaus, I.-G. Anghelescu, V. Arolt, B. T. Baune, C. Konrad, A. Thiel, A. J. Fallgatter, C. Figge, M. von Hagen, M. Koller, F. U. Lang, M. E. Wigand, T. Becker, M. Jäger, D. E. Dietrich, S. Stierl, H. Scherk,

C. Spitzer, H. Folkerts, S. H. Witt, F. Degenhardt, A. J. Forstner, M. Rietschel, M. M. Nöthen, P. Falkai, T. G. Schulze, and U. Heilbronner. A longitudinal approach to biological psychiatric research: The PsyCourse study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 180(2):89–102, Aug. 2018. doi:10.1002/ajmg.b.32639.

[9] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, Nov. 2018. doi:10.1093/nar/gky1120.

[10] W. S. Bush and J. H. Moore. Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, dec 2012. doi:10.1371/journal.pcbi.1002822.

[11] E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli. Longitudinal studies. *Journal of Thoracic Disease*, 7(11):E537–E540, 2015. doi:10.3978/j.issn.2072-1439.2015.10.63.

[12] Y.-F. Chiu, A. E. Justice, and P. E. Melton. Longitudinal analytical approaches to genetic data. *BMC Genetics*, 17(S2), Feb. 2016. doi:10.1186/s12863-015-0312-y.

[13] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan. Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2):121–133, Feb. 2011. doi:10.1038/nprot.2010.182.

[14] R. B. Davies. Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, 29(3):323, 1980. doi:10.2307/2346911.

[15] C. A. de Leeuw, J. M. Mooij, T. Heskes, and D. Posthuma. MAGMA: Generalized gene-set analysis of GWAS data. *PLOS Computational Biology*, 11(4):e1004219, Apr. 2015. doi:10.1371/journal.pcbi.1004219.

[16] C. A. de Leeuw, B. M. Neale, T. Heskes, and D. Posthuma. The statistical properties of gene-set analysis. *Nature Reviews Genetics*, 17(6):353–364, Apr. 2016. doi:10.1038/nrg.2016.29.

[17] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55 (4):997–1004, Dec. 1999. doi:10.1111/j.0006-341x.1999.00997.x.

[18] A. Diamond. Executive functions. *Annual Review of Psychology*, 64(1):135–168, Jan. 2013. doi:10.1146/annurev-psych-113011-143750.

[19] R. Diestel. *Graph Theory*. Graduate Texts in Mathematics. Springer, New York, NY, 2 edition, Feb. 2000.

[20] P. J. Diggle, P. J. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford Statistical Science Series. OUP Oxford, 2002. ISBN 9780198524847.

[21] J. Engels. Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56(10):968–976, Oct. 2003. doi:10.1016/s0895-4356(03)00170-7.

[22] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, M. Milacic, C. D. Roca, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, G. Viteri, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, Nov. 2017. doi:10.1093/nar/gkx1132.

[23] L. Fahrmeir, T. Kneib, and S. Lang. *Regression*. Springer Berlin Heidelberg, 2009. doi:10.1007/978-3-642-01837-4.

[24] S. Freytag, J. Manitz, M. Schlather, T. Kneib, C. I. Amos, A. Risch, J. Chang-Claude, J. Heinrich, and H. Bickeböller. A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Human Heredity*, 76(2):64–75, 2013. doi:10.1159/000357567.

[25] N. P. Friedman and A. Miyake. Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, 86:186–204, Jan. 2017. doi:10.1016/j.cortex.2016.04.023.

[26] N. P. Friedman, A. Miyake, S. E. Young, J. C. DeFries, R. P. Corley, and J. K. Hewitt. Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental Psychology: General*, 137(2):201–225, May 2008. doi:10.1037/0096-3445.137.2.201.

[27] N. P. Friedman, A. Miyake, L. J. Altamirano, R. P. Corley, S. E. Young, S. A. Rhea, and J. K. Hewitt. Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. *Developmental Psychology*, 52(2):326–340, Feb. 2016. doi:10.1037/dev0000075.

[28] N. A. Furlotte, E. Eskin, and S. Eyheramendy. Genome-wide association mapping with longitudinal data. *Genetic Epidemiology*, 36(5):463–471, May 2012. doi:10.1002/gepi.21640.

[29] Q. Gao, Y. He, Z. Yuan, J. Zhao, B. Zhang, and F. Xue. Gene- or region-based association study via kernel principal component analysis. *BMC Genetics*, 12(1), Aug. 2011. doi:10.1186/1471-2156-12-75.

[30] M. A. García-Campos, J. Espinal-Enríquez, and E. Hernández-Lemus. Pathway analysis: State of the art. *Frontiers in Physiology*, 6, dec 2015. doi:10.3389/fphys.2015.00383.

[31] T. Ge, J. W. Smoller, and M. R. Sabuncu. Kernel machine regression in neuroimaging genetics. In *Machine Learning and Medical Imaging*, pages 31–68. Elsevier, 2016. doi:10.1016/b978-0-12-804076-8.00002-5.

[32] J. L. Haines, M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, K. L. Spencer, S. Y. Kwan, M. Noureddine, J. R. Gilbert, N. Schnetz-Boutaud, A. Agarwal, E. A. Postel, and M. A. Pericak-Vance. Complement factor h variant increases the risk of age-related macular degeneration. *Science*, 308(5720):419–421, apr 2005. doi:10.1126/science.1110359.

[33] A. S. Hatoum, C. L. Morrison, E. C. Mitchell, M. Lam, C. E. Benca-Bachman, A. E. Reineberg, R. H. Palmer, L. M. Evans, M. C. Keller, and N. P. Friedman. Genome-wide association study shows that executive functioning is influenced by GABAergic processes and is a neurocognitive genetic correlate of psychiatric disorders. *Biological Psychiatry*, 93(1):59–70, Jan. 2023. doi:10.1016/j.biopsych.2022.06.034.

[34] U. Heilbronner, K. Adorjan, H. Anderson-Schmidt, M. Budde, A. L. Comes, K. Gade, M. Heilbronner, J. L. Kalman, M. O. Kohshour, S. Papiol, D. Reich-Erkelenz, S. K. Schaupp, E. C. Schulte, F. Senner, T. Vogl, P. Falkai, and T. G. Schulze. The psycourse codebook, version 5.0, 2021. URL `https://data.ub.uni-muenchen.de/id/eprint/251`.

[35] J. N. Hellwege, J. M. Keaton, A. Giri, X. Gao, D. R. V. Edwards, and T. L. Edwards. Population stratification in genetic association studies. *Current Protocols in Human Genetics*, 95(1), oct 2017. doi:10.1002/cphg.48.

[36] A. E. Hendricks, J. Dupuis, M. W. Logue, R. H. Myers, and K. L. Lunetta. Correction for multiple testing in a gene region. *European Journal of Human Genetics*, 22(3): 414–418, July 2013. doi:10.1038/ejhg.2013.144.

[37] S. Hilbert, T. T. Nakagawa, P. Puci, A. Zech, and M. Bühner. The digit span backwards task. *European Journal of Psychological Assessment*, 31(3):174–180, July 2015. doi:10.1027/1015-5759/a000223.

[38] P. Holmans. Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. In *Computational Methods for Genetics of Complex Traits*, pages 141–179. Elsevier, 2010. doi:10.1016/b978-0-12-380862-2.00007-2.

[39] J. G. Ibrahim and G. Molenberghs. Missing data methods in longitudinal studies: a review. *TEST*, 18(1):1–43, Feb. 2009. doi:10.1007/s11749-009-0138-x.

[40] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, F. Loney, B. May, M. Milacic, K. Rothfels, C. Sevilla, V. Shamovsky, S. Shorser, T. Varusai, J. Weiser, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, Nov. 2019. doi:10.1093/nar/gkz1031.

[41] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1): D457–D462, oct 2015. doi:10.1093/nar/gkv1070.

[42] B. Kerner, K. E. North, and M. D. Fallin. Use of longitudinal data in genetic studies in the genome-wide association studies era: summary of group 14. *Genetic Epidemiology*, 33(S1):S93–S98, 2009. doi:10.1002/gepi.20479.

[43] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, Feb. 2012. doi:10.1371/journal.pcbi.1002375.

[44] T. Kircher, M. Wöhr, I. Nenadic, R. Schwarting, G. Schratt, J. Alferink, C. Culmsee, H. Garn, T. Hahn, B. Müller-Myhsok, A. Dempfle, M. Hahmann, A. Jansen, P. Pfefferle, H. Renz, M. Rietschel, S. H. Witt, M. Nöthen, A. Krug, and U. Dannlowski. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *European Archives of Psychiatry and Clinical Neuroscience*, 269(8):949–962, Sept. 2018. doi:10.1007/s00406-018-0943-x.

[45] N. B. Larson, S. McDonnell, L. C. Albright, C. Teerlink, J. Stanford, E. A. Ostrander, W. B. Isaacs, J. Xu, K. A. Cooney, E. Lange, J. Schleutker, J. D. Carpten, I. Powell, J. E. Bailey-Wilson, O. Cussenot, G. Cancel-Tassin, G. G. Giles, R. J. MacInnis, C. Maier, A. S. Whittemore, C.-L. Hsieh, F. Wiklund, W. J. Catolona, W. Foulkes, D. Mandal, R. Eeles, Z. Kote-Jarai, M. J. Ackerman, T. M. Olson, C. J. Klein, S. N. Thibodeau, and D. J. Schaid. gsSKAT: Rapid gene set analysis and multiple testing correction for rare-variant association studies using weighted linear kernels. *Genetic Epidemiology*, 41(4):297–308, Feb. 2017. doi:10.1002/gepi.22036.

[46] N. B. Larson, J. Chen, and D. J. Schaid. A review of kernel methods for genetic association studies. *Genetic Epidemiology*, 43(2):122–136, Jan. 2019. doi:10.1002/gepi.22180.

[47] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, D. C. Christiani, M. M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, Aug. 2012. doi:10.1016/j.ajhg.2012.06.007.

[48] C. M. Lewis and J. Knight. Introduction to genetic association studies. *Cold Spring Harbor Protocols*, 2012(3):pdb.top068163–pdb.top068163, Mar. 2012. doi:10.1101/pdb.top068163.

[49] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, Dec. 2015. doi:10.1016/j.cels.2015.12.004.

[50] D. Liu, X. Lin, and D. Ghosh. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088, May 2007. doi:10.1111/j.1541-0420.2007.00799.x.

[51] B. Maher. Personal genomes: The case of the missing heritability. *Nature*, 456(7218): 18–21, Nov. 2008. doi:10.1038/456018a.

[52] D. Malzahn, A. Schillert, M. Müller, and H. Bickeböller. The longitudinal non-parametric test as a new tool to explore gene-gene and gene-time effects in cohorts. *Genetic Epidemiology*, 34(5):469–478, June 2010. doi:10.1002/gepi.20500.

[53] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct. 2009. doi:10.1038/nature08494.

[54] J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511, June 2010. doi:10.1038/nrg2796.

[55] J. N. S. Matthews. Summary measures analysis of longitudinal data, July 2005.

[56] A. Mishra and S. MacGregor. A novel approach for pathway analysis of GWAS data highlights role of BMP signaling and muscle cell differentiation in colorectal cancer susceptibility. *Twin Research and Human Genetics*, 20(1):1–9, Jan. 2017. doi:10.1017/thg.2016.100.

[57] A. Miyake, N. P. Friedman, M. J. Emerson, A. H. Witzki, A. Howerter, and T. D. Wager. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology*, 41 (1):49–100, Aug. 2000. doi:10.1006/cogp.1999.0734.

[58] G. Molenberghs and G. Verbeke. *Linear Mixed Models for Longitudinal Data*. Springer New York, 2000. doi:10.1007/978-1-4419-0300-6.

[59] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006. doi:10.1038/ng1847.

[60] I. Rodchenkov, O. Babur, A. Luna, B. A. Aksoy, J. V. Wong, D. Fong, M. Franz, M. C. Siper, M. Cheung, M. Wrana, H. Mistry, L. Mosier, J. Dlin, Q. Wen, C. O'Callaghan, W. Li, G. Elder, P. T. Smith, C. Dallago, E. Cerami, B. Gross, U. Dogrusoz, E. Demir, G. D. Bader, and C. Sander. Pathway commons 2019 update: integration, analysis and exploration of pathway data. *Nucleic Acids Research*, Oct. 2019. doi:10.1093/nar/gkz946.

[61] F. E. Satterthwaite. An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2(6):110, Dec. 1946. doi:10.2307/3002019.

[62] D. J. Schaid. Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. *Human Heredity*, 70(2):109–131, 2010. doi:10.1159/000312641.

[63] D. J. Schaid. Genomic similarity and kernel methods II: Methods for genomic information. *Human Heredity*, 70(2):132–140, 2010. doi:10.1159/000312643.

[64] B. Schölkopf, A. Smola, E. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319, 1998.

[65] B. Schölkopf, A. Smola, and K.-R. Müller. Kernel principal component analysis. In *Lecture Notes in Computer Science*, pages 583–588. Springer Berlin Heidelberg, 1997. doi:10.1007/bfb0020217.

[66] K. Sikorska, N. M. Montazeri, A. Uitterlinden, F. Rivadeneira, P. H. Eilers, and E. Lesaffre. GWAS with longitudinal phenotypes: performance of approximate procedures. *European Journal of Human Genetics*, 23(10):1384–1391, Feb. 2015. doi:10.1038/ejhg.2015.1.

[67] L. Smigielski, S. Papiol, A. Theodoridou, K. Heekeren, M. Gerstenberg, D. Wotruba, R. Buechler, P. Hoffmann, S. Herms, K. Adorjan, H. Anderson-Schmidt, M. Budde, A. L. Comes, K. Gade, M. Heilbronner, U. Heilbronner, J. L. Kalman, F. Klöhn-Saghatolislam, D. Reich-Erkelenz, S. K. Schaupp, E. C. Schulte, F. Senner, I.-G. Anghelescu, V. Arolt, B. T. Baune, U. Dannlowski, D. E. Dietrich, A. J. Fallgatter, C. Figge, M. Jäger, G. Juckel, C. Konrad, V. Nieratschker, J. Reimer, E. Reininghaus, M. Schmauß, C. Spitzer, M. von Hagen, J. Wiltfang, J. Zimmermann, A. Gryaznova, L. Flatau-Nagel, M. Reitt, M. Meyers, B. Emons, I. S. Haußleiter, F. U. Lang, T. Becker, M. E. Wigand, S. H. Witt, F. Degenhardt, A. J. Forstner, M. Rietschel, M. M. Nöthen, T. F. M. Andlauer, W. Rössler, S. Walitza, P. Falkai, T. G. Schulze, and E. Grünblatt. Polygenic risk scores across the extended psychosis spectrum. *Translational Psychiatry*, 11(1), Nov. 2021. doi:10.1038/s41398-021-01720-0.

[68] D. F. Stern. Signaling by ERBB4. *Reactome - a curated knowledgebase of biological pathways*, 69, June 2019. doi:10.3180/r-hsa-1236394.3. URL `https://doi.org/10.3180/r-hsa-1236394.3`.

[69] E. Uffelmann, Q. Q. Huang, N. S. Munung, J. de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1), aug 2021. doi:10.1038/s43586-021-00056-9.

[70] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 years of GWAS discovery: Biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, jul 2017. doi:10.1016/j.ajhg.2017.06.005.

[71] K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *The American Journal of Human Genetics*, 81(6): 1278–1283, Dec. 2007. doi:10.1086/522374.

[72] K. Wang, M. Li, and H. Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, Nov. 2010. doi:10.1038/nrg2884.

[73] L. Wang, P. Jia, R. D. Wolfinger, X. Chen, and Z. Zhao. Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics*, 98(1): 1–8, July 2011. doi:10.1016/j.ygeno.2011.04.006.

[74] B. Wendel, S. Papiol, T. F. M. Andlauer, J. Zimmermann, J. Wiltfang, C. Spitzer, F. Senner, E. C. Schulte, M. Schmauß, S. K. Schaupp, J. Repple, E. Reininghaus, J. Reimer, D. Reich-Erkelenz, N. Opel, I. Nenadić, S. Meinert, C. Konrad, F. Klöhn-Saghatolislam, T. Kircher, J. L. Kalman, G. Juckel, A. Jansen, M. Jäger, M. Heilbronner, M. von Hagen, K. Gade, C. Figge, A. J. Fallgatter, D. E. Dietrich, U. Dannlowski, A. L. Comes, M. Budde, B. T. Baune, V. Arolt, I.-G. Anghelescu, H. Anderson-Schmidt, K. Adorjan, P. Falkai, T. G. Schulze, H. Bickeböller, and U. Heilbronner. A genome-wide association study of the longitudinal course of executive functions. *Translational Psychiatry*, 11(1), June 2021. doi:10.1038/s41398-021-01510-8.

[75] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, July 2011. doi:10.1016/j.ajhg.2011.05.029.

[76] M. C. Wu, A. Maity, S. Lee, E. M. Simmons, Q. E. Harmon, X. Lin, S. M. Engel, J. J. Molldrem, and P. M. Armistead. Kernel machine SNP-set testing under multiple candidate kernels. *Genetic Epidemiology*, 37(3):267–275, Mar. 2013. doi:10.1002/gepi.21715.

[77] Q. Yan, D. E. Weeks, H. K. Tiwari, N. Yi, K. Zhang, G. Gao, W.-Y. Lin, X.-Y. Lou, W. Chen, and N. Liu. Rare-variant kernel machine test for longitudinal data from population and family samples. *Human Heredity*, 80(3):126–138, 2015. doi:10.1159/000445057.

[78] K.-H. Yuan and P. M. Bentler. Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology*, 63(2): 273–291, May 2010. doi:10.1348/000711009x449771.

[79] A. Ziegler and I. R. König. *A Statistical Approach to Genetic Epidemiology*. Wiley, Weinheim, Germany, second edition, Mar. 2010. ISBN 978-3-527-32389-0. doi:10.1002/9783527633654.

A. References of Original Work

A.1. Articles

Wendel B, Papiol S, Andlauer TFM, Zimmermann J, Wiltfang J, Spitzer C, Senner F, Schulte EC, Schmauß M, Schaupp SK, Repple J, Reininghaus E, Reimer J, Reich-Erkelenz D, Opel N, Nenadić I, Meinert S, Konrad C, Klöhn-Saghatolislam F, Kircher T, Kalman JL, Juckel G, Jansen A, Jäger M, Heilbronner M, von Hagen M, Gade K, Figge C, Fallgatter AJ, Dietrich DE, Dannlowski U, Comes AL, Budde M, Baune BT, Arolt V, Anghelescu IG, Anderson-Schmidt H, Adorjan K, Falkai P, Schulze TG, Bickeböller H, Heilbronner U:

A Genome-Wide Association Study of the Longitudinal Course of Executive Functions.

Translational Psychiatry (2021); 11(1)

URL: `https://www.nature.com/articles/s41398-021-01510-8`

DOI: 10.1038/s41398-021-01510-8

Wendel B, Heidenreich M, Budde M, Heilbronner M, Oraki Kohshour M, Papiol S, Falkai P, Schulze TG, Heilbronner U, Bickeböller H:

Kalpra: a kernel approach for longitudinal pathway regression analysis integrating network information with an application to the longitudinal PsyCourse Study.

Frontiers in Genetics (2022); 13

URL: `https://doi.org/10.3389/fgene.2022.1015885`

DOI: 10.3389/fgene.2022.1015885

A.2. Software

Wendel B, Heidenreich M, Bickeböller H

kalpra: Kernel Approach for Longitudinal Pathway Regression Analysis

URL: `https://gitlab.gwdg.de/bernadette.wendel/kalpra`

ARTICLE OPEN

Check for updates

A genome-wide association study of the longitudinal course of executive functions

Bernadette Wendel[1✉], Sergi Papiol[2,3], Till F. M. Andlauer[4], Jörg Zimmermann[5], Jens Wiltfang[6,7,8], Carsten Spitzer[9], Fanny Senner[2,3], Eva C. Schulte[2,3], Max Schmauß[10], Sabrina K. Schaupp[2], Jonathan Repple[11], Eva Reininghaus[12], Jens Reimer[13,14], Daniela Reich-Erkelenz[2], Nils Opel[11], Igor Nenadić[15,16], Susanne Meinert[11], Carsten Konrad[17], Farahnaz Klöhn-Saghatolislam[2,3], Tilo Kircher[15,16], Janos L. Kalman[2,3,18], Georg Juckel[19], Andreas Jansen[15,16,20], Markus Jäger[21], Maria Heilbronner[2], Martin von Hagen[22], Katrin Gade[6], Christian Figge[23], Andreas J. Fallgatter[24], Detlef E. Dietrich[25,26,27], Udo Dannlowski[11], Ashley L. Comes[2,18], Monika Budde[2], Bernhard T. Baune[28,29,30], Volker Arolt[11], Ion-George Anghelescu[31], Heike Anderson-Schmidt[2,6], Kristina Adorjan[2,3], Peter Falkai[3], Thomas G. Schulze[2,3], Heike Bickeböller[1] and Urs Heilbronner[2]

Executive functions are metacognitive capabilities that control and coordinate mental processes. In the transdiagnostic PsyCourse Study, comprising patients of the affective-to-psychotic spectrum and controls, we investigated the genetic basis of the time course of two core executive subfunctions: set-shifting (Trail Making Test, part B (TMT-B)) and updating (Verbal Digit Span backwards) in 1338 genotyped individuals. Time course was assessed with four measurement points, each 6 months apart. Compared to the initial assessment, executive performance improved across diagnostic groups. We performed a genome-wide association study to identify single nucleotide polymorphisms (SNPs) associated with performance change over time by testing for SNP-by-time interactions using linear mixed models. We identified nine genome-wide significant SNPs for TMT-B in strong linkage disequilibrium with each other on chromosome 5. These were associated with decreased performance on the continuous TMT-B score across time. Variant rs150547358 had the lowest P value $= 7.2 \times 10^{-10}$ with effect estimate beta = 1.16 (95% c.i.: 1.11, 1.22). Implementing data of the FOR2107 consortium (1795 individuals), we replicated these findings for the SNP rs150547358 (P value = 0.015), analyzing the difference of the two available measurement points two years apart. In the replication study, rs150547358 exhibited a similar effect estimate beta = 0.85 (95% c.i.: 0.74, 0.97). Our study demonstrates that longitudinally measured phenotypes have the potential to unmask novel associations, adding time as a dimension to the effects of genomics.

Translational Psychiatry (2021)11:386 ; https://doi.org/10.1038/s41398-021-01510-8

INTRODUCTION

The term "executive functions" (EFs) describes a group of higher-level cognitive abilities [1], including the regulation of thoughts and actions in daily life [1, 2]. As humans age, EFs pass different developmental stages, in which great variability is observed both within and between individuals [3, 4]. EFs naturally decline with

[1]Department of Genetic Epidemiology, University Medical Center Göttingen, Georg-August-University Göttingen, Göttingen 37073, Germany. [2]Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, Munich 80336, Germany. [3]Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich 80336, Germany. [4]Department of Neurology, University Hospital rechts der Isar, School of Medicine, Technical University of Munich, Munich 81675, Germany. [5]Psychiatrieverbund Oldenburger Land gGmbH, Karl-Jaspers-Klinik, Bad Zwischenahn 26160, Germany. [6]Department of Psychiatry and Psychotherapy, University Medical Center Göttingen, Göttingen 37075, Germany. [7]German Center for Neurodegenerative Diseases (DZNE), Göttingen 37075, Germany. [8]Neurosciences and Signaling Group, Institute of Biomedicine (iBiMED), Department of Medical Sciences, University of Aveiro, Aveiro 3810-193, Portugal. [9]Department of Psychosomatic Medicine and Psychotherapy, University Medical Center Rostock, Rostock 18147, Germany. [10]Department of Psychiatry and Psychotherapy, Bezirkskrankenhaus Augsburg, Augsburg 86156, Germany. [11]Institute for Translational Psychiatry, University of Münster, Münster 48149, Germany. [12]Department of Psychiatry and Psychotherapeutic Medicine, Research Unit for Bipolar Affective Disorder, Medical University of Graz, Graz 8036, Austria. [13]Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany. [14]Department of Psychiatry, Health North Hospital Group, Bremen 28102, Germany. [15]Department of Psychiatry and Psychotherapy, Philipps-University Marburg, Marburg 35039, Germany. [16]Centre for Mind, Brain and Behaviour, University of Marburg, Marburg 35032, Germany. [17]Department of Psychiatry and Psychotherapy, Agaplesion Diakonieklinikum, Rotenburg 27356, Germany. [18]International Max Planck Research School for Translational Psychiatry (IMPRS-TP), Max Planck Institute of Psychiatry, Munich 80804, Germany. [19]Department of Psychiatry, Ruhr University Bochum, LWL University Hospital, Bochum 44791, Germany. [20]Core-Facility Brainimaging, Faculty of Medicine, University of Marburg, Marburg, Germany. [21]Department of Psychiatry II, Ulm University, Bezirkskrankenhaus Günzburg, Günzburg 89312, Germany. [22]Clinic for Psychiatry and Psychotherapy, Clinical Center Werra-Meißner, Eschwege 37269, Germany. [23]Karl-Jaspers Clinic, European Medical School Oldenburg-Groningen, Oldenburg 26160, Germany. [24]Department of Psychiatry and Psychotherapy, University Tübingen, Tübingen 72076, Germany. [25]AMEOS Clinical Center Hildesheim, Hildesheim 31135, Germany. [26]Center for Systems Neuroscience (ZSN), Hannover 30559, Germany. [27]Department of Psychiatry, Medical School of Hannover, Hannover 30625, Germany. [28]Department of Psychiatry, University of Münster, Münster 48149, Germany. [29]Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Melbourne, Australia. [30]The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC, Australia. [31]Department of Psychiatry and Psychotherapy, Mental Health Institute Berlin, Berlin 14050, Germany. ✉email: bernadette.wendel@med.uni-goettingen.de

Received: 2 December 2020 Revised: 4 June 2021 Accepted: 15 June 2021
Published online: 10 July 2021

advanced age [4–6] in a gender-specific manner [7] and diminished EFs are also observed in the longitudinal course of severe mental disorders, such as schizophrenia [8]. In particular, EFs appear to be generally impaired in psychiatric patients suffering from schizophrenia, depression [4], or bipolar disorder [9]. Deficits are also associated, for example with decreased abilities to perform routine tasks [4]. Neurobiologically, EFs are linked intimately to the prefrontal cortex, as exemplified by the famous case of Phineas Gage [10].

There are many definitions of an EF [3], as it represents an umbrella term for multiple cognitive processes [2]. An influential theory of EFs is the "unity and diversity" concept [3, 11] that describes EFs as a "collection of related but separable abilities" [3]. EFs are differentiated into three latent core skills [3, 4, 11]: (i) set-shifting, allowing an individual to approach tasks flexibly and adjust to new conditions [3, 4], (ii) updating (or working memory), with respect to the monitoring, manipulating, and updating of information [4, 11], and (iii) inhibition, enabling an individual to control behavior, emotions, and responses [4, 11]. In general, EFs rank among the "most heritable psychological traits" [3]. On the behavioral genetic level, a highly heritable latent (common) factor affecting all EF aspects accounted for 99% of the variance common to all three skills [3]. Regarding specific EF components, the heritability estimates of set-shifting assessed by the Trail Making Test (TMT) range from 0.34 to 0.65 [12] and the estimates of updating measured by digit span tests range from 0.27 to 0.62 [12] (these results were obtained in twin studies). Recently, several genome-wide association studies (GWASs) on EFs have been undertaken [13–18]; however, genome-wide significance was not attained [2, 12]. Moreover, the genetic basis of variation over time is yet to be elucidated [19].

Here, we performed two longitudinal GWASs for the set-shifting and updating EF abilities assessed by the Trail Making Test, part B (TMT-B) and the Verbal Digit Span backwards (VDS-B), respectively, to identify genetic variation associated with the course of EFs across time. We used a linear mixed model (LMM) to model the dependence structure of the longitudinal PsyCourse Study [20] with four measurements across time. To validate our findings, we also performed a replication study using data from the FOR2107 consortium [21], which assessed two measurements over time.

MATERIALS AND METHODS

Discovery sample: PsyCourse Study

The PsyCourse Study is a multicenter longitudinal study that combines multilevel omics and longitudinal data [20]. We included 1338 genotyped individuals (dataset version 3.0) recruited in different centers in Germany and Austria, comprising patients from the affective-to-psychotic spectrum (377 bipolar I disorder, 100 bipolar II disorder, 420 schizophrenia, 95 schizoaffective disorder, 6 brief psychotic disorder, 9 schizophreniform disorder, and 73 with recurrent depression) and 258 psychiatrically healthy controls. The study protocol was approved by the respective ethics committee for each study center and was carried out following the rules of the Declaration of Helsinki of 1975, revised in 2008 (see ref. [20]). All study participants provided written consent [20]. The patients were diagnosed using parts of the Structured Clinical Interview for DSM (SCID-I) and were classified according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria. The patients were broadly differentiated in patients with predominantly affective symptoms (550 "affective", with recurrent depression, bipolar I and II disorders) and patients with predominantly psychotic symptoms (530, "psychotic", with schizophrenia, schizoaffective, brief psychotic and schizophreniform disorder) [20]. Deep phenotyping was performed during four visits, each ~6 months apart (see ref. [20]), thus corresponding to time t of the longitudinal course.

Set-shifting and updating were assessed with the Trail Making Test, part B (TMT-B) [22] and the Verbal Digit Span backwards (VDS-B) [23], respectively. The TMT-B requires an individual to connect numbers (numbers: 1–26) and letters of the alphabet in ascending alternating order. The test score was the time (in seconds (s)) needed to finish this exercise. As recommended by [24] participants with a time >300 s were set to 300 s. VDS-B measures the updating ability. Here, a trained interviewer verbally presented up to seven pairs of number sequences with increasing length, and the study participant was requested to repeat each sequence in backwards order, receiving a point score for each correctly repeated sequence. The maximum possible score for each sequence pair was 2. The process was terminated when an individual failed to repeat correctly both of the sequences in a pair of given length. The test score was the sum of all correctly repeated sequence pairs (range: 0–14).

Replication sample: FOR2107 consortium

To perform the replication study, we used data from the research consortium FOR2107 [21], a longitudinal cohort with two centers, Marburg and Münster (Germany), in which deep phenotyping was performed twice ~2 years apart [21]. In our analyses, we used a sample comprising 1795 individuals with genotype data available divided into five different diagnostic groups (851 affective: 107 bipolar disorder and 744 depression, 112 psychotic: 68 schizophrenia and 44 schizoaffective disorder, and 832 healthy controls). The participants were classified into the same three broad diagnostic groups (affective, psychotic, and controls) as in the discovery sample. Set-shifting was assessed by the TMT-B. In this cohort, participants with a time >180 s were excluded. For updating, we used the Letter–Number-Sequencing Test (LNST) as a substitute for the VDS-B. Here, a trained interviewer verbally presented an increasing sequence of letters and numbers, which the participant was requested to repeat, starting with the numbers in ascending order and ending with the letters in alphabetical order. The test was terminated when the individual repeated the same sequence incorrectly four times. The sum of the correctly repeated sequences was the test score, with a maximum of 24.

Genotyping and imputation

Discovery sample. The Illumina Infinium PsychArray (Illumina, USA) was used for genotyping purposes [20]. Genotypes were imputed with SHAPEIT2/IMPUTE2 using the 1000 Genomes Project Phase 3 data as a reference panel. Quality control (QC) was performed according to standard procedures, as described previously [25] (details Supplementary List 1) and poorly imputed genetic variants (INFO < 0.8) were excluded [20]. We included ~8.2 million SNPs with minor allele frequency (MAF) ≥ 0.01 in our analysis. Ancestry principal components (PCs) were computed with PLINK v1.9 [26] (http://pngu.mgh.harvard.edu/).

Replication sample. To replicate genome-wide significant SNPs of the discovery sample, we analyzed the genotypes of these nine significant SNPs (SNP_R). We additionally analyzed 187 suggestive SNPs (SNP_{NR}) with a P value $\leq 1 \times 10^{-5}$ in the discovery sample (99 for TMT-B, 88 for VDS-B/LNST) in an exploratory analysis. For the QC in the replication sample, please refer to Supplementary List 2.

Statistical analysis

We performed regression analysis, log-transforming the TMT-B values (lgTMT-B) to fulfill the linear mixed model requirement of normally distributed errors. We present effect estimates with 95% confidence intervals (c.i.s) transformed back to the original scale. Furthermore, we investigated missing data patterns across visits and diagnoses for violation of a missing-at-random (MAR) mechanism [27]. We computed the mean and standard deviation (s.d.) of EFs per visit and diagnostic group, testing for differences in means between diagnostic groups at each visit. For the discovery sample, we fitted LMMs to the longitudinal time course of lgTMT-B and VDS-B, investigating each phenotype first without the SNP terms, and subsequently including them. For each SNP, the fitted model for individual i at visit/time t_{ij} with $j = 1, 2, 3, 4$ was as follows:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 age_i + \beta_3 gender_i + \beta_4 diagnosis_i + \sum_{k=1}^{5} \beta_{4+k} PC_{ik} + b_{0i} + b_{1i} t_{ij} + c_i center_i + \beta_{10} SNP_i + \beta_{11} SNP_i * t_{ij} + \varepsilon_{ij}$$

The LMM adjusted for age_i, $gender_i$, $diagnosis_i$, PC_{ik}, i.e., age at visit 1, gender, diagnostic group (affective, psychotic, or control), and the top five PCs, for each individual i, the latter to correct for population stratification. We allowed for random intercepts and slopes b_{0i}, b_{1i} of the trajectories and a random center effect.

For the respective SNP under consideration, we integrated the main effect (SNP_i) and the SNP-by-time interaction ($SNP_i * t_{ij}$), where the latter is

Table 1. Characteristics at visit 1 in discovery sample and replication sample by diagnostic group.

Study sample	Phenotypes	Diagnostic groups mean (s.d.) or percentage (%)			Group difference
		Affective	Psychotic	Controls	*P* value
Discovery sample	Age	44.6 (13.4)	41.1 (12.1)	37.1 (15.6)	–
	Females	49.8 %	39.6 %	58.1 %	–
	TMT-B	83.9 (42.6)	92.3 (41.3)	59.4 (25.1)	$<2\times10^{-16}$
	VDS-B	6.2 (2.1)	5.5 (2.0)	7.3 (2.9)	$<2\times10^{-16}$
Replication sample	Age	37.6 (13.4)	38.4 (11.3)	34.1 (12.6)	–
	Females	63.9 %	44.6 %	63.0 %	–
	TMT-B	57.7 (23.9)	73.6 (30.9)	48.8 (18.6)	$<2\times10^{-16}$
	LNST	15.7 (3.3)	13.4 (3.5)	16.8 (3.2)	$<2\times10^{-16}$

The proportion of females (%), means of age (years), TMT-B, and VDS-B/LNST with standard deviation (s.d.).
We tested for differences in means between the diagnostic groups for lgTMT-B and VDS-B. Results are only displayed for visit 1 as results for the other visits proved to be similar.

tested (two-sided) for the influence of the SNP on the longitudinal course (see ref. [28]). The interaction term consisting of SNP × diagnosis × time has not been investigated due to the limited sample size. We assumed an additive genetic model with each considered SNP in dosage format. We set the genome-wide significance level to 5×10^{-8}, yielding replication SNPs (SNP_R), and set the level for suggestive significance to 1×10^{-5} for SNPs to be further explored (SNP_{NR}, not to be replicated). For the replication sample, we separately determined linkage disequilibrium (LD) blocks with $r^2 > 0.8$ for both SNP sets, correcting for multiple testing by dividing 5% by the number of LD blocks for the SNP set [29]. In the end, the SNP_R were contained in a single LD block, so the significance level for replication could be set to 5%. The significance levels for the exploratory analysis of the SNP_{NR} were set to $0.05/24 = 0.0021$ for lgTMT-B and $0.05/12 = 0.0042$ for VDS-B/LNST, respectively.

For the SNP analysis in the replication sample, we analyzed the difference (diff) of lgTMT-B (LNST) between the visits as outcome and SNP, age, gender, diagnosis, and PC's as covariates. We applied the difference model, as the LMM above contained too many parameters for the replication sample with only two measurements (in total: 613 individuals) and incomplete data resulting in low statistical power (data not shown; two-sided test). Here, the SNP effect may be interpreted as the difference between the average change between the genotypes, especially since SNP_R displayed only two genotypes.

We computed LD and haplotypes for Europeans with LDlink [30] and created a regional plot with gene identification using Locus-Zoom [31]. Finally, the average longitudinal course over time per genotype along with 95% c.i. is displayed for the top SNP.

All statistical analyses were performed with R, version 3.5.1 (https://www.r-project.org/). The LMM was fitted with the R package lme4 [32] and *P* values were computed using the Satterthwaite approximation of the lmerTest package [33, 34].

RESULTS

Behavioral characteristics of the EFs

Discovery sample. In comparison with controls, the disease groups were slightly older on average (Table 1). A total of 1272 (1297) individuals had at least one TMT-B (VDS-B) measurement, demonstrating a similar decrease of available data in each diagnostic group (Table 2). Missing value patterns did not hint at any violation of a missing-at-random (MAR) assumption (data not shown). Figure 1 illustrates the mean longitudinal course of TMT-B (left) and VDS-B (right) for each diagnostic group with 95% c.i.s; controls differed significantly from patients (see Fig. 1, c.i.s). Generally, executive performance increased over time, with differences between affective and psychotic patients decreasing over time. An improvement in the respective EF performance is reflected by a decreased TMT-B score for set-shifting and an increased VDS-B score for updating. The individual trajectories were highly variable (Supplementary Fig. 1). The mean difference between diagnostic groups was significant at each visit when adjusting for age and gender (see Table 1). Table 3 displays the time effect estimates in the LMM for each phenotype without SNP stratified by diagnostic group. For lgTMT-B, the time effect within each diagnostic group is highly significant and similar across groups. For VDS-B, the time effects for the two patient groups are similar, very small, and only nominally significant in the psychotic group, but larger and highly significant for controls.

Replication sample. We analyzed 1795 genotyped individuals with at least one TMT-B and LNST measurement (we deleted data for one individual who had a value larger than the maximum score of 24). Phenotypes were measured at both visits for 34.2%. The means of the diagnostic groups at each visit were significantly different (Table 1) during which the controls had again the best EF abilities, followed by affective and then psychotic individuals (Supplementary Fig. 2).

GWAS of the discovery sample

The QQ-plot (Supplementary Fig. 3) demonstrates that the genomic inflation factor was $\lambda = 1.0034$ for lgTMT-B and $\lambda = 0.9999$ for VDS-B, hence not indicating any inflation. As illustrated on the Manhattan plots (lgTMT-B Fig. 2A, VDS-B Fig. 2B) for the SNP-by-time interaction in the LMM, we identified nine genome-wide significant SNPs on chromosome 5 (all imputed) in one LD block ($r^2 > 0.85$) for lgTMT-B, and none for VDS-B. For lgTMT-B, 99 SNPs were suggestive, for VDS-B 88.

For the nine genome-wide significant SNPs of the GWAS, Supplementary Table 1 displays estimates for the effect of the SNP-by-time interaction on lgTMT with 95% c.i. and *P* values. The top SNP rs150547358 (P value $= 7.2\times10^{-10}$) had an effect of 1.16 (95% c.i. 1.11–1.22) seconds per measurement (spm) in the discovery sample on the original TMT-B scale. We present the mean plot for the top SNP in Fig. 2C, where the TMT-B score increases over time for heterozygotes with risk allele "C". Figure 2D displays the regional Manhattan plot with three genes in or near the nine significant SNPs. Four of them, including rs150547358, are located in an intron region of ring finger protein 180 (RNF180) (Supplementary Table 1). Other genes located nearby are regulator of G protein signaling 7 binding protein (RGS7BP) and 5-hydroxytryptamine receptor 1A (HTR1A), but neither contained any of the nine SNPs. For the SNP main effect, also included in the model, we did not observe any genome-wide significant SNPs (Supplementary Fig. 4; $P < 5\times10^{-8}$).

Difference analysis of the replication sample

The analysis of the differences also identified the top SNP, rs150547358, as significant ($P = 0.015$), and thus replicated this GWAS-significant LD block. The effect estimate for the top SNP

Table 2. Available data of TMT-B and VDS-B per visit for the discovery sample.

EF core skill	Diagnostic groups											
	Affective				Psychotic				Controls			
Visit (t)	1	2	3	4	1	2	3	4	1	2	3	4
TMT-B	506 (92%)	315 (57%)	234 (43%)	182 (33%)	456 (86%)	295 (56%)	252 (46%)	227 (48%)	258 (100%)	225 (82%)	178 (69%)	57 (22%)
VDS-B	503 (92%)	324 (59%)	234 (43%)	185 (34%)	479 (90%)	320 (60%)	265 (50%)	236 (45%)	257 (99.6%)	225 (87%)	178 (69%)	60 (23%)

Absolute numbers and percent of group total within the diagnostic group with 550 affective individuals, 530 psychotic individuals, and 258 controls.

was 0.85 (95% c.i. 0.74–0.97) on the original scale and the highest effect size in the scale of the analysis (greatest negative effect). The estimates for the other SNPs were slightly larger when transformed back to the original scale and also positive (see Supplementary Table 1 for the summary).

Exploratory analysis of the GWAS-suggestive SNP_{NR} in the replication sample yielded no significant results after multiple testing corrections for either phenotype (Supplementary Fig. 5).

DISCUSSION

We performed a GWAS on the longitudinal course of EFs and detected nine SNPs within the same LD block associated with change over a relatively short period of time (~1.5 years) in the EF core skill set-shifting. Importantly, we were able to replicate a significant result for this LD block in an independent sample, which was observed in a heterogeneous population including controls and different psychiatric disorders of the affective-to-psychotic spectrum across age groups. Analysis of TMT-B performance of C-allele carriers, in contrast to the AA genotype, revealed a pronounced slowing over time.

Recently, the analysis of longitudinal data has come to the fore in genetic research. Multiple methods have been developed to perform GWAS with longitudinal data [35–40] for binary as well as continuous phenotypes. These analysis methods are mostly applied to analyze long-term developments of the investigated phenotypes [41, 42], as most data comprise multiple measurements over a relatively long period of time. These longitudinal studies often detect group effects [8] based on age or baseline cognitive functions, for example. To date, short-term variability, for example with respect to the longitudinal course of schizophrenia has been found as reviewed [8], but without considering a potential genetic effect. In our longitudinal GWAS, we enter uncharted territory as we study short-term courses of cognitive phenotypes in relation to the genetic background. The discovery sample, the PsyCourse Study, is unique in this sense, as it assesses the phenotypes multiple times in a very heterogeneous sample over a relatively short period of time (18 months). Here, the main interest is the observation of short-term changes specific to a phenotype, such as EF skills, and the use of newly identified characteristics to detect genotype–phenotype associations. The genetic variants found in this study may, if further replicated, be used to improve clinical evaluation of the longitudinal course of EF skills. Knowledge of the genetic status of a patient may, in the future, enhance the interpretation of the course of EF abilities e.g., during psychiatric treatment. Moreover, special training programs could support patients with a known genetic disposition to lack improvement over time. To our knowledge, no other study has performed such analyses to date.

Behavioral results

Prior to our GWAS, we studied the short-term courses of changes in cognitive abilities, focusing on the differences between the diagnostic groups considered. In the discovery sample, we observed an identical pattern for both phenotypes: psychotic individuals demonstrated the lowest EF abilities, followed by those with affective disorders and then the control individuals. This greater EF impairment in psychotic individuals compared to controls is well-documented, as exemplified by [43]. However, regarding the impairment difference between bipolar (affective) and schizophrenic (psychotic) patients, there are various studies [43–48] analyzing these differences. The hypothesis exists that bipolar patients demonstrate less severe impairment in comparison to schizophrenic patients [49]. Some studies [44, 46, 48] lend their support to this hypothesis, though not always statistically significant, whereas others detected similar levels of impairment in symptomatic patients [45, 47]. In our analysis, we observed a statistically significant difference between affective and psychotic

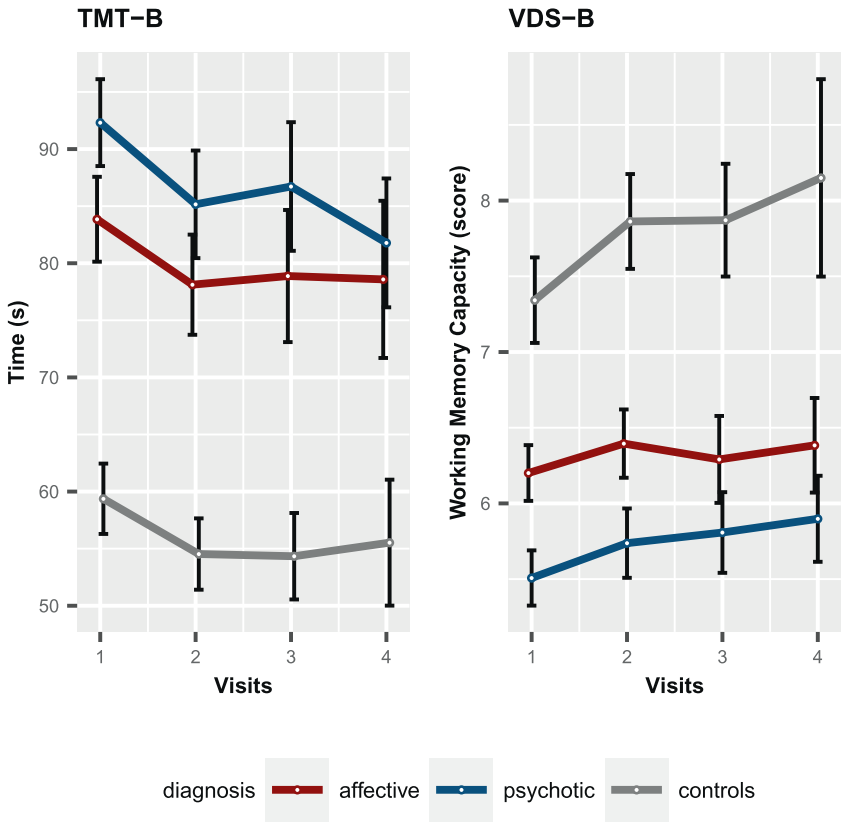


Fig. 1 Longitudinal course of TMT-B score (time in seconds, left) and VDS-B score (working memory capacity, right) for each diagnostic group in the discovery sample. Displayed are means with 95% confidence interval for each visit 1, 2, 3, 4, ~6 months apart.

Table 3. Results of the LMM of the discovery sample to test the time effect on lgTMT-B and VDS-B within each diagnostic group.

EF core skill		TMT-B			VDS-B		
Diagnostic groups	**Time effect (t)**	***β***	**95% c.i.**	***P* value**	***β***	**95% c.i.**	***P* value**
Affective		0.957	0.94, 0.97	9.8×10^{-09}	0.076	0, 0.15	0.053
Psychotic		0.950	0.94, 0.96	$<2 \times 10^{-16}$	0.086	0.02, 0.15	0.011
Controls		0.947	0.93, 0.96	6.1×10^{-11}	0.288	0.17, 0.41	2.7×10^{-06}

The effect estimates β of lgTMT-B are transformed back to their original scale.

individuals at visit 1 but detected a decline in these discrepancies over time. The abilities of these two diagnostic groups converged with patients from the psychotic group displaying an improvement in their skills and patients from the affective group presenting a more constant course. Documentation of the EF convergence is only possible thanks to the longitudinal design of the discovery sample and represents a great advantage of this study design.

Owing to the slightly different age structure of the two study samples, with the discovery sample being minimally older on average at visit 1, we further observed the impact of age reflected by the minimally lower average test score. That is, the discovery sample had lower VDS and greater TMT-B scores than the replication sample. The TMT-B mean scores may also be influenced further by the different cutoff thresholds of 300 s in the discovery sample and 180 s for the replication sample.

Genome-wide association studies

To our knowledge, the LD block comprising the nine SNPs we detected for the set-shifting ability has been not identified in any GWAS before. These SNPs are part of two common haplotypes, that is, 97.7% carry the haplotype consisting of the major alleles and 1.7% have the rare haplotype with only minor alleles in European populations [30]. However, we did not observe different allelic distributions between the three diagnostic groups (Supplementary Table 2). We displayed the longitudinal course for the two genotypes "AC" and "AA" of the top SNP rs150547358, observing a steady increase in the TMT-B score for "AC" and an almost unchanging course for "AA". Consequently, the minor allele C was associated with a decline in the set-shifting ability of ~5 s over a period of 18 months for AC with a large c.i. at the last visit owing to the small number of available heterozygous individuals. This result reflects a relatively high decrease in the ability over this short period. Furthermore, it portrays a highly interesting observation, which is further underpinned when we consider the genetic region of the nine SNPs. Variant rs150547358, the significantly replicated SNP, is one of four associated SNPs directly located in the ring finger protein 180 (RNF180) gene on chromosome 5q12.3. It is an E3 ubiquitin-protein ligase [50], whose product is involved in protein modification. RNF180 is

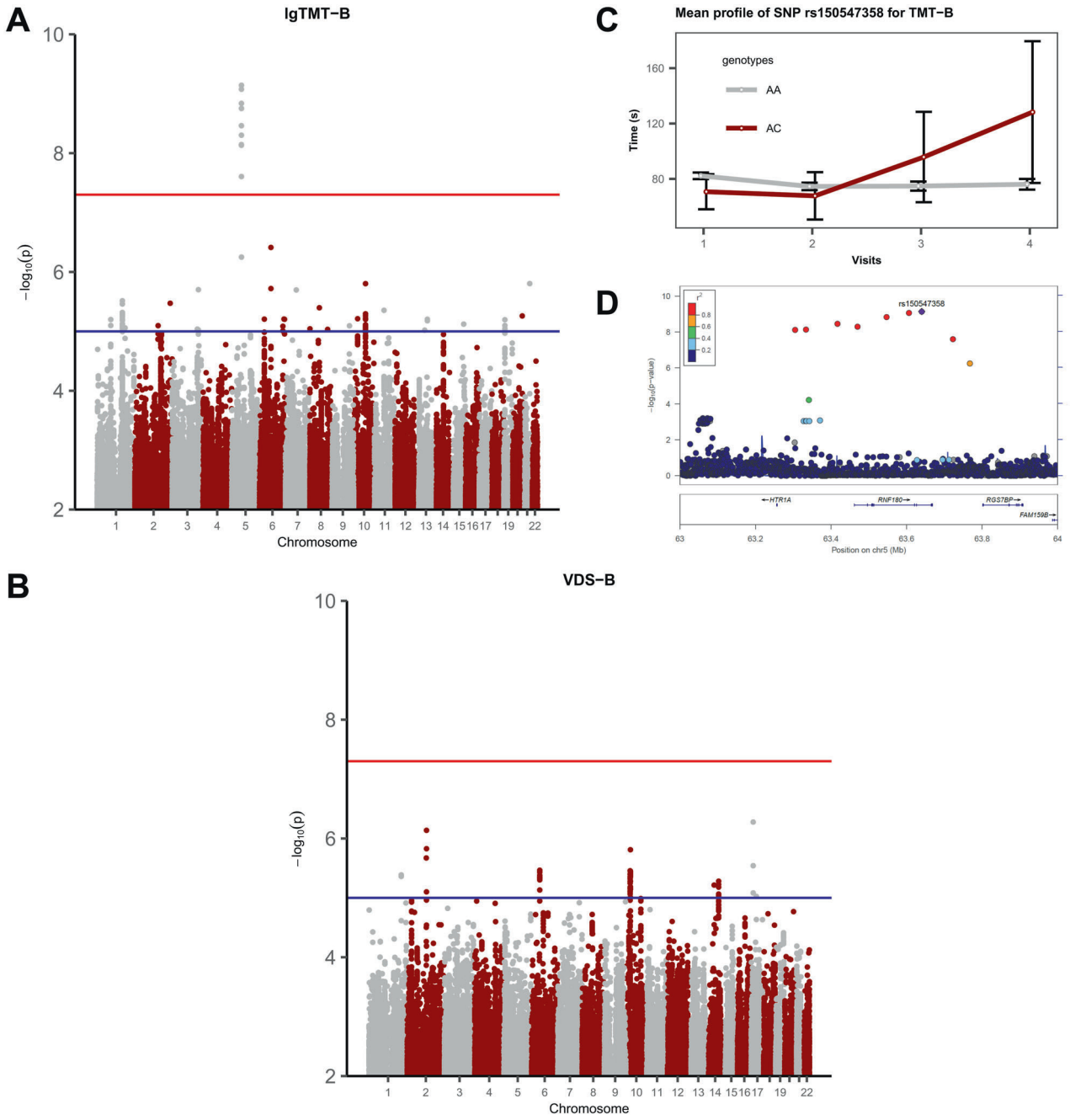


Fig. 2 Results of the genome-wide association studies of the discovery sample. A Manhattan plot of the GWAS of lgTMT-B in the discovery sample. The lines in (**A**) and (**B**) indicate the thresholds for the genome-wide significance of 5×10^{-8} (red) and for suggestive SNPs (blue, $P \leq 1 \times 10^{-5}$). **B** Manhattan plot of the GWAS of VDS-B in the discovery sample. **C** Mean profile of TMT-B by the top SNP rs150547358 genotypes for the discovery sample (1039 AA, 28 AC, 0 CC) with the 95% confidence intervals. **D** GWAS regional Manhattan plot of chromosome 5 for lgTMT-B of the discovery sample. Colors indicate the LD values (r^2) of SNPs with rs150547358 (in purple).

associated with the regulation of monoamine levels in different brain regions, for example, the prefrontal cortex (PFC) in RNF180 knockout mice [51]. The PFC is a critical part of the frontal lobe in the development of EFs [4, 52]. Another gene located in the nearby region, HTR1A (5-hydroxytryptamine receptor 1A), is an important receptor of serotonin (5-HT) also essential to the prefrontal lobe. More importantly, HTR1A is an autoreceptor, located on the cell bodies of serotonin-synthesizing neurons of the brainstem dorsal raphe nucleus, helping to maintain homeostasis in serotonergic function [53]. Furthermore, a genetic polymorphism in the 5-HT system has previously been implicated in EF performance [12].

In an additional exploratory gene-set analysis performed with MAGMA v1.06 as a part of the FUMA pipeline (https://fuma.ctglab.nl/) [54], we did not receive significant (Bonferroni-corrected P values ≤0.05) pathways for either phenotype.

Our results are a first step in the direction of understanding the molecular genetic influences on the longitudinal course of EFs. We were unable to consider the third core ability, inhibition, which also plays an important role for EF, because we could not fulfill a specific assessment requirement resulting from the multicenter and interview-based structure of the discovery sample [20]. Many unknown factors remain, such as the genetic aspects due to the correlation of the different EF abilities, as we only concentrated on

individual EF core skills in two separate analyses. According to the "unity but diversity" concept [11] that also concerns the genetic underpinnings of the EFs, a genetic study of a latent common factor needs to follow. Further, we need to acknowledge the problem of missing data which is a great challenge in longitudinal studies as presented in our samples. Here, selecting the correct analysis method, e.g., linear mixed models are imported but generally, more longitudinal studies with multiple time points and greater sample sizes will be required to unmask further time and genomics interactions [19].

CODE AND DATA AVAILABILITY

R code and data will be available upon reasonable request by the authors. The summary statistics of our analysis will be published in the GWAS Catalog (https://www.ebi.ac.uk/gwas/).

REFERENCES

1. Friedman NP, Miyake A, Altamirano LJ, Corley RP, Young SE, Rhea SA, et al. Stability and change in executive function abilities from late adolescence to early adulthood: a longitudinal twin study. Developmental Psychol. 2016;52:326–40.
2. Barnes JJM, Dean AJ, Nandam LS, O'Connell RG, Bellgrov MA. The molecular genetics of executive function: role of monoamine system genes. Biol Psychiatry. 2011;69:e127–e143.
3. Friedman NP, Miyake A, Young SE, DeFries JC, Corley RP, Hewitt JK. Individual differences in executive functions are almost entirely genetic in origin. J Exp Psychol: General. 2008;137:201–25.
4. Diamond A. Executive functions. Annu Rev Psychol. 2013;64:135–68.
5. Best JR, Miller PH, Jones LL. Executive functions after age 5: changes and correlates. Developmental Rev. 2009;29:180–200.
6. West R. Aging and the neural correlates of executive function. In: Wiebe SA, Karbach JExecutive function. New York: Routledge; 2017. p. 91–105.
7. van Hooren SA, Valentijn AM, Bosma H, Ponds RW, van Boxtel MP, Jolles J. Cognitive functioning in healthy older adults aged 64-81: a cohort study into the effects of age, sex, and education. Aging Neuropsychol Cognition. 2007;24:40–54.
8. Heilbronner U, Samara M, Leucht S, Falkai P, Schulze TG. The longitudinal course of schizophrenia across the lifespan. Harv Rev Psychiatry. 2016;24:118–28.
9. Martínez-Arán A, Vieta E, Colom F, Torrent C, Sánchez-Moreno J, Reinares M, et al. Cognitive impairment in euthymic bipolar patients: implications for clinical and functional outcome. Bipolar Disord. 2004;6:224–32.
10. Ratiu P, Talos IF. The tale of phineas gage, digitally remastered. N Engl J Med. 2004;351:e21.
11. Miyake A, Friedman NP, Emerson MJ, Witzki AH, Howerter A, Wager TD. The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: a latent variable analysis. Cogn Psychol. 2000;41:49–100.
12. Li JJ, Roberts DK. Genetic influences on executive functions across the life span. In: Wiebe SA, Karbach J, (eds.) Executive function. New York: Routledge; 2017. p. 106–23.
13. Luciano M, Hansell NK, Lahti J, Davies G, Medland SE, Räikkönen K, et al. Whole genome association scan for genetic polymorphisms influencing information processing speed. Biol Psychol. 2011;86:193–202.
14. Seshadri S, DeStefano AL, Au R, Massaro JM, Beiser AS, Kelly-Hayes M, et al. Genetic correlates of brain aging on MRI and cognitive test measures: a genome-wide association and linkage analysis in the Framingham study. BMC Med Genet. 2007;8:1–14.
15. Cirulli ET, Kasperaviciūte D, Attix DK, Need AC, Ge D, Gibson G, et al. Common genetic variation and performance on standardized cognitive tests. Eur J Hum Genet. 2010;18:815–20.
16. Need AC, Attix DK, McEvoy JM, Cirulli ET, Linney KL, Hunt P, et al. A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTB. Hum Mol Genet. 2009;18:4650–61.
17. Malone SM, Vaidyanathan U, Basu S, Miller MB, McGue M, Iacono WG. Heritability and molecular-genetic basis of the P3 event-related brain potential: a genome-wide association study. Psychophysiology. 2014;51:1246–58.
18. LeBlanc M, Kulle B, Sundet K, Agartz I, Melle I, Djurovic S, et al. Genome-wide study identifies PTPRO and WDR72 and FOXQ1-SUMO1P1 interaction associated with neurocognitive function. J Psychiatr Res. 2012;46:271–8.
19. Boyce WT, Sokolowski MB, Robinson GE. Genes and environments, development and time. Proc Natl Acad Sci USA. 2020;117:23235–41.
20. Budde M, Anderson-Schmidt H, Gade K, Reich-Erkelenz D, Adorjan K, Kalman JL, et al. A longitudinal approach to biological psychiatric research: the PsyCourse study. Am J Med Genet Part B: Neuropsychiatr Genet. 2018;180:89–102.
21. Kircher T, Wöhr M, Nenadic I, Schwarting R, Schratt G, Alferink J, et al. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. Eur Arch Psychiatry Clin Neurosci. 2018;269:949–62.
22. Bowie CR, Harvey PD. Administration and interpretation of the trail making test. Nat Protoc. 2006;1:2277–81.
23. Hilbert S, Nakagawa TT, Puci P, Zech A, Bühner M. The digit span backwards task. Eur J Psychologcal Assess. 2015;31:174–80.
24. Strauss E, Sherman EMS, Spreen O. A compendium of neuropsychological tests—administration, norms, and commentary. New York: Oxford University Press; 2006.
25. Andlauer TF, Buck D, Antony G, Bayas A, Bechmann L, Berthele A, et al. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. Sci Adv. 2016;2:e1501678.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
27. Molenbergh G, Verbeke, G. Linear mixed models for longitudinal data. Berlin, Heidelberg: Springer; 2000.
28. Sikorska K, Rivadeneira F, Groenen PJF, Hofman A, Uitterlinden AG, Eilers PHC, et al. Fast linear mixed model computations for genome-wide association studies with longitudinal data. Stat. Med. 2012 ;32:165–80.
29. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics. 2008;9:516.
30. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31:3555–7.
31. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:2336–7.
32. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1.
33. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. J Stat Softw. 2017;82:1–26.
34. Luke SG. Evaluating significance in linear mixed-effects models in R. Behav Res Methods. 2016;49:1494–502.
35. Sikorska K, Lesaffre E, Groenen PJF, Rivadeneira F, Eilers PHC. Genome-wide analysis of large-scale longitudinal outcomes using penalization GALLOP algorithm. Sci Rep. 2018;8:1–8.
36. Sikorska K, Montazeri NM, Uitterlinden A, Rivadeneira F, Eilers PH, Lesaffre E. GWAS with longitudinal phenotypes: performance of approximate procedures. Eur J Hum Genet. 2015;23:1384–91.
37. Wu W, Wang Z, Xu K, Zhang X, Amei A, Gelernter J, et al. Retrospective association analysis of longitudinal binary traits identifies important loci and pathways in cocaine use. Genetics. 2019;213:1225–36.
38. Rudra P, Broadaway KA, Ware EB, Jhun MA, Bielak LF, Zhao W, et al. Testing cross-phenotype effects of rare variants in longitudinal studies of complex traits. Genet Epidemiol. 2018;41:320–32.
39. Ning C, Wang D, Zhou L, Wei J, Liu Y, Kang H, et al. Efficient multivariate analysis algorithms for longitudinal genome-wide association studies. Bioinformatics. 2019;35:4879–85.
40. Lee Y, Park S, Moon S, Lee J, Elston RC, Lee W, et al. On the analysis of a repeated measure design in genome-wide association analysis. Int J Environ Res Public Health. 2014;11:12283–303.
41. Adkins DE, Clark SL, Copeland WE, Kennedy M, Conway K, Angold A, et al. Genome-wide meta-analysis of longitudinal alcohol consumption across youth and early adulthood. Twin Res Hum Genet. 2015;18:335–47.
42. Tang W, Kowgier M, Loth DW, Soler Artigas M, Joubert BR, Hodge E, et al. Large-scale genome-wide association studies and meta-analyses of longitudinal change in adult lung function. PLoS ONE. 2014;9:e100776.
43. Wobrock T, Ecker UK, Scherk H, Schneider-Axmann T, Falkai P, Gruber O. Cognitive impairment of executive function as a core symptom of schizophrenia. World J Biol Psychiatry. 2009;10:442–51.
44. Szoke A, Meary A, Trandafir A, Bellivier F, Roy I, Schurhoff F, et al. Executive deficits in psychotic and bipolar disorders - Implications for our understanding of schizoaffective disorder. Eur Psychiatry. 2008;23:20–25.
45. Amann B, Gomar JJ, Ortiz-Gil J, McKenna P, Sans-Sansa B, Sarró S, et al. Executive dysfunction and memory impairment in schizoaffective disorder: a comparison with bipolar disorder, schizophrenia and healthy controls. Psychological Med. 2012;42:2127–35.
46. Hill SK, Reilly JL, Keefe RS, Gold JM, Bishop JR, Gershon ES, et al. Neuropsychological Impairments in schizophrenia and psychotic bipolar disorder: findings from the bipolar-schizophrenia network on intermediate phenotypes (B-SNIP) study. Am J Psychiatry. 2013;170:1275–84.

47. Lewandowski KE, Cohen BM, Keshavan MS, Öngür D. Relationship of neurocognitive deficits to diagnosis and symptoms across affective and non-affective psychoses. Schizophrenia Res. 2011;133:212–7.
48. Reichenberg A, Harvey PD, Bowie CR, Mojtabai R, Rabinowitz J, Heaton RK, et al. Neuropsychological function and dysfunction in schizophrenia and psychotic affective disorders. Schizophrenia Bull. 2008;35:1022–9.
49. Lynham AJ, Hubbard L, Tansey KE, Hamshere ML, Legge SE, Owen MJ, et al. Examining cognition across the bipolar/schizophrenia diagnostic spectrum. J Psychiatry Neurosci. 2018;43:245–53.
50. Ogawa M, Mizugishi K, Ishiguro A, Koyabu Y, Imai Y, Takahashi R, et al. Rines/RNF180, a novel RING finger gene-encoded product, is a membrane-bound ubiquitin ligase. Genes Cells. 2008;13:397–409.
51. Kabayama M, Sakoori K, Yamada K, Ornthanalai VG, Ota M, Morimura N, et al. Rines E3 ubiquitin ligase regulates MAO-A levels and emotional responses. J Neurosci. 2013;33:12940–53.
52. Best JR, Miller PH. A developmental perspective on executive function. Child Dev. 2010;81:1641–60.
53. McDevitt RA, Neumaier F. Regulation of dorsal raphe nucleus function by serotonin autoreceptors: a behavioral perspective. J Chem Neuroanat. 2011;41:234–46.
54. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. 2017;8:1–11.

ACKNOWLEDGEMENTS

Thomas G. Schulze and Peter Falkai are supported by the Deutsche Forschungsgemeinschaft (German Research Foundation; DFG) within the framework of the projects http://www.kfo241.de and http://www.PsyCourse.de (SCHU 1603/4-1, 5-1, 7-1; FA241/16-1). Bernadette Wendel and Heike Bickeböller are supported by the DFG (KFO241, BI 576 15-1). Tilo Kircher was also supported by the DFG (FOR2107 KI588/14-1 and FOR2107 KI588/14-2). The genotyping was funded in part by the German Federal Ministry of Education and Research (BMBF) through the Integrated Network IntegraMent (Integrated Understanding of Causes and Mechanisms in Mental Disorders), under the auspices of the e:Med Program with a grant awarded to Thomas G. Schulze (01ZX1614K). Thomas G. Schulze received additional support from the German Federal Ministry of Education and Research (BMBF) within the framework of the BipoLife network (01EE1404H) and the Dr. Lisa Oehler Foundation (Kassel, Germany). Sergi Papiol was supported by a 2016 NARSAD Young Investigator Grant (25015) from the Brain and Behavior Research Foundation. This work was further funded by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to UD) and the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to UD). Igor Nenadić was supported by the Deutsche Forschungsgemeinschaft (DFG, grants NE2254/1-2, NE2254/2-1, NE2254/3-1, NE2254/4-1). Jens Wiltfang is supported by an Ilídio Pinho professorship, iBiMED (UIDB/04501/2020) at the University of Aveiro, Portugal. Andreas Jansen was supported by the DFG (FOR2107 JA-1890/7-1 and FOR2108 JA-1890/7-2). We acknowledge support by the Open Access Publication Funds of the Göttingen University. The authors would like to thank Frederike Stein for her support in sharing the replication data, and Andrew Entwistle for proofreading the manuscript.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-021-01510-8.

Correspondence and requests for materials should be addressed to B.W.

Reprints and permission information is available at http://www.nature.com/reprints

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

Supplementary information

Supplementary List 1.

For the **PsyCourse** study the following quality control steps were performed.
Sequence of Quality Control steps:

1. Removal of SNPs with call rates **<98%** or a MAF **<1%**
2. Removal of individuals with genotyping rates **<98%**
3. Removal of gender mismatches
4. Removal of genetic duplicates
5. Removal of cryptic relatives with $\widehat{\pi}$ **≥ 12.5**
6. Removal of genetic outliers with a distance from the mean of >**4** SD in the **first eight** MDS components
7. Removal of individuals with a deviation of the autosomal or X-chromosomal heterozygosity from the mean >**4** SD
8. Removal of non-autosomal variants
9. Removal of SNPs with call rates **<98%** or a MAF **<1%** or Hardy-Weinberg Equilibrium (HWE) test p-values **<1×10-6**
10. Removal of A/T and G/C SNPs
11. Update of variant IDs and positions to the IDs and positions in the 1000 Genomes Phase 3 reference panel
12. Alignment of alleles to the reference panel
13. Removal of duplicated variants and variants not present in the reference panel

Imputation was conducted using SHAPEIT2 (https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html) (1) and IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) (2,3) using the 1000 Genomes Phase 3 as reference panel. Genetic marker with a poor imputation quality (INFO < 0.8) were excluded (4).

Supplementary List 2.

FOR2107

Genotyping, quality control, and imputation
Genotyping was conducted using the Infinium PsychArray BeadChip, as described previously (5). The quality control (QC) of genetic data was conducted in PLINK v1.90b6.10 (6) and R v3.5.2, as described previously (7). Pre-imputation QC of genotype data consisted of the following steps:

1. Removal of SNPs with call rates <98% or a minor allele frequency (MAF) <1%
2. Removal of individuals with genotyping rates <98%
3. Removal of sex mismatches
4. Removal of genetic duplicates
5. Removal of cryptic relatives with pi-hat≥12.5
6. Removal of genetic outliers with a distance from the mean of >4 SD in the first eight multidimensional scaling (MDS) ancestry components
7. Removal of individuals with a deviation of the autosomal or X-chromosomal heterozygosity from the mean >4 SD
8. Removal of non-autosomal variants
9. Removal of SNPs with call rates <98% or a MAF <1% or Hardy-Weinberg Equilibrium (HWE) test *p*-values $<1\times10^{-6}$
10. Removal of A/T and G/C SNPs
11. Update of variant IDs and positions to the IDs and positions in the 1000 Genomes Phase 3 reference panel
12. Alignment of alleles to the reference panel
13. Removal of duplicated variants and variants not present in the reference panel

For the calculation of ancestry components (used to determine genetic outliers and as covariates in the analyses), pre-imputation genotype data were used. Additional variant filtering steps were removal of variants with a MAF <0.05 or HWE *p*-value $<10^{-3}$; removal of variants mapping to the extended MHC region (chromosome 6, 25-35 Mbp) or to a typical inversion site on chromosome 8 (7-13 Mbp); linkage disequilibrium (LD) pruning (command --indep-pairwise 200 100 0.2). Next, the pairwise identity-by-state (IBS) matrix of all individuals was calculated using the command --genome on the filtered genotype data. Multidimensional scaling (MDS) analysis was performed on the IBS matrix using the eigendecomposition-based algorithm in PLINK v1.90b6.10.

After imputation, variants with a MAF <1%, an HWE test $p<1\times10^{-6}$, and an INFO metric <0.8 were removed. Imputation was conducted using SHAPEIT v2 (r837) (1), IMPUTE2 v2.3.2 (2,3), and the 1000 Genomes Phase 3 reference panel.

In total, imputed genetic data were available for 2,248 individuals.

Variants before QC: 596,861; variants after QC: 284,691; variants after imputation: 8,565,143.

References

1. Delaneau, O., Zagury, J.F., Machini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nature Methods. 2013 Jan; 10(1): 5-6.

2. Howie, B.N., Donnelly, P., Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genetics. 2009 Jun; 5(6): e1000529.

3. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nature Genetics. 2012 Jul; 44(8): 955-959.

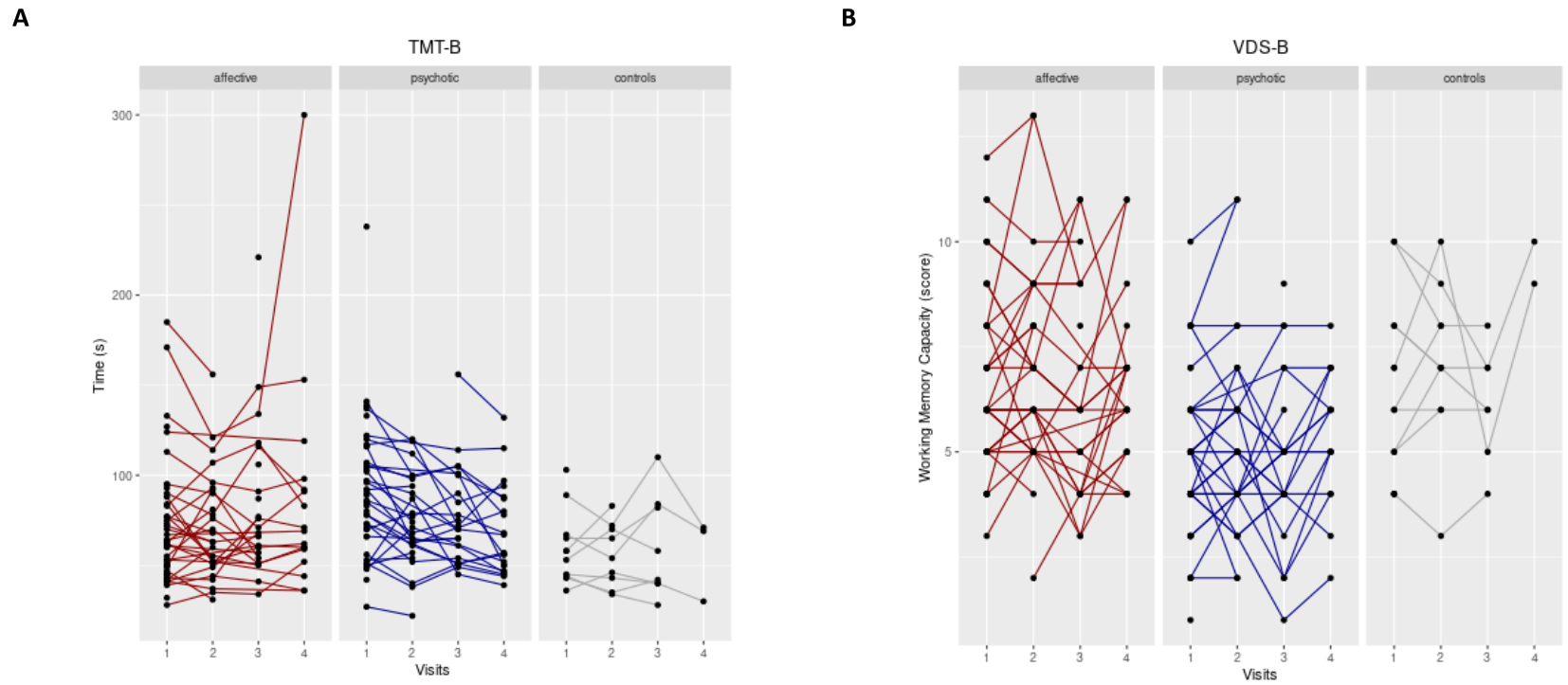
4. Budde, M. et al. A Longitudinal Approach to Biological Psychiatric Research: The PsyCourse Study. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2018 Aug; 180(2): 89-102.

5. Meller, T. et al. Associations of schizophrenia risk genes ZNF804A and CACNA1C with schizotypy and modulation of attention in healthy subjects. Schizophrenia Research. 2019 Jun; 208: 67-75.

6. Chang, C.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015 Feb; 4(1).

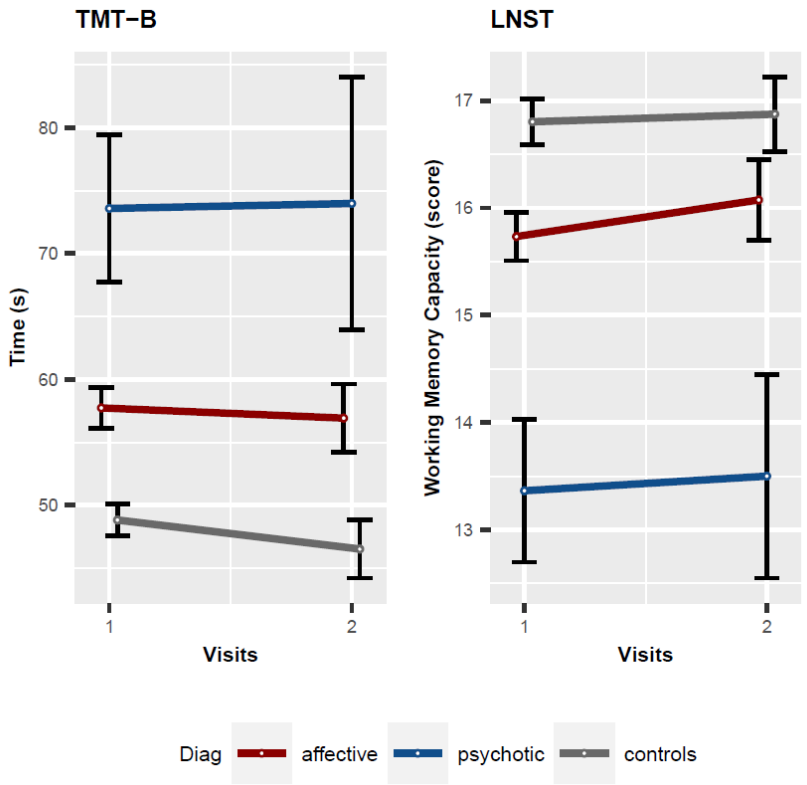
7. Andlauer, T.F.M. et al. Novel multiple sclerosis susceptibility loci implicated in epigenetic regulation. Science Advances. 2016 Jun; 2(6): e1501678.

Supplementary Figure 1



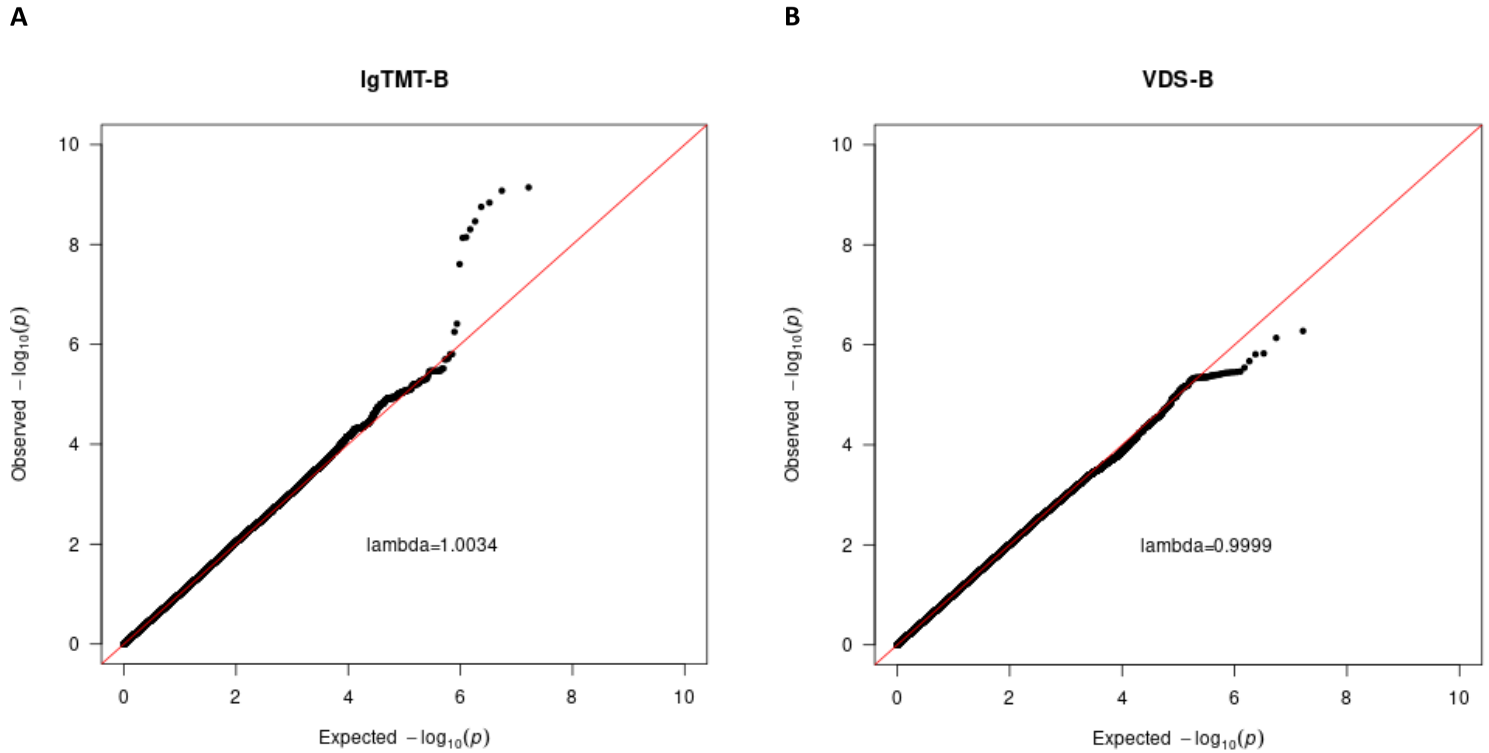
Supplementary Figure 1. Spaghetti plot of the longitudinal course of approximately 320 randomly selected individuals of (A) TMT-B score (time in seconds) and (B) VDS-B score (working memory capacity) of the discovery sample. The trajectories were separated for each diagnostic group.

Supplementary Figure 2



Supplementary Figure 2. Longitudinal course of TMT-B score (time in seconds, left) and LNST (working memory capacity, right) for each diagnostic group in the replication sample. Displayed are means with 95% confidence intervals for both visits 1 and 2, two years apart.

Supplementary Figure 3



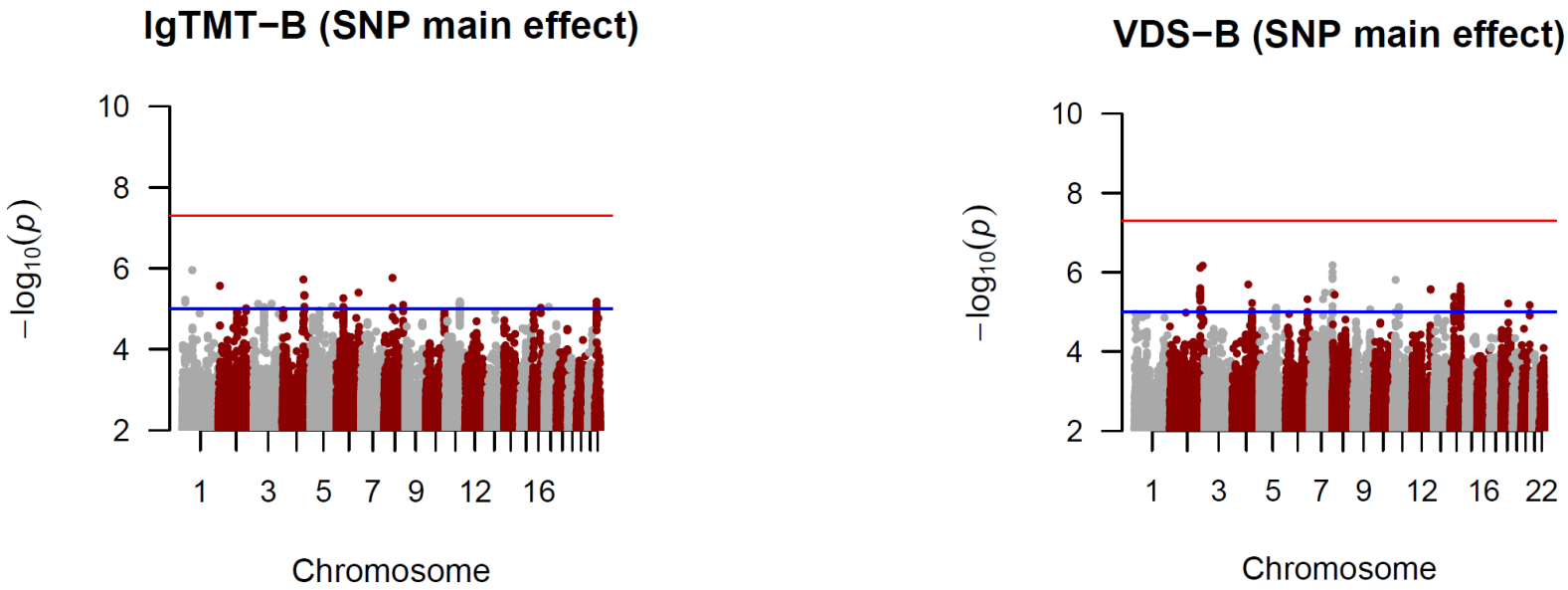
Supplementary Figure 3. QQ plot of the GWAS of (A) lgTMT-B and (B) VDS-B for the discovery sample with the genomic inflation factor lambda.

Supplementary Table 1

Supplementary Table 1. Overview of the genome-wide significant SNPs of the GWAS of the lgTMT-B with the effect estimates $\widehat{\beta}$ (95% c.i.s) in original scale for SNP-by-time interaction terms. Results are given for the GWAS (LMM) in the discovery sample (DS) and for the difference analysis in the replication sample (RS). Replicated SNP ($p_{RS} < 0.05$) is in bold. The SNPs are ordered according to their location on chromosome 5, the gene context and the distance to the next gene was received with FUMA (https://fuma.ctglab.nl/).

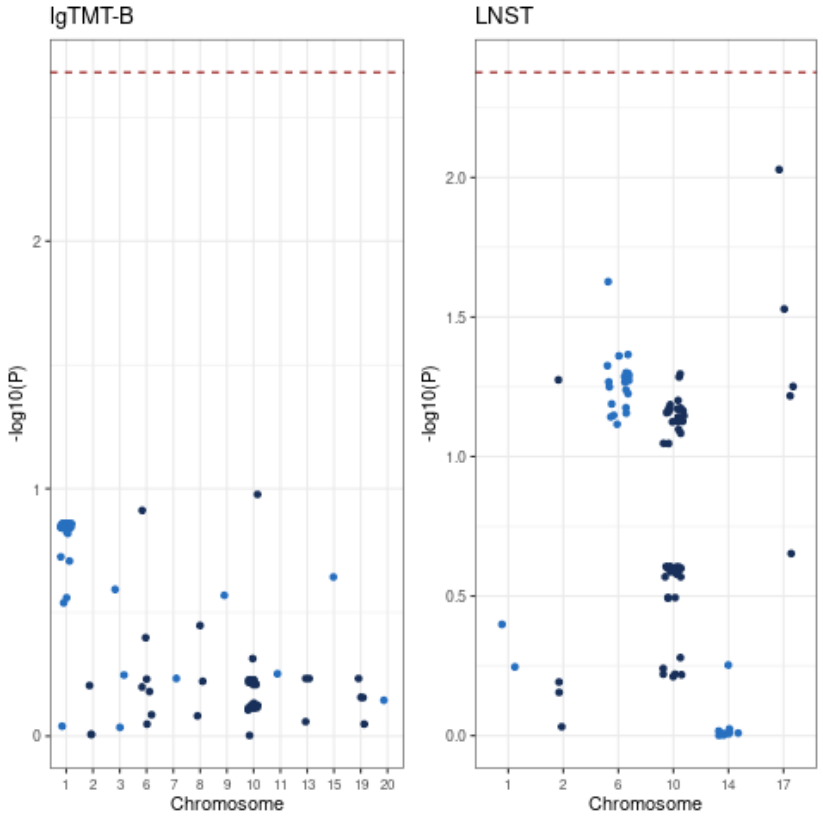
CHR	SNP	Location (BP)	$\widehat{\beta}_{DS}$ (95% c.i.)	p_{DS}	$\widehat{\beta}_{RS}$ (95% c.i.)	p_{RS}	MAF_{DS}	MAF_{RS}	A1/A2	Distance (BP)	Gene context
5	rs62368988	63304980	1.15 (1.09,1.20)	$7.4x10^{-09}$	0.89 (0.79,1.01)	0.075	0.016	0.020	T/C	28109	intergenic
5	rs62369014	63333947	1.15 (1.09,1.20)	$7.2x10^{-09}$	0.89 (0.79,1.01)	0.075	0.016	0.020	T/G	57076	intergenic
5	rs62369048	63417124	1.15 (1.10,1.20)	$3.4x10^{-09}$	0.89 (0.79,1.01)	0.074	0.016	0.020	A/G	44546	intergenic
5	rs146654929	63470382	1.15 (1.10,1.20)	$5.0x10^{-09}$	0.89 (0.79,1.01)	0.076	0.016	0.020	G/A	0	RNF 180
5	rs62372500	63547213	1.16 (1.10,1.21)	$1.5x10^{-09}$	0.89 (0.79,1.01)	0.075	0.015	0.019	T/C	0	RNF 180
5	rs191575088	63606345	1.16 (1.10,1.21)	$8.4x10^{-10}$	0.89 (0.79,1.01)	0.081	0.016	0.019	G/A	0	RNF 180
5	**rs150547358**	63640195	**1.16 (1.11,1.22)**	$\mathbf{7.2x10^{-10}}$	**0.85 (0.74,0.97)**	**0.015**	**0.015**	**0.018**	C/A	0	RNF 180
5	5:63686382:AT	63686382	1.15 (1.10,1.21)	$1.8x10^{-09}$	0.90 (0.80,1.02)	0.112	0.016	0.020	A/AT	13956	intergenic
5	rs62369421	63722706	1.14 (1.08,1.19)	$2.5x10^{-08}$	0.90 (0.80,1.01)	0.064	0.017	0.022	T/G	0	ncRNA_intronic

Supplementary Figure 4



Supplementary Figure 4. Manhattan plot of the GWAS of lgTMT-B (left) and the VDS-B (right) in the discovery sample testing the SNP main effect. The lines indicate the thresholds for genome-wide significance of 5×10^{-8} (red) and for suggestive SNPs (blue, $p\leq 1\times10^{-5}$).

Supplementary Figure 5



Supplementary Figure 5. Manhattan plot of the difference analysis of the SNP_{NR} (SNP not be replicated) for the lgTMT-B (left) and the LNST (right) in the replication sample (FOR2107 consortium), containing only suggestive (not significant) SNPs of the GWAS in the discovery sample (PsyCourse Study). The dashed red line presents the significance level (lgTMT-B: 0.0021; LNST: 0.0042) corrected for multiple testing.

Supplementary Table 2

Supplementary Table 2. Distribution of the genotypes for SNP rs150547358 in the discovery and the replication sample. Entries refer to the number of individuals with wildtype/C-carriers in the respective diagnostic groups. The p-values marked by [1] are results of a Fisher's exact test. The other p-values are results of χ^2-tests.

Visit	Discovery sample					Replication sample				
	Total	Affective	Psychotic	Controls	p-value	Total	Affective	Psychotic	Controls	p-value
1	1175/34	486/16	439/13	250/5	0.4994	595/16	278/12	39/0	278/4	0.0984[1]
2	803/22	300/11	285/7	218/4	0.4429	595/16	278/12	39/0	278/4	
3	638/18	223/8	242/7	173/3	0.5786[1]					
4	449/12	175/5	217/7	57/0	0.5982[1]					

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 06 December 2022
DOI 10.3389/fgene.2022.1015885

OPEN ACCESS

EDITED BY
Ka-Chun Wong,
City University of Hong Kong, Hong Kong SAR, China

REVIEWED BY
Madan Kundu,
Daiichi Sankyo, United States
Jixiang Yu,
City University of Hong Kong, Hong Kong SAR, China

*CORRESPONDENCE
Bernadette Wendel,
bernadette.wendel@med.uni-goettingen.de

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 10 August 2022
ACCEPTED 24 November 2022
PUBLISHED 06 December 2022

CITATION
Wendel B, Heidenreich M, Budde M, Heilbronner M, Oraki Kohshour M, Papiol S, Falkai P, Schulze TG, Heilbronner U and Bickeböller H (2022), Kalpra: A kernel approach for longitudinal pathway regression analysis integrating network information with an application to the longitudinal PsyCourse Study.
Front. Genet. 13:1015885.
doi: 10.3389/fgene.2022.1015885

Kalpra: A kernel approach for longitudinal pathway regression analysis integrating network information with an application to the longitudinal PsyCourse Study

Bernadette Wendel[1]*, Markus Heidenreich[1], Monika Budde[2], Maria Heilbronner[2], Mojtaba Oraki Kohshour[2], Sergi Papiol[2,3], Peter Falkai[3], Thomas G. Schulze[2,4,5], Urs Heilbronner[2] and Heike Bickeböller[1]

[1]Department of Genetic Epidemiology, University Medical Center Göttingen, Georg-August-University Göttingen, Göttingen, Germany, [2]Institute of Psychiatric Phenomics and Genomics (IPPG), University Hospital, LMU Munich, Munich, Germany, [3]Department of Psychiatry and Psychotherapy, University Hospital, LMU Munich, Munich, Germany, [4]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, United States, [5]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, MD, United States

A popular approach to reduce the high dimensionality resulting from genome-wide association studies is to analyze a whole pathway in a single test for association with a phenotype. Kernel machine regression (KMR) is a highly flexible pathway analysis approach. Initially, KMR was developed to analyze a simple phenotype with just one measurement per individual. Recently, however, the investigation into the influence of genomic factors in the development of disease-related phenotypes across time (trajectories) has gained in importance. Thus, novel statistical approaches for KMR analyzing longitudinal data, i.e. several measurements at specific time points per individual are required. For longitudinal pathway analysis, we extend KMR to long-KMR using the estimation equivalence of KMR and linear mixed models. We include additional random effects to correct for the dependence structure. Moreover, within long-KMR we created a topology-based pathway analysis by combining this approach with a kernel including network information of the pathway. Most importantly, long-KMR not only allows for the investigation of the main genetic effect adjusting for time dependencies within an individual, but it also allows to test for the association of the pathway with the longitudinal course of the phenotype in the form of testing the genetic time-interaction effect. The approach is implemented as an R package, *kalpra*. Our simulation study demonstrates that the power of long-KMR exceeded that of another KMR method previously developed to analyze longitudinal data, while maintaining (slightly conservatively) the type I error. The network kernel improved the performance of long-KMR compared to the linear kernel. Considering different pathway densities, the power of the network kernel decreased with increasing pathway density. We applied long-KMR to cognitive data on

executive function (Trail Making Test, part B) from the PsyCourse Study and 17 candidate pathways selected from Reactome. We identified seven nominally significant pathways.

KEYWORDS

pathway analysis, kernel machine regression, longitudinal data, network, PsyCourse Study

1 Introduction

Pathway analyses or gene-set analyses are association studies, which test whole gene sets or pathways for association with a phenotype of interest (Holmans, 2010; Mooney and Wilmot, 2015). In contrast to a genome-wide association analysis (GWAS) in which a great number of individual SNP association tests are performed, a smaller group of genes or SNPs is tested simultaneously. Thus, the multiple testing problem of a GWAS is tremendously mitigated. In the last two decades, many different general approaches and particular tools have been developed for pathway analysis (Holmans, 2010; Mooney and Wilmot, 2015; de Leeuw et al., 2016).

In this paper, we focus on kernel machine regression (KMR) (Liu et al., 2007; Ge et al., 2016), a machine-learning algorithm (Liu et al., 2007) with great flexibility. KMR is a semi-parametric regression analysis (Liu et al., 2007) initially designed to analyze case-control studies (Liu et al., 2008; Wu et al., 2010; Wu et al., 2011) or quantitative data (Liu et al., 2007; Wu et al., 2011; Ge et al., 2016). KMR models the environmental (non-genetic) parameters parametrically and the high-dimensional genetic data (e.g., genotype information) non-parametrically (Liu et al., 2007; Freytag et al., 2013). The genetic data are transformed into a similarity matrix containing for every pair of individuals quantitative values, which describe the genetic similarities of the pairs of individuals (Schaid, 2010; Ge et al., 2016). This matrix is denoted as kernel matrix. The transformation is performed by a kernel function, which can have different forms depending on the desired similarity concept (Freytag et al., 2013). There are many possibilities to model a pathway as the only requirement of the kernel function is to be positive semidefinite (Schaid, 2010; Schaid, 2010) For example, a popular kernel is the linear kernel (Wu et al., 2010; Freytag et al., 2013). New kernels have been also defined, e.g., a kernel adjusting for size bias (Freytag et al., 2012) or a kernel integrating the network information of a pathway (Freytag et al., 2013). The latter was possible thanks to the development of different pathway databases, e.g., Reactome (Jassal et al., 2019), Pathway Commons (Rodchenkov et al., 2019) or KEGG (Kanehisa et al., 2017). Different versions and extensions of KMR have been developed to address various research questions (for a summary see (Larson et al., 2019)). KMR analyzing more complex phenotype data, e.g., family samples (Malzahn et al., 2014; Yan et al., 2015) is just one example.

Longitudinal studies assess multiple, thus correlated, measurements over time for each single individual (Molenberghs and Verbeke, 2000; Caruana et al., 2015). They enable researchers to study the time course of the investigated phenotype. A number of statistical methods have been and are still being developed especially in this context. An important aspect of longitudinal studies is the frequently high number of missing data or unequal measurement points (Caruana et al., 2015). A popular method to overcome this challenge are linear mixed models (LMM) (Molenberghs and Verbeke, 2000) in which so-called random effects are added to correct for the dependence structure of the different measurements. A random effect enables the modeling of an individual development for each subject. LMMs can handle missing phenotype data under the assumption that the data are at least missing at random (MAR) (Molenberghs and Verbeke, 2000).

In the genetic context, these LMMs can be applied to perform longitudinal GWASs (Wendel et al., 2021). Using this, we previously (Wendel et al., 2021) investigated the genetic influence of individual SNPs on the course over time of executive functions, which control and coordinate mental processes. These GWASs demonstrated the versatility of LMMs in genetic association studies. Thus, the next step is to investigate pathways for association with longitudinal phenotypes, for example, the genomic basis of the longitudinal course of executive functions. For this, we can exploit that LMMs share an estimation equivalence with KMR models (Liu et al., 2007).

The aim of this work is to develop a longitudinal pathway analysis to test for the association between genetic factors and the longitudinal phenotype applying KMR and simultaneously allowing integrating network information. To be able to analyze longitudinal data, we extended KMR to long-KMR. Other authors have also studied longitudinal data (Yan et al., 2015; Ge et al., 2016; Wang et al., 2016) and created a KMR extension (Yan et al., 2015; Yan et al., 2018). However, in this extension only single genes can be tested for association (Yan et al., 2018) and these genes can solely be modeled with a weighted linear kernel. In our longitudinal pathway analysis, the whole pathway can be modeled with different kernels respectively prior to testing. For example, a linear kernel or a network-based kernel (Freytag et al., 2013), which enables the integration of network information in KMR can be applied. Moreover, different genetic effects including main, interaction,

and joint genetic (main and interaction) effects can be considered. Thus, in long-KMR, we can model and test not only a main genetic effect, but most importantly also a genetic time-interaction effect. The latter translates to an association of the pathway with the trajectory of the considered phenotype.

In a simulation study, we assessed the properties of long-KMR regarding several aspects. We considered longitudinal studies with two and four measurement points. We compared the performance of long-KMR when applying a linear kernel or a network kernel. We also studied the influence of the pathway topology on the performance of the network kernel with a focus on the density of the pathway.

Finally, as a real-world application, we use long-KMR on the data from our previous longitudinal GWASs (Wendel et al., 2021) on executive functions of the PsyCourse Study (Budde et al., 2018). For this phenotype we chose several candidate pathways to be investigated with long-KMR.

In summary, in this paper we first present the theoretical aspects of long-KMR and the network kernel. We then describe the simulation approach used to evaluate our method, and, lastly, provide a real-world example of long-KMR.

2 Material and methods

In this section, we introduce the KMR analysis and its extension to analyze longitudinal data. We describe our simulation approach to investigate the type I error rate and power. Lastly, we present an application of long-KMR as example and give details on the PsyCourse Study data and the pathways used.

2.1 Kernel machine regression models

Let us assume y_i is a quantitative phenotype for individual i ($i = 1, \ldots, n$) with one measurement point per individual. We assume for the entire article that the pathway tested is represented as genotypes of the SNPs part of the pathway. The SNPs are coded as 0, 1, or 2, representing the number of minor alleles of the SNP in individual i. The genetic information for individual i of all selected SNPs s is stored as genotype vector g_i. We regress g_i on our phenotype of interest by applying the following model:

$$y_i = x_i\beta + h(g_i) + \varepsilon_i,$$

where y_i is the phenotype of interest for individual i, x_i represents potential covariates, β is the regression coefficient vector, and h is a non-parametric function. This function $h \in H_K$, where H_K is a reproducing kernel Hilbert space with an inner product (Schaid, 2010; Ge et al., 2016). The reproducing kernel Hilbert space is generated by a positive semidefinite kernel function k (Liu et al., 2007; Ge et al., 2016). The mathematical characteristics of the reproducing kernel Hilbert space (e.g., inner product) allows approximating h as a linear combination of the kernel function k (Liu et al., 2007; Schaid, 2010; Ge et al., 2016). The "kernel trick" (Ge et al., 2016) specifies hereby that any positive semidefinite kernel function can be used as k. We define the corresponding kernel matrix K as $K \coloneqq k(g_i, g_j)$ for any pair of individuals i and j of the associated kernel function k (Schaid, 2010; Ge et al., 2016). Here, we transform the high-dimensional $n \times s$ genotype matrix into a $n \times n$ similarity matrix. The kernel matrix K describes the similarity between each pair of individuals. By choosing a kernel, we can specify how to model the concept of genetic similarity. For example, we can use the popular linear kernel (LIN), which computes the similarity for each pair of individuals i and j by multiplying their genotype vectors g_i and g_j. The kernel matrix contains the elements $K(g_i, g_j) = g_i^T g_j$ (matrix notation: $K = GG^T$). The linear kernel assumes that each SNP contributes a random independent value in an additive manner (Freytag et al., 2013). For the above model, we assume that the random error $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $h \sim \mathcal{N}(0, \tau K)$ with K being the kernel matrix and τ a variance component. The null hypothesis of our association test is H_0: $h = 0$ being equivalent to H_0: $\tau = 0$ (Liu et al., 2007; Wu et al., 2011). To test for association, we perform a variance component test (Liu et al., 2007; Wu et al., 2011).

The KMR model can be read as a LMM with h being interpreted as a random effect (Liu et al., 2007; Ge et al., 2016). The above model for a quantitative phenotype with one measurement per individual can also be described as LMM (in matrix notation) (Liu et al., 2007; Ge et al., 2016):

$$y = X\beta + h + \varepsilon,$$

where y is the vector of phenotypes for n individuals, X is the design matrix, β is the regression coefficient vector of the fixed effects, h $\sim \mathcal{N}(0, \tau K)$ is the random effect vector with K being the kernel matrix, the random error ε is normally distributed. The variance component test for this model (Liu et al., 2007; Ge et al., 2016) is:

$$Q_{cross} = \frac{1}{2\sigma_\varepsilon^2}\left(y - X\hat{\beta}_0\right)^T K\left(y - X\hat{\beta}_0\right),$$

where $\hat{\beta}_0$ are the estimates of the fixed effects under H_0. For the longitudinal extension, we adjust for the dependence structure of the multiple measurements in the longitudinal data by including additional random effects (Molenberghs and Verbeke, 2000). Now we assume that y_i is a quantitative longitudinal phenotype for individual i ($i = 1, \ldots, n$) with m measurement points. The long-KMR model for individual i is:

$$y_i = X_i\beta + h(g_i) + Z_i b_i + \varepsilon_i,$$

where y_i is the phenotype vector of individual i, β is the fixed effect vector, and b_i the random effect vector. We assume that only two random effects are added (random intercept and slope

for time). Thus, we assume that $b_i \sim \mathcal{N}(0, D_i)$ with D_i being a 2×2 covariance matrix and $\varepsilon_i \sim \mathcal{N}(0, R_i)$ with R_i being a $m \times m$ covariance matrix, b_i and ε_i are uncorrelated. X_i and Z_i are two designs matrices for the fixed and random effects, respectively. The genotype vector g_i and function h are given as above. To obtain the test statistic of the extended variance component test, we followed the steps proposed by (Liu et al., 2007; Yan et al., 2015). Therefore, we look at the longitudinal model in matrix notation considering the whole dataset:

$$y = X\beta + h(G) + Zb + \varepsilon$$

where y is the phenotype vector, $h \sim \mathcal{N}(0, \tau \boldsymbol{K})$, $b \sim \mathcal{N}(0, D)$ $(D = diag(D_1, \ldots, D_n))$ and $\varepsilon \sim \mathcal{N}(0, R)$ with $R = diag(R_1, \ldots, R_n)$. The design matrices are $X = (X_1, \ldots, X_n)^T$ for the fixed effects and $Z = diag(Z_1, \ldots, Z_n)$ for the random effects, with β and b being the fixed and random effect vectors, respectively. The null hypothesis remains H_0: $\tau = 0$. The altered test statistic is:

$$Q_{long} = \frac{1}{2}\left(y - X\hat{\beta}_0\right)^T \hat{\Sigma}_0^{-1} \boldsymbol{K} \hat{\Sigma}_0^{-1}\left(y - X\hat{\beta}_0\right),$$

where $\hat{\beta}_0$ are the estimates of the fixed effects under H_0 and $\hat{\Sigma}_0^{-1}$ are the inverse of the covariance-variance matrix under H_0 with $\hat{\Sigma}_0 = \hat{R}_0 + Z\hat{D}_0 Z^T$. The test statistic is a quadratic form and follows a mixture of χ^2 distributions with $Q_{long} \sim \sum_{l=1}^{L} \lambda_l \chi_1^2$, where λ_l are the eigenvalues of $\frac{1}{2} V_0^{\frac{1}{2}} \hat{\Sigma}_0^{-1} \boldsymbol{K} \hat{\Sigma}_0^{-1} V_0^{\frac{1}{2}}$ with $V_0 = \hat{\Sigma}_0 - X(X^T \hat{\Sigma}_0^{-1} X)^{-1} X^T$ (Yan et al., 2015; Ge et al., 2016). We computed the p-values with the Davies method (Davies, 1980).

Next, we will apply long-KMR to test a genetic (G) interaction effect with time (t). Here, we multiply the time vector of individual i, $t_i = 0, \ldots, m-1$ with the genotype vector g_i of individual i. In addition to the main genetic kernel ($h(G)$) this extended model contains a kernel modelling the genetic time interaction effect ($t \times G$, further denoted as time-interaction effect). In matrix notation (whole dataset) the model is:

$$y = X\beta + h_1(G) + h_2(t \times G) + Zb + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, R)$, $h_1(G) \sim \mathcal{N}(0, \tau_1 \boldsymbol{K_1})$ and $h_2(t \times G) \sim \mathcal{N}(0, \tau_2 \boldsymbol{K_2})$. The notation follows the previous long-KMR in matrix notation. When fitting the LMM in this interaction model, we have to integrate $\boldsymbol{K_1}$ as random effect in form of a variance-covariance matrix. This is complex and computationally very extensive. We use two different approaches to reduce the computation time. For the first approach, we only include $h_2(t \times G)$ in our model without adjusting for the main genetic effect ($h_1(G)$) altering the LMM independent of any kernel matrix under the null hypothesis. For the second approach, we adjust for the main genetic effect by performing a principal component analysis (PCA) on $\boldsymbol{K_1}$. This so-called kernel principal component analysis (KPCA, (Schölkopf et al., 1997; Schölkopf et al., 1998)) has been previously applied in different situations (Schölkopf et al., 1997; Schölkopf et al., 1998; Gao et al., 2011). We replace $h_1(G)$ by a number of top principal components, which are added as fixed effects. By only including additional fixed effects, we avoid complex variance structures while adjusting for the main genetic effect. In both approaches, we are interested in testing $\boldsymbol{K_2}$, modeling the time-interaction effect for association. The null hypothesis is defined as H_0: $\tau_2 = 0$. The test statistic of long-KMR is slightly altered, as $\boldsymbol{K}$ of Q_{long} is exchanged with $\boldsymbol{K_2}$ modeling the time-interaction effect.

2.2 Network kernel

In long-KMR, we can also integrate network information on the studied pathway by applying the network-based kernel (Freytag et al., 2013) (noted as network kernel in the following). The network kernel is defined as $K = GANA^TG^T$, where G is the genotype matrix with the genotypes for each individual, A is an annotation matrix and N is an adjacency matrix of the pathway. The annotation matrix contains elements $a_{\rho\gamma} \in (0, 1)$ describing whether a SNP ρ ($\rho = 1, \ldots, s$) is mapped to the gene γ (=1) or not (=0). The assignment of a SNP to a gene is defined by its genomic location. We can adjust for different gene sizes (= number of SNPs mapped) by dividing $a_{\rho\gamma}$ by the square root of the number of SNPs mapped to gene γ (*Freytag et al.*, 2013). The size-adjusted annotation matrix replaces A in the network kernel. We distinguish these network kernels by denoting the unadjusted kernel as NET and the size-adjusted network kernel as ANET [similar to (Freytag et al., 2013)]. The elements of the quadratic adjacency matrix for a pathway are $n_{\gamma\gamma'} = 1$, if genes γ and γ' interact with each other, or zero otherwise. By definition (Freytag et al., 2013), the genes all interact with themselves; thus, the main diagonal of N contains only '1's. We do not distinguish between the different types of gene interaction (e.g., activation and inhibition) owing to the characteristics of the studied pathways (more details later). We slightly modify the network (topology) of the pathway to ensure a positive semidefinite kernel. We do not describe the details of these modifications here; please refer to (Freytag et al., 2013) for more details.

2.3 Simulation study

We studied type I error rates and power in different scenarios to assess the performance of long-KMR for different genetic effects and the network kernel. The type I error rate is defined as the proportion of simulations that have a p-value < α in the simulations of the model with no genetic effects (null model). Here we set α to equal 5%, 1%, 0.5%, and 0.1%, respectively. In the scenarios in which we simulated genetic effects, we determined the power as the proportion of simulations with a

TABLE 1 Models of simulation study.

Model names	Kernel	Statistical model (without genetic effect)	Phenotype data	Genetic effect model		
				Main genetic	Time-interaction	Joint
Two-measurement models						
KMR-LIN-ANCOVA	Linear kernel	ANCOVA: $y_2 = \beta_0 X_i + \beta_1 y_1 + \varepsilon$	complete data	*	*	—
KMR-LIN-m2	Linear kernel	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	—
KMR-LIN-m2_25MAR			25% missing data (MAR)	*	*	—
KMR-NET-d0.8-m2	Network kernel (pathway d = 0.8)	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	—
KMR-NET-d0.8-m2_25MAR			25% missing data (MAR)	*	*	—
KMR-NET-d0.5-m2	Network kernel (pathway d = 0.5)	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	—
KMR-NET-d0.5-m2_25MAR			25% missing data (MAR)	*	*	—
KMR-NET-d0.2-m2	Network kernel (pathway d = 0.2)	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	—
KMR-NET-d0.2-m2_25MAR			25% missing data (MAR)	*	*	—
Four-measurement models						
KMR-LIN-m4	Linear kernel	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	*
KMR-LIN-m4_25MAR			25% missing data (MAR)	*	*	—
KMR-LIN-m4_50MAR			50% missing data (MAR)	*	*	—
KMR-NET-d0.8-m4	Network kernel (pathway d = 0.8)	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	*
KMR-NET-d0.8-m4_25MAR			25% missing data (MAR)	*	*	—
KMR-NET-d0.8-m4_50MAR			50% missing data (MAR)	*	*	—
KMR-NET-d0.5-m4	Network kernel (pathway d = 0.5)	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	*
KMR-NET-d0.5-m4_25MAR			25% missing data (MAR)	*	*	—
KMR-NET-d0.5-m4_50MAR			50% missing data (MAR)	*	*	—
KMR-NET-d0.2-m4	Network kernel (pathway d = 0.2)	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	*	*
KMR-NET-d0.2-m4_25MAR			25% missing data (MAR)	*	*	—
KMR-NET-d0.2-m4_50MAR			50% missing data (MAR)	*	*	—
KMgene (comparison model) (Yan et al., 2018)	Weighted linear kernel	KMR (LMM): $y = X\beta + Zb + \varepsilon$	complete data	*	—	—

For each model, the kernel, the applied statistical models and the used phenotype data sets are displayed. For network kernel the pathway density (d) is given. The phenotype data can be complete or with 25/50% of values missing at random (MAR). The addressed genetic effects (main genetic, time-interaction and joint effect) are indicated with an asterisk "*".

p-value <5% threshold. In total, we compared the power for three different genetic effect models. We simulated two single-effect models containing either a main genetic effect or a time-interaction effect. We also created a more complex model, the joint model, which comprises a main genetic effect and a time-interaction effect. The joint model was only studied in a limited number of scenarios, as portrayed in Table 1. For the single-effect models, we had the same scenarios to evaluate the type I error rates and the power. We assessed the influence of the number of measurement points comparing two-measurement models with four-measurement models. The type I error rates and respective power of the linear kernel (LIN) and the network kernel (NET) were compared. For the latter, we only used the unadjusted network kernel (NET), as all genes had the same size. For two measurement points representing a pre/post-analysis, we applied the ANCOVA model (Table 1) to compare their performances with long-KMR. For the four-measurement models, we compared the performance of long-KMR with the previously

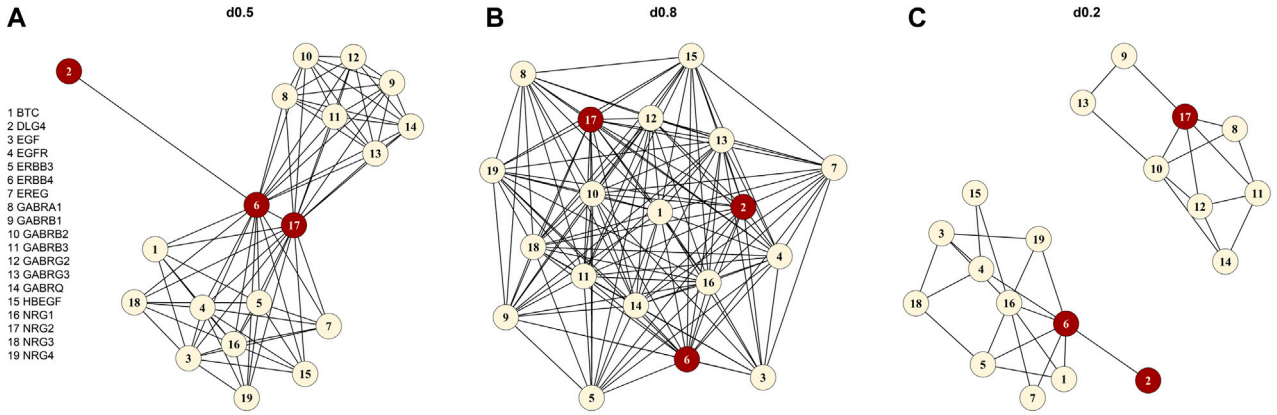


FIGURE 1
Graphical illustration of the "signaling by ERBB4" pathway **(A)** original form without the chemical compound (d0.5), **(B)** the simulated high-density pathway (d0.8), **(C)** the artificially created low-density pathway (d0.2). The red vertices are the three defined causal genes (NRG2, ERBB4 and DLG4).

published KMgene package (Yan et al., 2018). Further, we compared the analysis of complete phenotype data with incomplete phenotype data with 25% or 50% of the data missing [assuming missing at random mechanism (MAR)].

To evaluate the performance of the network kernel on pathways with different characteristics, we focused on the density (=d) of a pathway. This density is a graph-theoretical characteristic defined as the ratio of the number of present connections divided by the maximum number of possible connections in a pathway ($d \in [0, 1]$). When we consider a pathway as a graph in which the genes are the nodes and the connections of the genes are the edges linking the nodes, the density can be computed straightforwardly. We determined the density of the original pathway after downloading the pathway from the Reactome database, applying the *igraph* package (Csardi and Nepusz, 2006). We selected the "signaling by ERBB4" pathway [R-HSA-1236394, (Stern, 2019)] as foundation pathway for our simulation study. The selection process for "signaling by ERBB4" is described in detail in the section *Pathway Data*. The "signaling by ERBB4" pathway has a density of 0.46 but we denoted the pathway as d0.5 after rounding up $d = 0.46$ (Figure 1A). In addition, we created two artificial pathway topologies with different density originating from the original "signaling by ERBB4" pathway. We generated a high-density pathway with $d = 0.81$ (denoted as d0.8, Figure 1B) and a low-density pathway with $d = 0.20$ (d0.2, Figure 1C). Table 1 lists all the models studied with an overview of the different settings.

We sampled genotypes for 10,000 individuals with HAPGEN2 (Su et al., 2011) using common (MAF ≥ 0.05) variants of chromosome one of the CEU sample of the International HapMap Project (HapMap 3 release 2) (Altshuler et al. 2010). In analogy to our foundation pathway "signaling by ERBB4" with 19 genes (Figure 1), we created 19 "pseudo-" genes all with a size of 50 SNPs (in total: 950 SNPs). The 950 SNPs were simulated in the region between 742 kbp and 112,709 kbp with a separation of 500 kbp between SNPs of the single "pseudo-" genes to prevent LD. We assign the simulated SNPs to a "pseudo-" gene. For each simulation setting, we created 100 smaller genotype matrices each containing 950 SNPs and 1,000 individuals. To achieve this, we randomly drew genotypes for 1,000 individuals from the previously simulated 10,000 individual sample (elementary matrix). For each of the 100 genotype matrices, we simulated 1,000 quantitative phenotypes according to the LMM below, resulting in a total number of 100,000 replications [similar to (Yan et al., 2015)].

For the null model corresponding to the null hypothesis of no genetic effects, we simulated the quantitative phenotypes according to the following LMM for an individual i ($i = 1, \ldots, 1000$):

$$y_i = 0.5^* X_{1i} + 0.25^* X_{2i} + 0.2^* t_i + u_i,$$

where X_{1i} is a binary time-invariant variable with a probability of 0.5 (e.g., sex of individual i), X_{2i} is normally distributed and time-invariant with $\mathcal{N}(50, 5)$ (e.g., age at first measurement point) and $t_i = 0, \ldots, m-1$ where m equals the total number of measurement points ($m = 2$ or $m = 4$). Random error and random effects are modelled by u_i, which follow a multivariate normal distribution with mean zero and $Var(y_i)$. $Var(y_i)$ is defined as follows:

$$Var(y_i) = Z_i \begin{pmatrix} \sigma^2_{intercept} & \sigma_{cov} \\ \sigma_{cov} & \sigma^2_{time} \end{pmatrix} Z_i^T + \sigma^2_\varepsilon I_{m\times m}$$

where $I_{m\times m}$ is the identity matrix, $\sigma^2_{intercept} = \sigma^2_{time} = \sigma^2_\varepsilon = 1$ and $\sigma_{cov} = -0.5$. We selected the parameters similarly to (Yan et al., 2015). For the missing phenotype simulations, we assumed MAR

and generated the missing phenotypes with the R package mice (van Buuren and Groothuis-Oudshoorn, 2011).

For the power simulations, we added genetic effects to our null model to simulate the phenotypes. All models comprised three causal "pseudo-" genes each with three causal SNPs (in total: nine causal SNPs). The effect sizes β_k for each SNP had the same value. The effect size for the joint model was 0.04. For the single-effect models, we studied three different scenarios with three different effect sizes $\beta = 0.04$, 0.06, and 0.08. To compare the different network topologies, we defined the genes NRG2, ERBB4, and DLG4 of the "signaling by ERBB4" pathway as the causal genes (red nodes, Figure 1) based on their central position in the pathway. The main genetic effect model adds a sum consisting of the additive effect of the causal SNPs to the phenotype ($\sum_{k=1}^{9}\beta_k{}^*SNP_{ik}$) for each individual i. The time-interaction effect includes only the sum of the product of the causal SNPs and the time ($\sum_{k=1}^{9}\beta_k{}^*(SNP_{ik}{}^*t_{ij})$) at each time point j for individual i. The joint model comprised both sums ($\sum_{k=1}^{9}\beta_k{}^*SNP_{ik} + \sum_{k=1}^{9}\beta_k{}^*(SNP_{ik}{}^*t_{ij})$). In the first model, the main genetic kernel ($h(G)$) is tested. The latter models test the time-interaction kernel ($h(t \times G)$) for association. In the joint model, the main genetic kernel was computed with the linear kernel. Here, we performed a principal component analysis on the main genetic kernel to adjust for the main genetic effect to simplify computational complexity and gain speed. We added the top two principal components as fixed effects to our model.

To compare the type I error rate and power of long-KMR with KMgene (Yan et al., 2015; Yan et al., 2018) we performed a simulation with KMgene for 1,000 individuals and four measurement points. Here, only the main genetic effect model was simulated because of the characteristics of KMgene (Yan et al., 2018). For every simulated gene (in total: 19, each with 50 SNPs), we obtained a gene-level p-value, which we combined with the Fisher's method (Fisher, 1925; Larson et al., 2017) to receive a pathway p-value. This p-value combination was performed with the R package *metap* (Dewey, 2022).

2.4 Application to real data

2.4.1 The PsyCourse Study

The PsyCourse Study is a longitudinal, multi-center study comprising patients with diagnoses from the affective-to-psychotic spectrum and neurotypical individuals. A large battery of different phenotypes, including demographics, cognition, self- and observer rating scales, are assessed at up to four measurement points each 6 months apart (Budde et al., 2018). For our application, we analyzed 1,594 genotyped individuals including patients from the affective-to-psychotic spectrum (411 bipolar I disorder, 113 bipolar II disorder, 466 schizophrenia, 90 schizoaffective disorder, 10 schizophreniform disorder, 6 brief psychotic disorder and 94 with recurrent depression) and 404 control individuals. The diagnoses were determined according to the criteria in the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV); a subset of individuals suffering from schizophrenia (45 individuals) was diagnosed according to ICD-10 criteria. Different centers in Germany and Austria conducted the recruitment of the study participants. All individuals provided written informed consent, and the study protocol was approved by the respective ethics committees at each study center [see ref. (Budde et al., 2018)]. Based on their symptoms, the individuals were broadly distinguished into an "affective" group (618, predominantly affective symptoms including bipolar disorder I and II and recurrent depression) or a "psychotic" group (572, predominantly psychotic symptoms encompassing schizophrenia, schizoaffective, schizophreniform, and brief psychotic disorder).

As phenotype of interest, we chose the Trail Making Test, part B (TMT-B) (Bowie and Harvey, 2006). TMT-B is applied to assess set-shifting, one of the three latent core skills of executive functions (Diamond, 2013; Friedman et al., 2016), a specific group of cognitive abilities. During the test, an individual is required to connect numbers (numbers: 1–26) and letters of the alphabet in ascending alternating order, for which the time (in seconds) to finish this task is measured to represent the test score. Study participants with a time >300 s were set to 300 s according to the recommendation by (Strauss et al., 2006). The higher the TMT-B score of an individual is, the greater the cognitive impairment.

Genotyping was performed with the Illumina Infinium Global Screening Array-24 Kit (version 3.0 or version 1.0) and the imputation took place on the Michigan imputation server (Das et al., 2016) with the haplotype reference consortium as reference panel. Quality control (QC) steps were performed according to standard procedures described elsewhere (Smigielski et al., 2021). In the analysis, we included approximately 3.5 million imputed SNPs with a MAF >0.05. We used PLINK v1.9 (Chang et al., 2015) (https://www.cog-genomics.org/plink/) to compute the ancestry principal components.

2.4.2 Pathway data

We focused on pathways on the Reactome database (Jassal et al., 2019) downloaded from Pathway Commons database Version 12 (Rodchenkov et al., 2019) (Reactome version 69, date: 01|14|22). First, we selected pathways based on different keywords connected to executive functions including dopamine, serotonin, GABA, glutamate, NDMA, synaptic, voltage-gated potassium channels, plasticity, and prefrontal cortex. The keywords resulted in 130 pathways, which we reduced to the 17 pathways finally studied (Table 2). We selected the 17 pathways according to different criteria. First, we only used pathways that we were able to download. The pathway had to be between 15 and 100 genes in size, and the number of chemical compounds (CHEBI) in the pathway had to be at most five. For

TABLE 2 Selected pathways investigated in the real-data example.

Pathway name	Reactome identifier (R-HSA-xxx)	URL	Pathway Characteristics		
			No. Genes	Average degree	Density (d)
NCAM1 interactions	419037	https://reactome.org/content/detail/R-HSA-419037	37	3.40	0.093
Receptor-type tyrosine-protein phosphatases	388844	https://reactome.org/content/detail/R-HSA-388844	20	3.10	0.163
MECP2 regulates neuronal receptors and channels	9022699	https://reactome.org/content/detail/R-HSA-9022699	18	3.00	0.177
EPHB-mediated forward signaling	3928662	https://reactome.org/content/detail/R-HSA-3928662	33	7.45	0.233
Synaptic adhesion-like molecules*	8849932	https://reactome.org/content/detail/R-HSA-8849932	22	4.67	0.233
Transcriptional Regulation by MECP2	8986944	https://reactome.org/content/detail/R-HSA-8986944	17	4.00	0.250
Neurexins and neuroligins	6794361	https://reactome.org/content/detail/R-HSA-6794361	57	14.39	0.257
EPH-Ephrin signaling	2682334	https://reactome.org/content/detail/R-HSA-2682334	22	5.45	0.260
Regulation of MECP2 expression and activity	9022692	https://reactome.org/content/detail/R-HSA-9022692	31	8.00	0.267
Signaling by ERBB4*	1236394	https://reactome.org/content/detail/R-HSA-1236394	19	8.32	0.462
Trafficking of AMPA receptors	399719	https://reactome.org/content/detail/R-HSA-399719	17	7.88	0.493
NCAM signaling for neurite out-growth*	375165	https://reactome.org/content/detail/R-HSA-375165	21	11.00	0.524
Assembly and cell surface presentation of NMDA receptors	9609736	https://reactome.org/content/detail/R-HSA-9609736	24	12.08	0.525
Interaction between L1 and Ankyrins	445095	https://reactome.org/content/detail/R-HSA-445095	29	20.28	0.724
Negative regulation of NMDA receptor-mediated neuronal transmission	9617324	https://reactome.org/content/detail/R-HSA-9617324	21	16.86	0.843
Long-term potentiation*	9620244	https://reactome.org/content/detail/R-HSA-9620244	23	19.22	0.874
Ion channel transport*	983712	https://reactome.org/content/detail/R-HSA-983712	24	21.83	0.949

The pathways are listed according to ascending density and with links to their Reactome entry. The foundation pathway in our simulation study is printed in bold, the pathways with a p-value <0.1 in our application are further discussed and are labelled with an asterisk "*".

each pathway, we specified the density (d) by applying the *igraph* package (Csardi and Nepusz, 2006) and included only pathways with $d \leq 0.95$. The 17 selected pathways with specific characteristics e.g., number of genes, density, and average degree are displayed in Table 2. The average degree of a pathway is the average number of connections of a gene (=node). Originating from the list of 17 pathways, we chose "signaling by ERBB4" (Stern, 2019) (https://reactome.org/content/detail/R-HSA-1236394) as foundation pathway for our simulation study. This pathway was selected for its moderate size of 19 genes and because it only contains one CHEBI. The network consists of only one graph component, also denoted as connected (i.e., any gene can be reached from any other gene *via* a path). Most importantly, the pathway has an intermediate density of 0.46, which was a good basis for further artificial pathways we generated with high and low densities. "Signaling by ERBB4" is connected to schizophrenia (Banerjee et al., 2010) and schizophrenia endophenotypes, e.g. cognitive functions (Banerjee et al., 2010; Tian et al., 2017; Shi and Bergson, 2020) and thus is biologically very interesting. We deleted the CHEBI, as SNPs are the genomic basis in our analysis and a CHEBI cannot be assigned.

2.4.3 Statistical analysis

Each of the 17 pathways was tested for association with TMT-B. To fulfil the normality assumption, the TMT-B was log-transformed (lgTMT-B). We included the following fixed effects in the model: sex, age at first measurement point, diagnostic

group (affective, psychotic, and control), time, and the top five ancestry principal components. A random intercept and a random slope for the time effect were also added. We tested each pathway for a potential main genetic and a time-interaction effect. The linear kernel (LIN), the unadjusted network (NET), and the size-adjusted network kernel (ANET) were applied. We assigned a SNP to a gene of a pathway based on its genomic location with a mapping window of ± 500 kbp on each side of the gene. For the multiple testing correction, we considered the overlap of the tested pathways and computed the number of effective pathways (P_{eff}) according to (Hendricks et al., 2013; Larson et al., 2017). We computed a 17 × 325 matrix W for the 17 tested pathways and the 325 genes comprised in the 17 pathways with

$$w_{ry} = \begin{cases} \frac{1}{\sqrt{|P_r|}}, & if\ gene\ y \in pathway\ r \\ 0, & otherwise \end{cases},$$

where $|P_r|$ is the number of genes contained in pathway r. From the product of this matrix with its transpose, we computed the eigenvalues to obtain P_{eff} according to the Gao approach (Hendricks et al., 2013). We determined the number of eigenvalues required to fulfil $\frac{\sum_{r=1}^{P_{eff}}|\lambda_r|}{\sum_{r=1}^{P}|\lambda_r|} \geq c$, setting c to 0.95 leading to $P_{eff} = 15$. We set c to 0.95, as it was sufficient for us that the effective number of pathways explains 95% of the total variance. The adjusted significance level was computed as $\alpha_{Gao} = \frac{0.05}{15} = 0.0033$.

2.5 Code availability

We performed all analyses with R (R Core Team, 2021), which we also used to implement the KMR for quantitative longitudinal data and cross-sectional binary and quantitative data as an R package kalpra (kernel approach for longitudinal pathway regression analysis) available at https://gitlab.gwdg.de/bernadette.wendel/kalpra. In addition to the linear and network kernels, a quadratic kernel is also available. The pathway information can be directly downloaded and transformed into an annotation and adjacency matrix. The computational aspects for some example analyses are provided in Supplementary Table S1.

3 Results

3.1 Simulation studies

The type I error rate in our simulation study is defined, as mentioned above, as the proportion of simulations for which we obtained a p-value < α (α = 5%, 1%, 0.5%, and 0.1%) in the null simulations without genetic effects. The type I error rates were maintained overall at the different α thresholds for the models in our simulation scenarios. We did detect individual type I error rates only slightly exceeding the respective significance levels, e.g., 5% and 1%, for three models (KMR-NET-d0.5-m2, KMR-NET-d0.8-m2, and KMR-NET-d0.8-m4). However, all the values lie in the range of expected random variations (confidence intervals of the null model simulations carried out, data not shown). KMR-NET-d0.8-m2 presented the largest increase in type I error of 5.09% at the significance level of 5%. Table 3 displays the type I error rates. The error rates of the different network kernels (all densities) were overall higher compared to the linear kernel and were closer to the nominal level. The combined pathway p-values of the KMgene analysis revealed an inflation of the error rates. The error rates for the analyses of the missing aspect for the different network kernel were also maintained (Supplementary Table S2). Figure 2 displays a QQ-plot of the distribution of the multiple error rates for all models analyzing complete data distinguished between the two-measurement and four-measurement models including KMgene.

For the two single-genetic-effect models (main genetic and time-interaction model), the power comparison of the two-measurement models revealed that the LMM had the highest power independent of effect size and for either kernel. ANCOVA had the lowest power for the two-measurement models in comparison. Table 4 displays the results for the effect size β = 0.04. Increasing the number of measurement points resulted in an improvement of the power for long-KMR, in particular for the time-interaction effect (an increase from 21% to 52% (time-interaction effect) compared to 33%–36% (main genetic effect)). Overall, the time-interaction effect yielded a higher power for the four-measurement models, especially for the smaller effect sizes 0.04 (Table 4) and 0.06 (Supplementary Table S3). An additional power benefit compared to the linear kernel was achieved when applying the network kernel. For the main genetic effect with effect size 0.04, the network kernel in the two-measurement models demonstrate a higher power than KMR-LIN-m4. However, the power gain for the network kernel depends on the pathway density. The power increases with decreasing density (d0.2 > d0.5 > d0.8). A direct comparison of the power for the linear and network kernels for the four-measurement models is displayed in Figures 3A,B for the main genetic effect and the time-interaction effect, respectively. The power differences between KMR-LIN-m4 and KMR-NET-d0.8-m4, the pathway with the highest density, fluctuated in the different settings (Table 4; Supplementary Tables S3,S4). In the joint modeling of main genetic and time-interaction effects, the network kernel with the lowest density displayed the highest power. Table 5 illustrates the results for a genetic effect size of 0.04. Here, at the significance level of 5% the linear kernel had the second highest power followed by KMR-NET-d0.5-m4 and KMR-NET-d0.8-m4 (Table 5). As displayed in Figure 3C, the power of LIN-m4

TABLE 3 Type I error rates of the simulation studies.

Models	Estimated type I error rate (%)			
	α = 5%	α = 1%	α = 0.5%	α = 0.1%
KMR-LIN-ANCOVA	4.33	0.79	0.37	0.07
KMR-LIN-m2	4.37	0.78	0.38	0.07
KMR-NET-d0.8-m2	5.09	1.00	0.50	0.09
KMR-NET-d0.5-m2	5.02	1.01	0.48	0.07
KMR-NET-d0.2-m2	4.96	0.96	0.50	0.09
KMR-LIN-m4_25MAR	4.47	0.84	0.40	0.08
KMR-LIN-m4_50MAR	4.38	0.82	0.41	0.09
KMR-LIN-m4	4.32	0.78	0.38	0.07
KMR-NET-d0.8-m4	4.93	0.92	0.44	0.11
KMR-NET-d0.5-m4	4.83	0.92	0.45	0.07
KMR-NET-d0.2-m4	4.48	0.94	0.45	0.08
KMgene* Yan et al. (2018)	5.31	1.13	0.57	0.11

Simulated type I error for tests at significance levels of α = 5%, 1%, 0.5% and 0.1% are displayed. The simulations are based on 100,000 runs each with 1,000 individuals. *For comparability, the single gene-level *p*-values of KMgene are combined to a pathway *p*-value using Fisher's method.

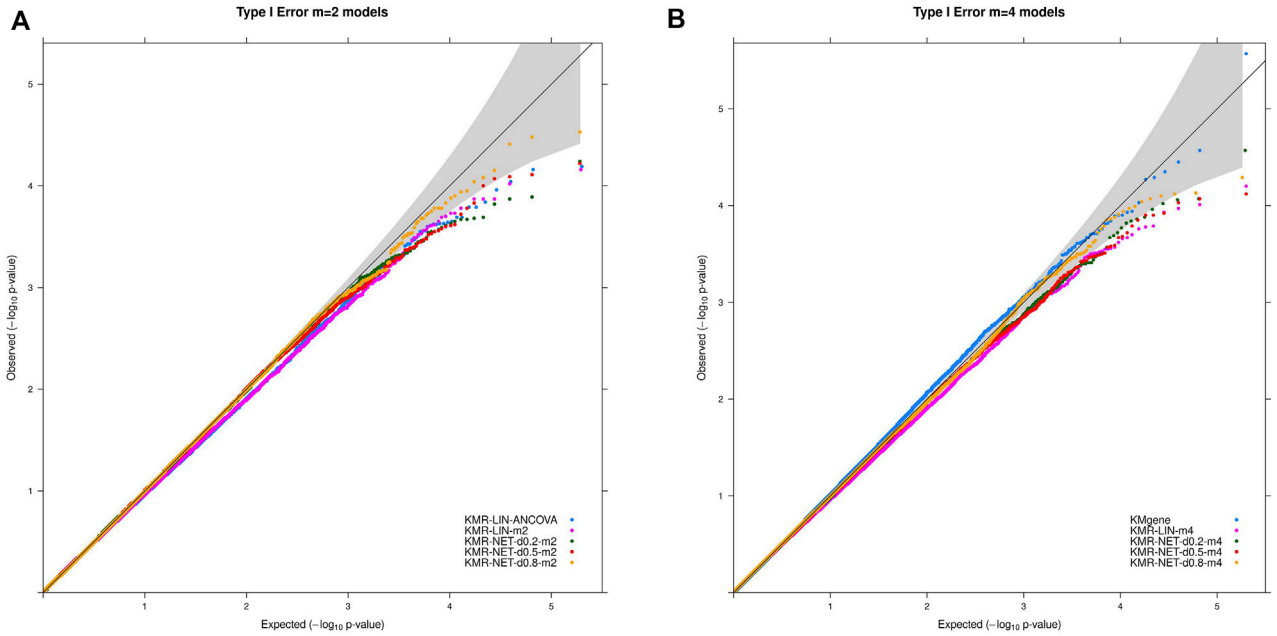


FIGURE 2
QQ-plots for the type I error rate for our simulation studies divided for the **(A)** two-measurement and **(B)** four-measurement models and the comparison model KMgene.

and KMR-NET-d0.5-m4 are very similar at different significance levels.

The analyses performed with different percentages of missing data revealed similar features for the single-effect models, with a general decrease of power compared to the analysis of a complete phenotype data set (Table 4). In general, the power increased with increasing effect sizes (β = 0.04, 0.06, and 0.08). For example, for KMR-LIN-m4 testing the main genetic effect, the power increased from 36% (β = 0.04) to 78% (β = 0.06, Supplementary Table S3), and then to 98% (β = 0.08,

TABLE 4 Power results of the simulation study.

	Genetic effect					
	Main genetic effect			Time-interaction effect		
Models	Complete	25MAR	50MAR	Complete	25MAR	50MAR
KMR-LIN-ANCOVA	12.48% [12.28; 12.69]	—	—	5.16% [05.02; 05.30]	—	—
KMR-LIN-m2	33.09% [32.79; 33.38]	25.26% [24.99; 25.53]	—	21.28% [21.02; 21.53]	07.12% [06.96; 07.28]	—
KMR-NET-d0.8-m2	39.46% [39.15; 39.76]	31.66% [31.37; 31.95]	—	27.14% [26.85; 27.43]	18.92% [18.68; 19.17]	—
KMR-NET-d0.5-m2	42.10% [41.78; 42.41]	33.70% [33.41; 33.99]	—	28.68% [28.38; 28.97]	19.87% [19.62; 20.12]	—
KMR-NET-d0.2-m2	43.68% [43.36; 43.99]	35.18% [34.88; 35.48]	—	29.81% [29.49; 30.12]	20.68% [20.43; 20.94]	—
KMR-LIN-m4	35.68% [35.38; 35.96]	31.00% [30.72; 31.29]	23.10% [22.83; 23.36]	51.72% [51.41; 52.03]	44.61% [44.30; 44.91]	31.36% [31.07; 31.65]
KMR-NET-d0.8-m4	42.24% [41.93; 42.56]	37.66% [37.36; 37.96]	29.42% [29.14; 29.70]	57.62% [57.32; 57.93]	50.66% [50.35; 50.97]	38.57% [38.27; 38.87]
KMR-NET-d0.5-m4	44.83% [44.51; 45.15]	39.86% [39.56; 40.16]	31.21% [30.93; 31.50]	61.05% [60.75; 61.35]	54.10% [53.79; 54.41]	40.77% [40.47; 41.08]
KMR-NET-d0.2-m4	46.86% [46.54; 47.18]	41.76% [41.46; 42.06]	32.87% [32.58; 33.16]	63.28% [62.98; 63.59]	56.12% [55.81; 56.43]	42.43% [42.13; 42.74]
KMgene* (Yan et al., 2018)	31.02% [30.74; 31.31]	—	—	—	—	—

Simulated power to detect an effect of size 0.04 with a test at significance levels of α = 5% is displayed. The simulations are based on 100,000 runs each with 1,000 individuals. Power estimates together with 95% confidence interval are presented for genetic main and time-interaction effects. Phenotype data were either complete or with 25/50% of values missing at random (MAR). Model names correspond with Table 1.

*For comparability, the single gene-level *p*-values of KMgene are combined to a pathway *p*-value using Fisher's method.

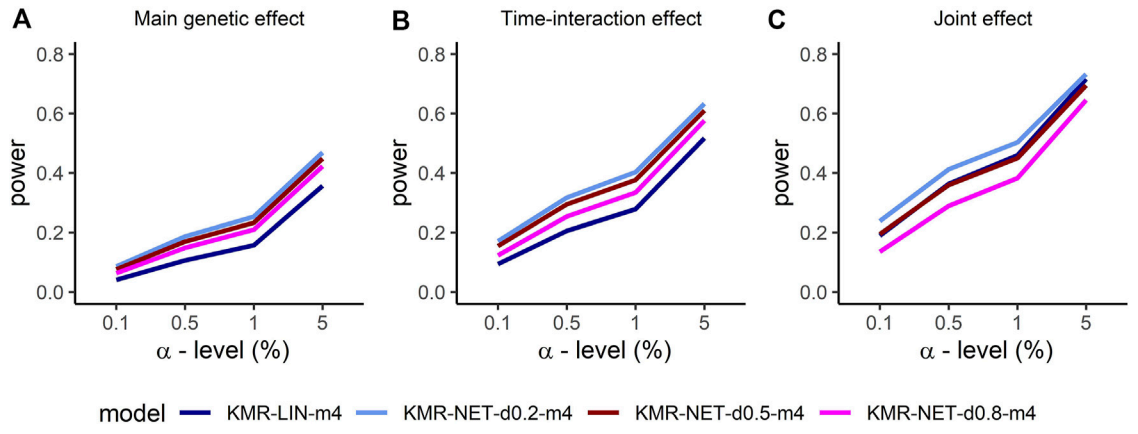


FIGURE 3
Comparison of the power for the four-measurement models (KMR-LIN, KMR-NET-d0.2, KMR-NET-d0.5, KMR-NET-d0.8) with complete phenotype data and effect size β = 0.04 for the different genetic effect models with **(A)** the main genetic **(B)** the time-interaction and **(C)** the joint effect in our simulation study.

Supplementary Table S4). In the pathway analysis, the comparison model KMgene yielded a significantly lower power compared to KMR-LIN-m4 for the same simulation scenario (N = 1,000, m = 4). This effect increased with increasing effect size β.

3.2 Application to the PsyCourse Study

In our real-world data sample, we analyzed 1,518 individuals with at least one TMT-B measurement including 591 "affective," 533 "psychotic," and 394 mentally healthy individuals. The mean

TABLE 5 Power comparison for main, time-interaction and joint effects.

Models	Genetic effect		
	Main genetic effect	Time-interaction effect	Joint effect
KMR-LIN-m4	35.68% [35.38; 35.96]	51.72% [51.41; 52.03]	71.55% [71.27; 71.83]
KMR-NET-d0.8-m4	42.24% [41.93; 42.56]	57.62% [57.32; 57.93]	64.47% [64.16; 64.78]
KMR-NET-d0.5-m4	44.83% [44.51; 45.15]	61.05% [60.75; 61.35]	69.44% [69.14; 69.74]
KMR-NET-d0.2-m4	46.86% [46.54; 47.18]	63.28% [62.98; 63.59]	73.23% [72.95; 73.50]

Simulated power to detect an effect of size 0.04 with a test at significance levels of $\alpha = 5\%$ is displayed. The simulations are based on 100,000 runs each with 1,000 individuals. Power estimates together with 95% confidence interval are presented for genetic main, time-interaction and joint effects. Model names correspond with Table 1. The adjustment for the main genetic effect in the joint genetic effect was performed by adding the top two principal components of a PCA on a main linear kernel.

TABLE 6 Results of the real-data analyses without outlier.

Kernel type	Genetic effect tested	Pathway	*p*-value
Linear kernel (LIN)	Main genetic effect	**Synaptic adhesion-like molecules**	**0.0389**
		NCAM signaling for neurite out-growth	0.0739
		Ion channel transport	0.1029
		Regulation of MECP2 expression activity	0.2101
		MECP2 regulates neuronal receptors and channels	0.2237
Linear kernel (LIN)	Time-interaction effect	**Ion channel transport**	**0.0089**
		MECP2 regulates neuronal receptors and channels	0.2738
		Synaptic adhesion-like molecules	0.3202
		Long-term potentiation	0.3391
		Receptor-type tyrosine-protein phosphatases	0.3579
Network kernel (NET)	Main genetic effect	**Synaptic adhesion-like molecules**	**0.0171**
		NCAM signaling for neurite-growth	**0.0472**
		Signaling by ERBB4	**0.0496**
		Long-term potentiation	0.0910
		MECP2 regulates neuronal receptors and channels	0.1038
Network kernel (NET)	Time-interaction effect	Synaptic adhesion-like molecules	0.2282
		NCAM1 interactions	0.2551
		Long-term potentiation	0.2943
		Neurexins and neuroligins	0.3341
		Trafficking of AMPA receptors	0.4404
Size-adjusted network kernel (ANET)	Main genetic effect	**Synaptic adhesion-like molecules**	**0.0174**
		Signaling by ERBB4	**0.0419**
		NCAM signaling for neurite out-growth	0.0548
		Long-term potentiation	0.0886
		MECP2 regulates neuronal receptors and channels	0.1059
Size-adjusted network kernel (ANET)	Time-interaction effect	Synaptic adhesion-like molecules	0.2429
		NCAM1 interactions	0.2498
		Long-term potentiation	0.2998
		Neurexins and neuroligins	0.3629
		Trafficking of AMPA receptors	0.4866

The five top ranked pathways (according to *p*-value) are listed for each kernel and genetic effect (main genetic and time-interaction effect). Nominal significant (*p*-value <0.05) pathways are printed in bold.

age at the first measurement point was 41 years [sd: 13.8] 48% of the samples were female (for more details see Supplementary Table S5). At all four measurement points, the psychotic group attained the highest TMT-B score; only at measurement points 1 and 3 were the differences for the psychotic and affective groups significant (see CI of Supplementary Figure S1). The control

group demonstrated at each measurement point a significant difference and attained the lowest TMT-B score (Supplementary Figure S1). Previously, we identified a phenotypic outlier, an individual with the highest possible score at each measurement point assessed. Here we focused on the results without the outlier. The results did not change qualitatively when removing the outlier (data not shown).

Thirteen of the 17 tested pathways overlapped at least with one other pathway in at least one gene. The four independent (i.e., pathways not overlapping) were "ion channel transport," "EPH-ephrin signaling," "receptor-type tyrosine-protein phosphatases," and "regulation of MECP2 expression and activity" (Table 2). We did not find any pathways significantly associated with the phenotype TMT-B after multiple testing correction (p-value $<0.0033 = \alpha_{Gao}$) for either applied kernel (LIN, NET, and ANET). However, we identified seven pathways in total as achieving a p-value <0.05, which are represented in bold in Table 6 with the respective kernel used. For example, the "synaptic adhesion-like molecules" pathway is nominally significant for the main genetic effect for ANET, NET, and LIN. The "signaling by ERBB4" pathway, which poses as the foundation of our simulation, was nominally significant with all three kernels when testing the main genetic effect. For the time-interaction effect, we identified only one pathway, "ion channel transport," as nominally significant. This pathway has the smallest p-value of all pathways (0.0089). To compare the different kernels, we ranked all pathways according to their p-values. Table 6 lists the top five pathways for each kernel stratified by the main genetic and time-interaction effects. For both network kernels, we noted a very similar ranking of the top five pathways in the respective genetic effects, whereas for the linear kernel the detected pathways varied between the genetic effect models. Considering the p-value ranking, the "synaptic adhesion-like molecules" pathway stood out as the one with smallest p-value (rank 1) in all analyses.

4 Discussion

Here we present long-KMR, a topology-based pathway analysis method for longitudinal data, which applies kernel machine regression. The methodological basis of long-KMR is presented. To create long-KMR the connection of KMR and LMM are exploited. In addition, we use the network kernel (Freytag et al., 2013) integrating network information into the model. A simulation study is conducted to assess the performance of long-KMR. The models applied in the simulation study are displayed in Table 1. Different aspects are studied, including the influence of the number of measurement points and varying pathway densities. We modeled and tested a main genetic effect and a time-interaction effect for association, the latter testing the association of a pathway with the trajectory of the phenotype TMT-B. Furthermore, we considered an approach to analyze a joint model containing the main genetic effect and the time-interaction effect in a computationally effective way. Lastly, we applied long-KMR to a cognitive phenotype from the PsyCourse Study (Budde et al., 2018).

4.1 Simulation studies

4.1.1 Number of measurements per individual

As expected, the power of long-KMR increases with growing number of measurement points, in particular for the time-interaction effect. This can be traced back to the information that is added to the model at each measurement point, increasing the probability of detecting an effect. Thus, we also identified a larger power loss when analyzing the time-interaction effect with incomplete phenotype data (missing measurements).

4.1.2 Network kernel

The performance of long-KMR improves further when we apply the network kernel instead of the linear kernel, in particular in the single-effect models. We observe that the network kernel has at least the same power as the linear kernel. The power benefit of the network kernel is more pronounced when testing in the presence of smaller genetic effect sizes. For larger genetic effect sizes the power is already extremely high (approx. 98%–99%, Supplementary Table S4), thus the power increase is less noticeable. This power gain is due to the integration of additional pathway information on gene interactions and network topology (Freytag et al., 2013). Here, the topology characteristics of the pathway network play an important role. As the network kernel was developed to exploit the connection of a pathway (Freytag et al., 2013) we studied the influence of the pathway density, identifying a power increase with decreasing pathway density. The higher the density, the more the respective power of network and linear kernel converged. Mathematically, a pathway with many connections (high density) leads to a denser adjacency matrix N, i.e., N contains mainly '1's. Thus, we do not add a lot of specific information when multiplying GA with N (see definition of network kernel). We integrate more noise into the kernel (when N is highly dense) as we sum up the same effects (sum of rows) and only inflate the similarity values artificially (higher range). Thus, we exclude variations and cover potential effects with noise. Consequently, a candidate pathway should preferentially be studied with respect to its characteristics before applying the network kernel when performing long-KMR.

In the joint model including both main and time-interaction effect, the network kernel demonstrated a slightly different performance for different pathway densities. We consider four measurement points only. The network kernel with density 0.2 (lowest density) still has the highest power but only slightly higher when compared to the linear kernel (approx. 72%–73%). The network kernels with densities 0.5 and 0.8 have surprisingly

low power compared to the linear kernel. This phenomenon is perhaps due to the simulation of the genetic effect as purely linear effect and enhanced by the application of two different kernel functions in one model. We simulate the genetic effects in a linear fashion and thus we observe the performance of the network kernel in worst-case scenarios. Nevertheless, the network kernel improved long-KMR slightly when the pathway density is not too high. In general, long-KMR is preferable except when testing a very dense pathway. The latter should be acknowledged and considered when interpreting the results of long-KMR under a specific kernel.

It should also be taken into account that a possible misspecification of a pathway, for example, in the form of wrongly described gene connections leads to an inaccurate pathway topology and pathway characteristics, e.g., density. This can lead to power changes in the analysis. Thus, one of the greatest challenges to topology-based pathway analyses remains the possible inaccuracy and perhaps incompleteness of the studied pathways. Here, future work is required to minimize possible misclassifications. In the future, it would be also worthwhile analyzing other pathway characteristics, e.g., the betweenness centrality or diameter of the pathway, and their influence on power of the long-KMR with the network kernel. However, these aspects should also be considered beforehand in one-measurement settings in order to determine any indication of the performance being affected and thus keep the computational costs associated with the analysis of an extensive longitudinal scenario down to an acceptable level. Additionally, more complex simulation models could be considered, including e.g., genetic effect models in which causal SNP effects interact with each other and the causal SNPs vary between main and interaction effects. Here it is expected that these scenarios are even more advantageous for the network kernel. However, this exceeds the scope of this communication.

4.1.3 Comparison of long-KMR with ANCOVA and KMgene

When comparing the different two-measurement models either for the main effect or for the time-interaction effect, long-KMR has the higher power and is the preferred option, in spite of its longer computation time, in particular when using the network kernel. As expected ANCOVA has lower power. Note that the ANCOVA model only uses the second measurement point as dependent variable (Table 1) and loses information regarding the time effect. For the main genetic effect, we even observed that by applying the network kernel compared to the linear kernel, the power loss resulting from the smaller number of measurement points is reduced.

For the four-measurement models the comparison with KMgene (Yan et al., 2015; Yan et al., 2018) on pathway level reveals that our long-KMR has higher power. In addition, the KMgene type I error rates were slightly inflated (Table 3) for the Fisher method. Thus, we used a second *p*-value combination approach according to Stouffer (Larson et al., 2017), yielding even slightly more inflated *p*-values (data not illustrated). Thus, our approach represents the suitable choice when analyzing a whole pathway. KMgene remains a solid approach when analyzing single genes.

4.2 Application to the PsyCourse Study

In our application, a total of seven pathways were nominally significant (Table 6). Six of the seven pathways were associated with TMT-B when testing for the main effect. We looked more closely at the pathways with a *p*-value <0.1, i.e., "synaptic adhesion-like molecules," "signaling by ERBB4," "long-term potentiation," and "NCAM signaling for neurite growth." The first three pathways contain the gene DLG4. This synaptic gene encodes for the density protein 95 (PSD95) and plays a critical role in the activity regulation of NMDA (N-methyl-D-aspartate) receptors in schizophrenic patients (Cheng et al., 2010; Tian et al., 2017). It is important for learning and memory (Tian et al., 2017) and as a predictor of cognitive deficits (Fan et al., 2018). DLG4 is also part of the complex DLG4-NMDA-DLGAP1, which was associated with influencing executive functions, in particular the set-shifting abilities (cognitive flexibility) in attention deficit hyperactivity disorder individuals (Fan et al., 2018). NMDA receptors, which are highly influenced by DLG4, are important in many neuropsychiatric disorders that have a cognitive flexibility impairment (Fan et al., 2018), e.g., schizophrenia (Cheng et al., 2010). Two other schizophrenia susceptibility genes are NRG1 and ERBB4 (Banerjee et al., 2010; Tian et al., 2017), which are part of the "signaling by ERBB4" and "long-term potentiation" pathway together with DLG4. The signaling pathway of NRG1 and ERBB4 has been identified as influencing the transmission of glutamate and GABA (Banerjee et al., 2010), which are implicated in playing a role in executive functions (Hatoum et al., 2020). NRG1-ERBB4 signaling has also been discussed as a target of gene therapy in adults with neurodevelopmental disorders to reduce cognitive impairment, e.g., in executive functions (Shi and Bergson, 2020). They modulate different synaptic processes, such as long-term potentiation, and are essential for the development of the nervous system (Ledonne and Mercuri, 2019), proper brain function and cognitive processes (Ledonne and Mercuri, 2019). The "long-term potentiation" pathway is also strongly influenced by the above-mentioned NMDA glutamate receptors and is strongly involved with learning and memory processes (Lisman et al., 2012; Lüscher and Malenka, 2012). For the fourth pathway, "NCAM signaling for neurite outgrowth," the neural cell adhesion molecule (NCAM) also plays an important role in the nervous system (Li et al., 2013).

Of the seven pathways nominally significant, "ion channel transport" was the only pathway to prove significant for the time-interaction effect and when modelled with the linear kernel. This pathway had the lowest *p*-value (0.0089). Ion channels are implicated in influencing the susceptibility to or the pathogenesis of psychiatric diseases (Imbrici et al., 2013), and are integral to synaptic functioning.

Conclusion

Overall, we demonstrated that our longitudinal topology-based pathway analysis displays a power gain and a great flexibility to model pathways and genetic effects. Our approach enables the choice between the popular linear kernel and a network kernel that integrates pathway topology information. The latter demonstrated superiority depending on the density of the pathway of interest. The approach is implemented as the R package kalpra, which is available at https://gitlab.gwdg.de/bernadette.wendel/kalpra.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Due to the sensitivity of individual genetic information, the dataset presented in this study are available upon reasonable request. Requests to access these datasets should be directed to the authors.

Ethics statement

The studies involving human participants were reviewed and approved by the respective ethics committees at each study center. The patients/participants provided their written informed consent to participate in this study.

Author contributions

Conceptualization and core writing team: BW, UH, and HB; methodology, analysis tools, and analysis: BW; contribution to analysis tools: MH; data curation: MB, MH, SP, PF, TGS, and UH; writing-original draft: BW; writing-review and editing: all authors. All authors contributed to the article and approved the submitted version.

Funding

TGS and PF are supported by the Deutsche Forschungsgemeinschaft (German Research Foundation; DFG) within the framework of the projects http://www.kfo241.de and http://www.PsyCourse.de (SCHU 1603/4-1, 5-1, 7-1; FA241/16-1). BW and HB are supported by the DFG (KFO241, BI 576 15-1). TGS received additional support from the German Federal Ministry of Education and Research (BMBF) within the framework of the BipoLife network (01EE1404H), IntegraMent (01ZX1614K), e:Med Program (01ZX1614K) and the Lisa Oehler Foundation (Kassel, Germany). TGS was further supported by the grants GEPI-BIOPSY (01EW 2005) and MulioBio (01EW 2009) from ERA-NET Neuron (BMBF). SP was supported by a 2016 NARSAD Young Investigator Grant (25015) from the Brain and Behavior Research Foundation. UH was supported by European Union's Horizon 2020 Research and Innovation Program (PSY-PGx, Grant agreement No. 945151).

Acknowledgments

The authors would like to thank Andrew Entwistle for proofreading the manuscript. Further, we acknowledge support by the Open Access Publication Funds of the Göttingen University.

Conflict of interest

The 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.2022.1015885/full#supplementary-material

References

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467 , 7311, 52–58. doi:10.1038/nature09298

Banerjee, A., MacDonald, M. L., Borgmann-Winter, K. E., and Hahn, C.-G. (2010). Neuregulin 1- erbB4 pathway in schizophrenia: From genes to an interactome. *Brain Res. Bull.* 83, 132–139. doi:10.1016/j.brainresbull.2010.04.011

Bowie, C. R., and Harvey, P. D. (2006). Administration and interpretation of the Trail making test. *Nat. Protoc.* 1 (5), 2277–2281. doi:10.1038/nprot.2006.390

Budde, M., Anderson-Schmidt, H., Gade, K., Reich-Erkelenz, D., Adorjan, K., Kalman, J. L., et al. (2018). A longitudinal approach to biological psychiatric research: The PsyCourse study. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 180 (2), 89–102. doi:10.1002/ajmg.b.32639

Caruana, E. J., Roman, M., Hernández-Sánchez, J., and Solli, P. (2015). Longitudinal studies. *J. Thorac. Dis.* 7, E537–E540. doi:10.3978/j.issn.2072-1439.2015.10.63

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* 4, 7. doi:10.1186/s13742-015-0047-8

Cheng, M.-C., Lu, C.-L., Luu, S.-U., Tsai, S.-U., Hsu, S.-H., Chen, T.-T., et al. (2010). Genetic and functional analysis of the DLG4 gene encoding the post-synaptic density protein 95 in schizophrenia. *PLoS ONE* 5, e15107. doi:10.1371/journal.pone.0015107

Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal. Available: https://igraph.org/.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., et al. (2016). Next-generation genotype imputation service and methods. *Nat. Genet.* 48, 1284–1287. doi:10.1038/ng.3656

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *Appl. Stat.* 29, 323. doi:10.2307/2346911

de Leeuw, C. A., Neale, B. M., Heskes, T., and Posthuma, D. (2016). The statistical properties of gene-set analysis. *Nat. Rev. Genet.* 17, 353–364. doi:10.1038/nrg.2016.29

Dewey, M. (2022). metap: meta-analysis of significance values. R package version 1.8, Diamond, A. (2013 Executive Functions. *Annu. Rev. Psychol.* 64 (1), 135–168. doi:10.1146/annurev-psych-113011-143750

Fan, Z., Qian, Y., Lu, Q., Wang, Y., Chang, S., and Yang, L. (2018). DLGAP1 and NMDA receptor-associated postsynaptic density protein genes influence executive function in attention deficit hyperactivity disorder. *Brain Behav.* 8, e00914. doi:10.1002/brb3.914

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh, Scotland: Oliver & Boyd.

Freytag, S., Bickeböller, H., Amos, C. I., Kneib, T., and Schlather, M. (2012). A novel kernel for correcting size bias in the logistic kernel machine test with an application to rheumatoid arthritis. *Hum. Hered.* 74, 97–108. doi:10.1159/000347188

Freytag, S., Manitz, J., Schlather, M., Kneib, T., Amos, C. I., Risch, A., et al. (2013). A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. *Hum. Hered.* 76, 64–75. doi:10.1159/000357567

Friedman, N. P., Miyake, A., Altamirano, L. J., Corley, R. P., Young, S. E., Rhea, A., et al. (2016). Stability and change in executive function abilities from late adolescence to early adulthood: A longitudinal twin study. *Dev. Psychol.* 52 (2), 326–340. doi:10.1037/dev0000075

Gao, Q., He, Y., Yuan, Z., Zhao, J., Zhang, B., and Xue, F. (2011). Gene- or region-based association study via kernel principal component analysis. *BMC Genet.* 12, 75. doi:10.1186/1471-2156-12-75

Ge, T., Smoller, J. W., and Sabuncu, M. R. (2016). "Kernel machine regression in neuroimaging genetics," in *Machine learning and medical imaging* (Netherlands: Elsevier).

Hatoum, A. S., Morrison, C. L., Mitchell, E. C., Lam, M., Benca-Bachman, C. E., Reineberg, A. E., et al. (2020). Genome-wide association study of over 427,000 individuals establishes executive functioning as a neurocognitive basis of psychiatric disorders influenced by GABAergic processes. bioRxiv [Preprint]. doi:10.1101/674515

Heilbronner, U., Adorjan, K., Anderson-Schmidt, H., Budde, M., Comes, A. L., Gade, K., et al. (2021). The PsyCourse codebook. Version 5.0

Hendricks, A. E., Dupuis, J., Logue, M. W., Myers, R. H., and Lunetta, K. L. (2013). Correction for multiple testing in a gene region. *Eur. J. Hum. Genet.* 22, 414–418. doi:10.1038/ejhg.2013.144

Holmans, P. (2010). Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Adv. Genet.* 72, 141–179. doi:10.1016/B978-0-12-380862-2.00007-2

Imbrici, P., Conte Camerino, D., and Tricarico, D. (2013). Major channels involved in neuropsychiatric disorders and therapeutic perspectives. *Front. Genet.* 4, 76. doi:10.3389/fgene.2013.00076

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2019). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498. doi:10.1093/nar/gkz1031

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). Kegg: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi:10.1093/nar/gkw1092

Larson, N. B., Chen, J., and Schaid, D. J. (2019). A Review of kernel methods for genetic association studies. *Genet. Epidemiol.* 43, 122–136. doi:10.1002/gepi.22180

Larson, N. B., McDonnell, S., Albright, L. C., Teerlink, C., Stanford, J., Ostrander, E. A., et al. (2017). gsSKAT: Rapid gene set analysis and multiple testing correction for rare-variant association studies using weighted linear kernels. *Genet. Epidemiol.* 41, 297–308. doi:10.1002/gepi.22036

Ledonne, A., and Mercuri, N. B. (2019). On the modulatory roles of neuregulins/ErbB signaling on synaptic plasticity. *Int. J. Mol. Sci.* 21, 275. doi:10.3390/ijms21010275

Li, S., Leshchyns'ka, I., Chernyshova, Y., Schachner, M., and Sytnyk, V. (2013). The neural cell adhesion molecule (NCAM) associates with and signals through p21-activated kinase 1 (Pak1). *J. Neurosci.* 33, 790–803. doi:10.1523/JNEUROSCI.1238-12.2013

Lisman, J., Yasuda, R., and Raghavachari, S. (2012). Mechanisms of CaMKII action in long-term potentiation. *Nat. Rev. Neurosci.* 13, 169–182. doi:10.1038/nrn3192

Liu, D., Ghosh, D., and Lin, X. (2008). Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinforma.* 9, 292. doi:10.1186/1471-2105-9-292

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* 63, 1079–1088. doi:10.1111/j.1541-0420.2007.00799.x

Lüscher, C., and Malenka, R. C. (2012). NMDA receptor-dependent long-term potentiation and long-term depression (LTP/LTD). *Cold Spring Harb. Perspect. Biol.* 4, a005710. doi:10.1101/cshperspect.a005710

Malzahn, D., Friedrichs, S., Rosenberger, A., and Bickeböller, H. (2014). Kernel score statistic for dependent data. *BMC Proc.* 8, S41. doi:10.1186/1753-6561-8-S1-S41

Molenberghs, G., and Verbeke, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Mooney, M. A., and Wilmot, B. (2015). Gene set analysis: A step-by-step guide. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 168, 517–527. doi:10.1002/ajmg.b.32328

R Core Team (2021). R: A language and environment for statistical computing. Available: https://www.R-project.org/(Accessed September 3, 2021).

Rodchenkov, I., Babur, O., Luna, A., Aksoy, B. A., Wong, J. V., Fong, D., et al. (2019). Pathway commons 2019 update: Integration analysis and exploration of pathway data. *Nucleic Acids Res.* 48, D489–D497. doi:10.1093/nar/gkz946

Schaid, D. J. (2010). Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Hum. Hered.* 70, 109–131. doi:10.1159/000312641

Schaid, D. J. (2010). Genomic similarity and kernel methods II: Methods for genomic information. *Hum. Hered.* 70, 132–140. doi:10.1159/000312643

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). "Kernel principal component analysis," in *Lecture notes in computer science* (Germany: Springer Berlin Heidelberg).

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* 10, 1299–1319. doi:10.1162/089976698300017467

Shi, L., and Bergson, C. M. (2020). Neuregulin 1: An intriguing therapeutic target for neurodevelopmental disorders. *Transl. Psychiatry* 10, 190. doi:10.1038/s41398-020-00868-5

Smigielski, L., Papiol, S., Theodoridou, A., Heekeren, K., Gerstenberg, M., Wotruba, D., et al. (2021). Polygenic risk scores across the extended psychosis spectrum. *Transl. Psychiatry* 11, 600. doi:10.1038/s41398-021-01720-0

Stern, D. F. (2019). Signaling by ERBB4. Reactome - a curated knowledgebase of biological pathways. *Nucleic Acids Res* 33, D428–32. doi:10.3180/r-hsa-1236394.3

Strauss, E., Sherman, E. M. S., and Spreen, O. (2006). *A compendium of neuropsychological tests - administration, norms, and commentary*. New York: Oxford University Press.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: Simulation of multiple disease SNPs. *Bioinformatics* 27, 2304–2305. doi:10.1093/bioinformatics/btr341

Tian, J., Geng, F., Gao, F., Chen, Y.-H., Liu, J.-H., Wu, J.-L., et al. (2017). Down-regulation of neuregulin1/ErbB4 signaling in the Hippocampus is critical for

learning and memory. *Mol. Neurobiol.* 54, 3976–3987. doi:10.1007/s12035-016-9956-5

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–67. doi:10.18637/jss.v045.i03

Wang, Z., Xu, K., Zhang, X., Wu, X., and Wang, Z. (2016). Longitudinal SNP-set association analysis of quantitative phenotypes. *Genet. Epidemiol.* 41, 81–93. doi:10.1002/gepi.22016

Wendel, B., Papiol, S., Andlauer, T. F. M., Zimmermann, J., Wiltfang, J., Spitzer, C., et al. (2021). A genome-wide association study of the longitudinal course of executive functions. *Transl. Psychiatry* 11, 386. doi:10.1038/s41398-021-01510-8

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* 86, 929–942. doi:10.1016/j.ajhg.2010.05.002

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi:10.1016/j.ajhg.2011.05.029

Yan, Q., Fang, Z., and Chen, W. (2018). KMgene: A unified r package for gene-based association analysis for complex traits. *Bioinformatics* 34, 2144–2146. doi:10.1093/bioinformatics/bty066

Yan, Q., Weeks, D. E., Tiwari, H. K., Yi, N., Zhang, K., Gao, G., et al. (2015). Rare-Variant kernel machine test for longitudinal data from population and family samples. *Hum. Hered.* 80, 126–138. doi:10.1159/000445057

Supplementary Material

1 Supplementary Figures and Tables

1.1 Supplementary Tables

Supplementary Table 1. Run times of exemplary analyses

Models	**Kernel**	
	Linear Kernel	Network kernel
Simulation (N=1000) – analysis of a single pathway		
ANCOVA*	00.55 sec [00.54 sec]	00.97 sec [00.96 sec]
Long-KMR with m=2	08.24 sec [07.98 sec]	15.46 sec [15.18 sec]
Long-KMR with m=4	35.00 sec [22.86 sec]	91.01 sec [89.65 sec]
KMgene** with m=4	08.29 sec [08.05 sec]	-
Real data example (N=1517) – analysis of 17 pathways		
Baseline (m=1)	38.31 sec [37.85 sec]	26.45 sec [26.18 sec]
Longitudinal (m=4)	496.73 sec [488.73 sec]	861.74 sec [852.98 sec]

We provide the run time (=time actually needed to finalize the process), which is required to fit the model of the KMR analysis and to compute the p-value. The time in brackets is the CPU time in seconds (sec). For the simulated data, we determine the computation time for an individual pathway consisting of 19 genes with a pathway density of 0.5 (d0.5). The run times displayed for the real-data example provides information for the analysis of 17 pathways.

*The ANCOVA model is applied as pre/post-analysis (two measurements). Due to the characteristics of the ANCOVA model (see Table 1) it uses only one measurement as dependent variable thus it reduces to a KMR with one measurement (baseline).

**For KMgene, we provide the added computation times relating to the analysis of 19 single genes.

Note that the actual runtime can vary as it also depends on the computing power. When applying real-data, the run time is also largely depending data structure, e.g. the number of genotyped SNPs, size of pathway.

Supplementary Table 2. Type I error rates of the simulation studies.

Models	**Estimated type I error rate (%)**			
	$\alpha = 5\%$	$\alpha = 1\%$	$\alpha = 0.5\%$	$\alpha = 0.1\%$
KMR-LIN-m2_25MAR	4.44	0.78	0.37	0.06
KMR-NET-d0.8-m2_25MAR	4.95	0.99	0.48	0.09
KMR-NET-d0.5-m2_25MAR	4.93	0.99	0.46	0.07
KMR-NET-d0.2-m2_25MAR	4.98	1.00	0.47	0.07
KMR-LIN-m4_25MAR	4.47	0.84	0.40	0.08
KMR-LIN-m4_50MAR	4.38	0.82	0.41	0.09
KMR-NET-d0.2-m4_25MAR	4.95	1.00	0.49	0.09
KMR-NET-d0.2-m4_50MAR	4.93	0.98	0.48	0.09
KMR-NET-d0.5-m4_25MAR	4.94	0.93	0.42	0.07
KMR-NET-d0.5-m4_50MAR	4.88	0.93	0.47	0.08
KMR-NET-d0.8-m4_25MAR	4.88	0.95	0.48	0.10
KMR-NET-d0.8-m4_50MAR	4.86	0.95	0.48	0.09

Simulated type I error for tests at significance levels of $\alpha = 5\%$, 1%, 0.5% and 0.1% are displayed. The simulations are based on 100,000 runs each with 1000 individuals.

Supplementary Table 3. Power results of the simulation study with effect size **0.06**.

	Genetic effect					
	Main genetic effect			Time-interaction effect		
Models	Complete	25MAR	50MAR	Complete	25MAR	50MAR
KMR-LIN-ANCOVA	26.66% [26.39;26.94]	-	-	06.13% [05.98;06.28]	-	-
KMR-LIN-m2	74.01% [73.73;74.28]	59.91% [59.61;60.22]	-	51.26% [50.95;51.58]	11.33% [11.13;11.52]	-
KMR-NET-d0.8-m2	76.45% [76.18;76.72]	65.02% [64.72;65.32]	-	56.80% [56.49;87.11]	39.19% [38.90;39.50]	-
KMR-NET-d0.5-m2	79.80% [79.54;80.05]	68.74% [68.45;69.03]	-	60.28% [59.97;60.59]	41.91% [41.60;42.21]	-
KMR-NET-d0.2-m2	81.53% [81.28;81.77]	70.61% [70.32;70.89]	-	62.13% [61.83;62.44]	43.42% [43.11;43.73]	-
KMR-LIN-m4	77.47% [77.21;77.73]	70.63% [70.35;70.91]	55.52% [55.21;55.83]	92.66% [92.50;92.83]	87.35% [87.14;87.55]	71.08% [70.80;71.36]
KMR-NET-d0.8-m4	79.19% [79.29;79.79]	73.74% [73.46;74.01]	61.16% [60.86;61.47]	92.58% [92.42;92.74]	87.89% [87.68;88.09]	74.73% [74.46;75.00]
KMR-NET-d0.5-m4	82.72% [82.48;82.96]	77.32% [77.06;77.58]	64.84% [64.54;65.14]	94.26% [94.12;94.41]	90.40% [90.21;90.59]	78.26% [78.01;78.52]
KMR-NET-d0.2-m4	84.16% [83.93;84.40]	79.09% [78.83;79.34]	66.79% [66.50;67.09]	95.04% [94.91;95.18]	91.40% [91.22;91.58]	79.99% [79.74;80.24]
KMgene*	66.22% [65.93;66.51]			-	-	-

Simulated power to detect an effect of size 0.06 with a test at significance levels of $\alpha = 5\%$ is displayed. The simulations are based on 100,000 runs each with 1000 individuals. Power estimates together with 95% confidence interval are presented for genetic main and time-interaction effects. Phenotype data were either complete or with 25/50% of values missing at random (MAR). Model names correspond with Table 1. *For comparability, the single gene-level p-values of KMgene are combined to a pathway p-value using Fisher's method.

Supplementary Table 4. Power results of the simulation study with effect size **0.08**.

	Genetic effect					
	Main genetic effect			Time-interaction effect		
Models	Complete	25MAR	50MAR	Complete	25MAR	50MAR
KMR-LIN-ANCOVA	45.50% [48.19;48.81]	-	-	07.54% [07.37;07.70]	-	-
KMR-LIN-m2	96.62% [96.50;96.73]	89.75% [89.57;89.94]	-	82.88% [82.65;83.12]	18.23% [17.99;18.47]	-
KMR-NET-d0.8-m2	96.17% [96.05;96.29]	90.16% [89.97;90.35]	-	84.11% [83.88;84.34]	64.26% [63.96;64.56]	-
KMR-NET-d0.5-m2	97.26% [97.16;97.36]	92.34% [92.17;92.50]	-	86.86% [86.65;87.08]	68.03% [67.74;68.32]	-
KMR-NET-d0.2-m2	97.67% [97.57;97.76]	93.20% [93.04;93.36]	-	88.14% [87.93;88.34]	69.85% [69.56;70.13]	-
KMR-LIN-m4	97.57% [97.47;97.66]	95.30% [95.17;95.43]	86.50% [86.29;86.71]	99.84% [99.82;99.87]	99.40% [99.35;99.45]	95.57% [95.44;95.70]
KMR-NET-d0.8-m4	97.20% [97.10;97.30]	94.93% [94.80;95.07]	94.93% [94.80;95.07]	99.73% [99.70;99.76]	99.14% [99.08;99.20]	95.18% [95.05;95.31]
KMR-NET-d0.5-m4	98.04% [97.95;98.22]	96.29% [96.17;96.41]	90.01% [89.82;90.19]	99.83% [99.80;99.85]	99.43% [99.39;99.48]	96.57% [96.45;96.68]
KMR-NET-d0.2-m4	98.35% [98.27;98.43]	96.86% [96.75;96.97]	91.15% [90.97;91.32]	99.86% [99.84;99.89]	99.54% [99.50;99.58]	97.07% [96.97;97.18]
KMgene*	75.98% [75.71;76.24]			-	-	-

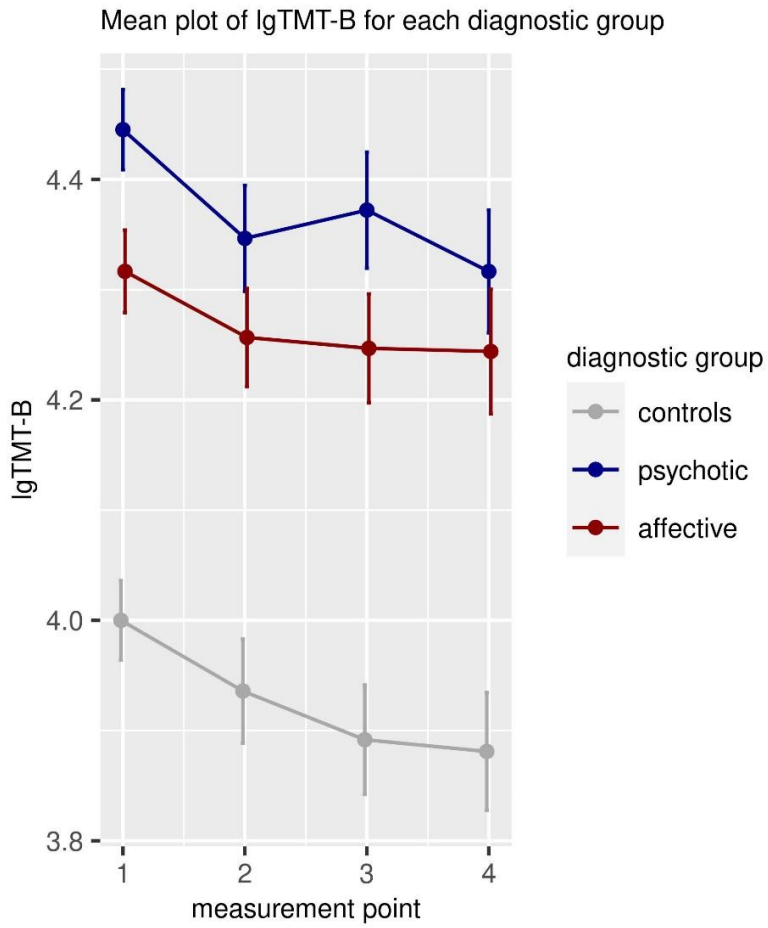
Simulated power to detect an effect of size 0.08 with a test at significance levels of $\alpha = 5\%$ is displayed. The simulations are based on 100,000 runs each with 1000 individuals. Power estimates together with 95% confidence interval are presented for genetic main and time-interaction effects. Phenotype data were either complete or with 25/50% of values missing at random (MAR). Model names correspond with Table 1. *For comparability, the single gene-level p-values of KMgene are combined to a pathway p-value using Fisher's method.

Supplementary Table 5. Phenotype information relating to the first measurement point of the PsyCourse Study

Phenotypes	Diagnostic groups mean (sd) or percentage		
	Affective	Psychotic	Controls
Female	51.4%	37.7%	58.6%
Age	44.9 [13.4]	43.5 [12.0]	36.8 [15.2]
TMT-B	83.5 [42.3]	93.1 [42.8]	58.6 [24.5]
Time effect on lgTMT-B			
β [95% CI]	0.96 [0.95;0.97]	0.95 [0.94;0.97]	0.96 [0.95;0.97]
p-value	4.93×10^{-11}	1.16×10^{-13}	8.62×10^{-15}

The mean and standard deviation (sd) of the age at first measurement and the TMT-B for each diagnostic group are provided. The LMM results testing the time effect on lgTMT-B within each diagnostic group are displayed. The effect estimates β of lgTMT-B are transformed back to their original scale.

1.2 Supplementary Figures



Supplementary Figure 1. Longitudinal course of lgTMT-B score (time in seconds) for each diagnostic group (affective, psychotic and controls). Displayed are means with 95% CI for each measurement point 1,2,3,4, approximately 6 months apart.

Package 'kalpra'

December 11, 2022

Title Kernel Approach for Longitudinal Pathway Regression Analysis

Version 1.0.0

Author Bernadette Wendel [aut], Markus Heidenreich [ctb], Heike Bickeböller [ctb]

Maintainer Bernadette Wendel <bernadette.wendel@med.uni-goettingen.de>

Description Perform a pathway analysis with a kernel machine regression. The pathway can be tested for association with a longitudinal quantitative, a cross-sectional binary or a cross-sectional quantitative phenotype.
Different kernels are available: linear kernel, quadratic kernel and network-based kernel. The latter integrates network information.
The pathway information can be downloaded from the Pathway Commons database.

License GPL (>= 2)

Encoding UTF-8

LazyData true

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

Depends R (>= 2.10)

Imports biomaRt, CompQuadForm, dplyr, gtools, igraph, matlab, Matrix, matrixcalc, methods, nlme, paxtoolsr, plyr, XML

Suggests testthat (>= 3.0.0)

Config/testthat/edition 3

R topics documented:

`adjustAnnotation` *Create size adjusted annotation matrix*

Description

The function deletes empty genes (= no SNPs mapped) if parameter `removeGene=TRUE` and if parameter `removeSNP=TRUE` it also deletes not mapped SNPs. The annotation matrix can also be size adjusted for each gene if `sizeadjusted=TRUE`.

Usage

```
adjustAnnotation(
  Annotationmatrix,
  sizeadjusted = FALSE,
  removeSNP = FALSE,
  removeGene = FALSE
)
```

Arguments

`Annotationmatrix`	an annotation matrix of a pathway
`sizeadjusted`	logical variable, size adjustment of the annotation matrix, default: `FALSE`
`removeSNP`	logical variable, SNPs which are not mapped to any gene are maintained or deleted, default: `FALSE` (see below)
`removeGene`	logical variable, empty genes (no SNPs mapped) are maintained or deleted, default: `FALSE` (see below)

Value

an annotation matrix without empty genes (genes without SNPs assigned)

Note

If the parameter `removeSNP=TRUE`, SNPs that are not mapped are deleted and the rows are removed. The dimension of the annotation matrix changes and thus, the dimensions of the genotype matrix need to be changed accordingly. This reduction needs to be performed by the user before computing the kernel matrix. This can be performed by matching the row names of the reduced annotation matrix (`removeSNP=TRUE`) with the column names of the genotype matrix. If the empty genes are deleted by `removeGene=TRUE` before creating the adjacency matrix, the rewiring of the adjacency matrix in regards to emtpy genes are incorrect. This steps should be performed after creating the adjacency matrix.

Author(s)

Bernadette Wendel

References

For details on the size adjustment see:

- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

Examples

```
Annotationmatrix1<-create.anno(GeneInformation, SNPBimFile, windowsize=20,method="position")
Annotation.adjusted<-adjustAnnotation(Annotationmatrix1)
Annotation.notadjusted<-adjustAnnotation(Annotationmatrix1, sizeadjusted = FALSE)
```

`CrossPhenotype` *Example cross-sectional phenotype (binary and quantitative)*

Description

The data set `CrossPhenotype` contains a simulated normally distributed phenotype and a binary phenotype for 250 individuals for one measurement point. The data contain columns with a complete phenotype and a phenotype with missing values for each phenotype (in total: 80x`NA` (quantitative) and 69x`NA` (binary), see below). In addition to the phenotype, we also simulated two types of covariables (a binary and a continuous variable).

Usage

```
CrossPhenotype
```

Format

A data frame with 250 rows and 7 columns:

IID Identifier of each individual, pattern: IID*n* with *n*=1,...,250

Age normally distributed variable, representing the age of an individual

Gender a binary (0 or 1) variable, representing the gender of an individual

Quantpheno quantitative phenotype

Quantmisspheno quantitative phenotype with 80 missing measurement points

Binpheno binary phenotype

Binmisspheno binary phenotype with 69 missing measurement points

Source

Simulated data set

`download.geneinfo` *Download gene information*

Description

The function downloads information of the genes part of the pathway studied. It uses the package **biomaRt**. The database and data set used to download the information is predefined with 'Ensembl'and 'hsapiens_gene_ensembl', respectively (cannot be changed by the user). It returns the following information: HGNC symbols, chromosome, start position and stop position.

Usage

```
download.geneinfo(pathway)
```

Arguments

pathway	list of one or multiple pathway(s) in SIF (list of data frames)

Value

a dataframe or a list with dataframes of gene information for each pathway (depends on input)

Author(s)

Bernadette Wendel

Examples

```
PathwaysURI<-searchPathway("neuronal system","reactome")
head(PathwaysURI);dim(PathwaysURI)

selectedPathway<-subset(PathwaysURI,name=="Transcriptional Regulation by MECP2")
selectedPathway
Pathway1<-get.pathway(selectedPathway, URI="uri", pathwayname="name", delete =TRUE)
geneInfo<-download.geneinfo(Pathway1)
head(geneInfo)

pathwayList<-get.pathway(PathwaysURI[c(2:7,10:12),],delete = FALSE)
class(pathwayList);length(pathwayList)

geneInfoList<-download.geneinfo(pathwayList)
class(geneInfoList)
```

GeneInformation *Example of gene information data (20 genes)*

Description

The data set GeneInformation is a data frame with simulated gene information for 20 genes of a simulated pathway.

Usage

```
GeneInformation
```

Format

A data frame with 20 rows and 4 columns:

hgnc_symbol hypothetical HGNC symbol of the gene

CHR simulated chromosome of the gene

startBP simulated start base pair position of the gene

endBP simulated end base pair position of the gene

Source

Simulated data set

Genotypematrix *Example genotypes for 250 individuals*

Description

Genotypematrix is a matrix with simulated genotypes for 250 individuals and 2000 SNPs. The genotype matrix has no missing values and contains only three values: 0,1, or 2 representing the number of minor alleles.

Usage

```
Genotypematrix
```

Format

A data frame with 250 rows and 2000 columns:

row each row contains the genotype for one individual

rownames contains the identifier of the individuals

column each column holds information for one SNP

columnnames contains the names of the SNPs

Source

Simulated data set

GenotypematrixInteraction *Example genotypes for 100 individuals*

Description

GenotypematrixInteraction is a matrix with simulated genotypes for 100 individuals and 2000 SNPs. The genotype matrix has no missing values and contains only three values: 0, 1, or 2 representing the number of minor alleles.

Usage

```
GenotypematrixInteraction
```

Format

A data frame with 100 rows and 2000 columns:

row each row contains the genotype for one individual

rownames contains the identifier of the individuals

column each column holds information for one SNP

columnnames contains the names of the SNPs

Source

Simulated data set

get.Annotation *Create an annotation matrix from pathway information*

Description

This function creates an annotation matrix containing information on which SNP is assigned to which gene of a pathway (0 or 1, if sizeadjusted=FALSE).

Usage

```
get.Annotation(
  pathway,
  SNPinfo,
  method = c("position", "LD"),
  SNP = "SNP",
  chrSNP = "CHR",
  bp = "BP",
  bp.2 = "BP2",
  gene = "hgnc_symbol",
  chrgene = "chromosome_name",
  startgene = "start_position",
  endgene = "end_position",
```

```
  windowsize = NULL,
  geneinfo = FALSE,
  sizeadjusted = FALSE,
  removeSNP = FALSE,
  removeGene = FALSE
)
```

Arguments

pathway	a single pathway in SIF (if geneinfo=FALSE) or a data frame with at least four columns with gene information (if geneinfo=TRUE)
SNPinfo	file contains information on SNPs, either *'.bim'* file or *'.blocks.det'* file from PLINK, for 'position' or 'LD', respectively
method	method to assign a SNP to a gene, either 'position' or 'LD'
SNP	column of SNP name ('position') or names of the SNPs of one LD block ('LD') in SNPinfo, default: 'SNP'
chrSNP	column in SNPinfo of chromosome of SNP, default: 'CHR'
bp	column in SNPinfo of base pair position (if method=position) or start base pair of LD block (if method=LD), default: 'BP'
bp.2	column in SNPinfo of end base pair of the LD block, only when method=LD, default: 'BP2'
gene	column in pathway (if geneinfo=TRUE) with gene name, default: 'hgnc_symbol'
chrgene	column in pathway (if geneinfo=TRUE) with chromosome of gene, default: 'chromosome_name'
startgene	column in pathway (if geneinfo=TRUE) with gene start, default: 'start_position'
endgene	column in pathway (if geneinfo=TRUE) with gene end, default: 'end_position'
windowsize	size of mapping window for SNPs, state size in kbp, default: 0
geneinfo	logical variable, if TRUE parameter pathway needs to contain information on genes in form of a data frame with gene name, chromosome, start- and end base pair, default:FALSE
sizeadjusted	logical variable, size adjustment of the annotation matrix, default: FALSE (see below)
removeSNP	logical variable, SNPs which are not mapped to any gene are maintained or deleted, default: FALSE (see below)
removeGene	logical variable, empty genes (no SNPs mapped) are maintained or deleted, default: FALSE (see below)

Details

If the parameter geneinfo=FALSE (default) then the information about the genes in the pathway presented in parameter pathway are downloaded by the function download.geneinfo().

If the parameter sizeadjusted=TRUE, the annotation matrix is size adjusted (see Freytag et al.) and the values of the annotation matrix values are <1.

If the parameter removeSNP=TRUE, SNPs which are not mapped to any gene are deleted and the rows are removed. The dimension of the annotation matrix changes and thus, the dimensions of the genotype matrix need to be changed accordingly.This reduction needs to be performed by the user before computing the kernel matrix. This can be performed by matching the row names of the reduced annotation matrix (removeSNP=TRUE) with the column names of the genotype matrix.

If the parameter `removeGene=TRUE` empty genes are deleted. If this step is performed before creating the adjacency matrix, the rewiring of the adjacency matrix in regards to empty genes are incorrect. This steps should be performed **after** creating the adjacency matrix.

Value

a single annotation matrix

Author(s)

Bernadette Wendel

References

For more details on PLINK see https://www.cog-genomics.org/plink/ and

- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JL: Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC: PLINK: a toolset for whole-genome assocation and population-based linkage analysis. Am J Hum Genet 2017, 81.

For details on size-adjustment see:

- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

Examples

```
Annotationmatrixfinal<-get.Annotation(pathway = GeneInformation, SNPinfo = SNPBimFile, method="position",
windowsize = 50, geneinfo = TRUE)
```

---

get.networkadjacency *Create adjacency matrix of a pathway*

Description

This function computes an adjacency matrix of the pathway and rewires the network according to Freytag et al., e.g. elements which are not genes are deleted and empty genes (no SNPs assigned) are deleted (if annotation matrix is provided). The neighbors of the deleted genes are connected, i.e. edges are added between nodes (a direct path) where an indirect path was before. We assume *undirected* graphs in all cases.

Usage

```
get.networkadjacency(pathway, Annotationmatrix = NULL, signed = FALSE)
```

Arguments

pathway	a *single* pathway dataframe or a list of pathway data frames
Annotationmatrix	a *single* pathway annotation matrix or a list of annotation matrices, default: NULL (see below)
signed	logical variable, sets weights according to 'activation'|'inhibition', default: FALSE

Details

If Annotationmatrix=NULL the adjacency matrix is created but not rewired. If the adjacency matrix should be rewired the annotation matrix provided needs to contain all genes (including empty genes).

Value

an adjacency matrix of a pathway or a list of adjacency matrices (depends on input format)

Author(s)

Bernadette Wendel and Markus Heidenreich

References

For details on the rewiring of a network:

- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

Examples

```
AdjacencyMatrix<-get.networkadjacency(Pathway, Annotationmatrix = NULL, signed = FALSE)

Annotationmatrixfinal<-get.Annotation(pathway = GeneInformation, SNPinfo = SNPBimFile,
method="position", windowsize = 50, geneinfo = TRUE)
AdjacencyMatrix1<-get.networkadjacency(Pathway, Annotationmatrix =Annotationmatrixfinal,
                signed=FALSE)
AdjacencyMatrix1
```

---

get.pathway *Download pathway in SIF*

Description

This function downloads one or multiple pathways from the Pathway Commons database. The pathway(s) is/are in SIF (Standard Interchange Format).

Usage

```
get.pathway(pathwaydataframe, URI = "uri", pathwayname = "name", delete = TRUE)
```

Arguments

pathwaydataframe	data frame with at least two columns: URI and pathway name(s)
URI	name of column yielding the URI of the pathway(s)
pathwayname	name of the column with the pathway name(s)
delete	logical variable, deletes empty pathways (no genes), default: TRUE

Value

a dataframe with one pathway or a list of dataframe where each pathway is in SIF and accessable by the pathway name

Author(s)

Bernadette Wendel

References

For more details on Pathway Commons database

- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research 2010, 39(Database):D685-D690.
- Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Research 2019

Examples

```
PathwaysURI<-searchPathway("neuronal system","reactome")
head(PathwaysURI);dim(PathwaysURI)

selectedPathway<-subset(PathwaysURI,name=="Transcriptional Regulation by MECP2")
selectedPathway
Pathway1<-get.pathway(selectedPathway, URI="uri", pathwayname="name", delete =TRUE)

pathwayList<-get.pathway(PathwaysURI[c(2:7,10:12),],delete = FALSE)
class(pathwayList);length(pathwayList)
```

KMR.Cross.bin — *Kernel machine regression (KMR) analysis for cross-sectional binary data*

Description

This function performs the null model fitting and p-value computation when the phenotype is binary and cross-sectional. Assumption: binomial distribution

Usage

```
KMR.Cross.bin(
  formula,
  phenotypedata,
  kernelmatrix,
  phenotype = "pheno",
  method = c("satt", "davies", "all"),
  lim = NULL,
  acc = NULL
)
```

Arguments

formula	formula of null model
phenotypedata	data frame with phenotype (do not delete rows with missing values) and covariables
kernelmatrix	kernel matrix - either a matrix or a list of matrices
phenotype	name of the column with phenotype, default: 'pheno'
method	method of p-value computation, possibilities: 'satt' for the Satterthwaite approximation, 'davies' for the Davies' algorithm or 'all' for both available methods
lim	maximum number of integration terms (more details see package **CompQuadForm** function davies). Values range from 1,000 (procedure called repeatedly) to 50,000 (procedure called only occasionally)
acc	error bound (more details see package **CompQuadForm** function davies). Suitable values for 'acc' range from 0.001 to 0.00005'

Details

The p-value computation can be performed with two different methods by option method. If parameter is 'davies' the p-value is computed as described by Davies. The function davies() from the package **CompQuadForm** is applied. The parameter 'lim' and 'acc' are only important for davies() and ignored otherwise. For parameter 'satt' the p-value is computed by using the Satterthwaite approximation as described by Schaid.

Value

Outcome object with the results of the kernel regression

Author(s)

Bernadette Wendel

References

For details on the variance component test

- Liu D, Ghosh D, Lin X: Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC Biometrics 2008, 9(1).

- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet 2011, 89:82-93.
- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.
- Ge T, Schmoller JW, Sabuncu MR: Kernel machine regression in neuroimaging genetics. In Machine Learning and Medical Imaging. page 31-68. Elsevier 2016.

The details on the p-value computation methods can be found in:

- Davies RB: Algorithm AS 155: The Distribution of a Linear Combination of chi square Random Variables. Applied Statistics 1980, 29(3):323.
- Schaid DJ: Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum Hered 2019, 70:109-131.

Examples

```
LinKernel<-Linearkernel(Genotypematrix)

crossbin<-KMR.Cross.bin(formula= Binpheno~Age+Gender, phenotypedata = CrossPhenotype,
                        kernelmatrix = LinKernel,
                        method="all", phenotype = "Binpheno")
summaryKMR(crossbin)
```

`KMR.Cross.quan` *Kernel machine regression (KMR) analysis for cross-sectional quantitative data*

Description

This function performs the null model fitting and p-value computation when the phenotype is quantitative and cross-sectional. Assumption: normal distribution

Usage

```
KMR.Cross.quan(
  formula,
  phenotypedata,
  kernelmatrix,
  phenotype = "pheno",
  method = c("satt", "davies", "all"),
  lim = NULL,
  acc = NULL
)
```

Arguments

formula	formula of null model
phenotypedata	data frame with phenotype (do not delete rows with missing values) and covariables
kernelmatrix	kernel matrix - either a matrix or a list of matrices
phenotype	name of the column with phenotype, default: 'pheno'
method	method of p-value computation, possibilities: 'satt' for the Satterthwaite approximation, 'davies' for the Davies' algorithm or 'all' for both available methods
lim	maximum number of integration terms (more details see package **CompQuadForm** function davies). Values range from 1,000 (procedure called repeatedly) to 50,000 (procedure called only occasionally)
acc	error bound (more details see package **CompQuadForm** function davies). Suitable values for 'acc' range from 0.001 to 0.00005'

Details

The p-value computation can be performed with two different methods by option method. If parameter is 'davies' the p-value is computed as described by Davies. The function davies() from the package **CompQuadForm** is applied. The parameter 'lim' and 'acc' are only important for davies() and ignored otherwise. For parameter 'satt' the p-value is computed by using the Satterthwaite approximation as described by Schaid.

Value

Outcome object with the results of the kernel regression

Author(s)

Bernadette Wendel

References

For details on the variance component test

- Liu D, Lin X, Ghosh D: Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. Biometrics 2007, 63(4):1079-1088.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet 2011, 89:82-93.
- Ge T, Schmoller JW, Sabuncu MR: Kernel machine regression in neuroimaging genetics. In Machine Learning and Medical Imaging. page 31-68. Elsevier 2016.

The details on the p-value computation methods can be found in:

- Davies RB: Algorithm AS 155: The Distribution of a Linear Combination of chi square Random Variables. Applied Statistics 1980, 29(3):323.
- Schaid DJ: Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum Hered 2019, 70:109-131.

Examples

```
QuadKernel<-Quadratickernel(Genotypematrix)

crossquant<-KMR.Cross.quan(formula = Quantpheno~Age, phenotypedata = CrossPhenotype,
                           kernelmatrix = QuadKernel,
                           method="all",phenotype = "Quantpheno")
summaryKMR(crossquant)
```

`KMR.KernelTime` *Kernel machine regression (KMR) with Kernel x Time interaction term and adjustment*

Description

This function allows the user to test for an interaction effect between the time factor and the kernel matrix. There are two possible random effect structures available (see below for more details).

Usage

```
KMR.KernelTime(
  fixedEff,
  phenotypedata,
  kernelmatrix,
  kerneltime,
  phenotype = "pheno",
  ID = "IID",
  timepoints = "time",
  estimation = c("REML", "ML"),
  method = c("satt", "davies"),
  randomEff = NULL,
  lim = NULL,
  acc = NULL
)
```

Arguments

`fixedEff`	formula of fixed effects
`phenotypedata`	data frame with phenotype and covariable
`kernelmatrix`	kernel matrix
`kerneltime`	kernel matrix of time interaction effect, i.e. genotypes are multiplied by time and used to compute kernel
`phenotype`	name of the column with phenotype, default: 'pheno'
`ID`	column with individual identifier, default: 'IID'
`timepoints`	column of time vector, default: 'time'
`estimation`	method of estimation, possibilities: 'REML' or 'ML'
`method`	method of p-value computation, possibilities: 'satt' for the Satterthwaite approximation, 'davies' for the Davies' algorithm or 'all' for both available methods

randomEff	formula of random effects, where only two options are possible, default: NULL, second possibility: 'slope'(see below)
lim	maximum number of integration terms (more details see package **CompQuadForm** function davies). values range from 1,000 (procedure called repeatedly) to 50,000 (procedure called only occasionally)
acc	error bound (more details see package **CompQuadForm** function davies). Suitable values for 'acc' range from 0.001 to 0.00005

Details

The p-value computation can be performed with two different methods by option method. If parameter is 'davies' the p-value is computed as described by Davies. The function davies() from the package **CompQuadForm** is applied. The parameter 'lim' and 'acc' are only important for davies() and ignored otherwise. For parameter 'satt' the p-value is computed by using the Satterthwaite approximation as described by Schaid.

Value

Outcome object with the results of the kernel regression

Note

There are only two possible random effect structures:

- The kernel matrix and a random intercept for correction of the dependence structure between the individuals ('randomEff=NULL')
- The kernel matrix and random intercept and a random time slope ('randomEff=slope')

Author(s)

Bernadette Wendel

References

For details on the extension see:

- Yan Q, Weeks DE, Tiwari HK, Yi N, Thang K, Gao G, Lin W-Y, Lou X-Y, Chen W, Liu N: Rare-Variant Kernel Machine Test for Longitudinal Data from Population and Family Samples. Human Heredity 2015, 80(3):126-138.

The details on the p-value computation methods can be found in:

- Davies RB: Algorithm AS 155: The Distribution of a Linear Combination of chi square Random Variables. Applied Statistics 1980, 29(3):323.
- Schaid DJ: Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum Hered 2019, 70:109-131.

Examples

```
WidePhenotypeLong<-long.format(WidePhenotype, i=4,j=6,m=3, columnmeasurement = "pheno")

longGenotype<-makeLongGenotype(GenotypematrixInteraction,m=3)
Genotypetime2<-makeLongGenotype(GenotypematrixInteraction,m=3,timevector = (0:2))
```

```
Kernelmatrix<-Linearkernel(longGenotype)
Kerneltime<-Linearkernel(Genotypetime2)

Interaction<-KMR.KernelTime(fixedEff=pheno~Age+Gender+time,WidePhenotypeLong,
                            kernelmatrix=Kernelmatrix,Kerneltime,
                            estimation="REML",method="all",randomEff = "slope")
summaryKMR(Interaction)
```

KMR.Long *Longitudinal kernel machine regression (KMR) analysis*

Description

The function performs the null model fitting and p-value computation for a longitudinal quantitative phenotype. Assumption: normal distribution.

Usage

```
KMR.Long(
  fixedEff,
  randomeffects,
  phenotypedata,
  kernelmatrix,
  phenotype = "pheno",
  ID = "IID",
  timepoints = "time",
  estimation = c("REML", "ML"),
  method = c("satt", "davies", "all"),
  lim = NULL,
  acc = NULL
)
```

Arguments

fixedEff	formula of fixed effects
randomeffects	formula of random effects (see below)
phenotypedata	data frame with phenotype (do not delete rows with missing values) and covariables
kernelmatrix	kernel matrix
phenotype	name of the column with phenotype, default: 'pheno'
ID	name of the column containing the identifier of the individuals, default: 'IID'
timepoints	name of the column containing the time, default: 'time'
estimation	method of estimation, possibilities: 'REML' or 'ML'
method	method of p-value computation, possibilities: 'satt' for the Satterthwaite approximation, 'davies' for the Davies' algorithm or 'all' for both available methods
lim	maximum number of integration terms (more details see package **CompQuadForm** function davies). Values range from 1,000 (procedure called repeatedly) to 50,000 (procedure called only occasionally)

acc — error bound (more details see package **CompQuadForm** function davies). Suitable values for 'acc' range from 0.001 to 0.00005

Details

The p-value computation can be performed with two different methods by option method. If parameter is 'davies' the p-value is computed as described by Davies. The function davies() from the package **CompQuadForm** is applied. The parameter 'lim' and 'acc' are only important for davies() and ignored otherwise. For parameter 'satt' the p-value is computed by using the Satterthwaite approximation as described by Schaid.

The random effects can be either a single random intercept or a random intercept and a random time slope to correct for the dependence of the longitudinal data. A more complex random effect structure is not possible.

The kernel matrix can model a main genetic effect or a time genetic interaction effect. The kernel need to be computed accordingly with makeLongGenotype().

Value

Outcome object with the results of the kernel regression

Author(s)

Bernadette Wendel

References

For detail on the expansion to longitudinal data:

- Yan Q, Weeks DE, Tiwari HK, Yi N, Thang K, Gao G, Lin W-Y, Lou X-Y, Chen W, Liu N: Rare-Variant Kernel Machine Test for Longitudinal Data from Population and Family Samples. Human Heredity 2015, 80(3):126-138.
- Ge T, Schmoller JW, Sabuncu MR: Kernel machine regression in neuroimaging genetics. In Machine Learning and Medical Imaging. page 31-68. Elsevier 2016.

The details on the p-value computation methods can be found in:

- Davies RB: Algorithm AS 155: The Distribution of a Linear Combination of chi square Random Variables. Applied Statistics 1980, 29(3):323.
- Schaid DJ: Genomic Similarity and Kernel Methods I: Advancements by Building on Mathematical and Statistical Foundations. Hum Hered 2019, 70:109-131.

Examples

```
longGenotype<-makeLongGenotype(Genotypematrix,m=4)
Kernellong<-Linearkernel(longGenotype)

longcompleteKMR<-KMR.Long(fixedEff = pheno~Age+Gender+time, randomeffects = ~1+time|IID,
                          phenotypedata = LongPhenotype,kernel = Kernellong,
                          method = "all",estimation = "REML")
summaryKMR(longcompleteKMR)
```

`Linearkernel` *Computation of a linear kernel*

Description

This function computes a linear kernel.

Usage

```
Linearkernel(Genotype)
```

Arguments

`Genotype`	a single *nxp* matrix of the genotype of *n* individuals and *p* SNPs (no missing values)

Value

an *nxn* kernel matrix (*n* - number of individuals)

Author(s)

Bernadette Wendel

References

For details on the linear kernel

- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X: Powerful SNP-Set Analysis for Case-Control Genome-Wide Association Studies. Am J Hum Genet 2010, 86:929-42.

Examples

```
KernelLin<-Linearkernel(Genotypematrix)#for cross-sectional data

longGenotype<-makeLongGenotype(Genotypematrix, m=4)
Kernellong<-Linearkernel(longGenotype)
```

LongPhenotype *Example quantitative longitudinal phenotype for 250 individuals*

Description

The data set LongPhenotype contains a simulated normally distributed phenotype for 250 individuals at four measurement points. The data are saved in long format. The data have a column with a complete phenotype and one column of a phenotype with missing values (in total: 294xNA, see below). In addition to the phenotype, we also simulated two types of covariables (a binary and a continuous variable).

Usage

```
LongPhenotype
```

Format

A data frame with 1000 rows and 6 columns:

IID Identifier of each individual, pattern: IID*n* with *n*=1,...,250

time a categorical variable representing the time factor: 0,1,2,3 (in total: four measurement points)

Age normally distributed variable, representing the age of an individual

Gender a binary (0 or 1) variable, representing the gender of an individual

pheno quantitative phenotype, simulated with a linear mixed model

missingpheno phenotype values with 294 missing measurement points

Source

Simulated data set

makeLongGenotype *Create genotype matrix in long format and computed interaction effect*

Description

This function transforms a genotype matrix from wide to long format if required. If the genetic time interaction effect is of interest, a time factor need to be provided by timevector and the genotypes are multiplied with the time.

Usage

```
makeLongGenotype(Genotype, m, ID = NULL, timevector = NULL)
```

Arguments

Genotype	genotype matrix in wide format (each row one individual)
m	number of time points
ID	name of column with individual identifiers, if default NULL row names are used
timevector	vector containing time points (e.g. 0,1,...) only required if the genetic time interaction is of interest, default: NULL

Value

genotype matrix in long format (*m* rows for each individual)

Author(s)

Bernadette Wendel

Examples

```
longGenotype<-makeLongGenotype(Genotypematrix, m=4)
dim(longGenotype)

longGenotypeInteraction<-makeLongGenotype(Genotypematrix, m=4, timevector=c(0,1,2,3))
dim(longGenotypeInteraction)
```

Networkkernel *Computation of a network kernel based on Freytag et al.*

Description

This function creates a network kernel matrix based on Freytag et al.. The annotation matrix and the adjacency matrix (both created previously) do not have to be sorted. This function will order the row/column names of the matrices before computing the network kernel. It will also delete empty genes from the annotation matrix. The adjacency matrix need to be already rewired (i.e. empty genes and elements which are no genes are already removed).

Usage

```
Networkkernel(Genotype, Annotation, N)
```

Arguments

`Genotype`	a single *nxp* matrix of the genotype of *n* individuals and *p* SNPs (no missing values)
`Annotation`	is an annotation matrix or a list of annotation matrices (order of annotation matrices identical to order of adjacency matrices (`N`), pathways need to match)
`N`	an adjacency matrix of a rewired pathway or a list of rewired adjacency matrices (order of adjacency matrices identical to order of annotation matrices (`Annotation`), pathways need to match)

Value

an *nxn* kernel matrix (*n* - number of individuals) or a list of kernel matrices depending on the input format of `Annotation` and `N`

Author(s)

Bernadette Wendel

References

For details on the network kernel

- Freytag S, Manitz J, Schlather M, Kneib T, Amos CI, Risch A, Chang-Claude J, Heinrich J, Bickeboeller H: A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum Hered. 2013, 76(2):64-75.

Examples

```
Annotationmatrixfinal<-get.Annotation(pathway = GeneInformation, SNPinfo = SNPBimFile,
                                      windowsize = 50, geneinfo = TRUE)
AdjacencyMatrix1<-get.networkadjacency(Pathway,Annotationmatrix = Annotationmatrixfinal,
                                       signed=FALSE)
NetworkKernel<-Networkkernel(Genotype = Genotypematrix,Annotation=Annotationmatrixfinal,
                             N=AdjacencyMatrix1)
```

Outcome-class *Object class storing the results of the kernel regression analysis*

Description

Object class storing the results of the kernel regression analysis

Slots

`inputformat` character string with general information

`fixedeffects` formula describing the fixed effects of the model

`randomeffects` information of the random effect structure

`estimationmethod` character string determining the method applied to estimate the parameter (`ML` or `REML`)

`pvalueinfo` character string on type of p-value computation method (`davies` or `satt`)

`resultdf` data frame containing the results of the kernel analysis

Author(s)

Bernadette Wendel

Pathway *Example pathway*

Description

The data set `Pathway` is a data frame with an hypothetical pathway of 42 interactions.

Usage

```
Pathway
```

Format

A data frame with 42 rows and 3 columns:

PARTICIPANT_A gene which interacts with gene in column 3

INTERACTION_TYPE type of gene interaction

PARTICIPANT_B gene which interacts with gene in column 1

Source

Simulated data set

pathway.characteristics *Determine graph theoretical characteristics of a pathway graph*

Description

This function determines specific graph theoretical characteristics of a pathway. The following features are assessed: number of nodes and edges, average node degree, density, diameter, betweenness and transitivity. All of the characteristics are computed with functions of the **igraph** package.

Usage

```
pathway.characteristics(pathway)
```

Arguments

`pathway`	a single graph in SIF (Standard Interchange Format) or a single adjacency matrix.

Value

a data frame of the pathway characteristics.

Author(s)

Bernadette Wendel

Examples

```
pathwayinfo<-pathway.characteristics(Pathway)
pathwayinfo
```

`pathway.plot` *Graphical illustration of a pathway*

Description

This function allows to plot a pathway.

Usage

```
pathway.plot(
  pathway,
  direction = "undirected",
  vertexcolor = "darkblue",
  vertexframe = vertexcolor,
  labeldist = -2.5,
  labelcex = 0.8,
  labelcolor = "black",
  labelfont = 2,
  edgewidth = 2.5,
  edgecol = "black",
  curved = FALSE,
  rescale = TRUE,
  pathwayname = NULL,
  title = NULL,
  vertexname = NULL,
  layout = NULL,
  weighted = NULL
)
```

Arguments

`pathway`	single pathway in a SIF format or an **igraph** object
`direction`	select between '`undirected`' and '`directed`' graph, default: '`undirected`'
`vertexcolor`	color of the vertices, default: '`darkblue`'
`vertexframe`	color of the frame of the vertices, default: `vertexcolor`
`labeldist`	distance of the label from the center of the vertex, default: `-2.5`
`labelcex`	font size for vertex labels, default: `0.8`
`labelcolor`	color of the vertex labels, default: '`black`'
`labelfont`	font for the vertex labels, default: `2`
`edgewidth`	width of the edges, default: `2.5`
`edgecol`	color of edges, default: '`black`'
`curved`	edge curve, range between (0,1), default: `FALSE` (=0), see **igraph**

rescale	logical constant, whether to rescale the coordinates, default: TRUE, see **igraph** documentation
pathwayname	name of the plotted graph
title	title of the plotted graph
vertexname	optional vector of names for vertices (user-defined)
layout	layout type of illustration, default: layout.fruchterman.reingold
weighted	logical variable, default: NULL, (if the adjacency matrix has weights it needs to be set to TRUE)

Value

plot of the pathway

Note

This function holds only part of the options of the plot()- function of the package **igraph**. For more option see **igraph**.

Author(s)

Bernadette Wendel

Examples

```
PathwaysURI<-searchPathway("neuronal system","reactome")
selectedPathway<-subset(PathwaysURI,name=="Transcriptional Regulation by MECP2")

Pathway1<-get.pathway(selectedPathway, URI="uri", pathwayname="name", delete =TRUE)
pathway.plot(Pathway1,direction="undirected",pathwayname="Transcriptional Regulation by MECP2",
            title="Display of the Regulation pathway")

pathway.plot(Pathway,direction = "undirected", pathwayname = "articifial")
```

Quadratickernel *Computation of a quadratic kernel*

Description

This function creates a kernel based on the quadratic kernel computation.

Usage

```
Quadratickernel(Genotype)
```

Arguments

Genotype	a single *nxp* matrix of the genotype of *n* individuals and *p* SNPs (no missing values)

Value

an *nxn* kernel matrix (*n* - number of individuals)

Author(s)

Bernadette Wendel

References

For details on the quadratic kernel

- Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin Y, Engel SM, Molldrem JJ, Armistead PM: Kernel machine SNP-set testing under multiple candidate kernels. Genet Epidemiol. 2013, 37(3):267-275.

Examples

```
KernelQuad<-Quadratickernel(Genotypematrix)
```

`reduceGenotype` *Downsizing of genotype matrix*

Description

The genotype matrix (dimensions) is reduced to match the dimensions of the phenotype data when measurements are missing. The genotype matrix is also transformed in long format. If the genetic time interaction effect is of interest, a time factor needs to be provided by `timevector` and the genotypes are multiplied with the time.

Usage

```
reduceGenotype(
  phenotypedata,
  Genotype,
  phenotype = "pheno",
  currentGenotypeformat = "long",
  m = NULL,
  timevector = NULL
)
```

Arguments

`phenotypedata`	phenotype data in long format with missing values (do not delete rows with missing values), row names are used to match with row names of `Genotype`
`Genotype`	an *nxp* matrix of the genotypes of all individuals, are used to match with row names of `phenotypedata`
`phenotype`	column with phenotype data
`currentGenotypeformat`	format of the input genotype matrix (Must be either '`long`' or '`wide`'), default: '`long`'

m	number of time points, only required if the format of the genotype matrix is 'wide'
timevector	vector containing time points if a genetic time interaction effect is of interest, default: NULL

Value

a reduced form (smaller dimensions) of the genotype matrix, if the phenotype data is longitudinal then the genotype matrix will also be transformed into long format

Author(s)

Bernadette Wendel

Examples

```
Genotypecross<-reduceGenotype(Genotypematrix,phenotypedata = CrossPhenotype,
                              phenotype="Quantmisspheno",
                              currentGenotypeformat = "wide",m=1)
dim(Genotypecross)
Kernelcross<-Linearkernel(Genotypecross)
dim(Kernelcross)
```

---

searchPathway *Scan Pathway Commons database for pathway(s)*

Description

This function scans the database Pathway Commons by a keyword returning a data frame of the found pathways with information, e.g. name and URI. The function uses the function `searchPc()` of the package **paxtoolsr** searching the Pathway Commons database.

Usage

```
searchPathway(keyword, pathwaysource, type = NULL, organism = NULL)
```

Arguments

keyword	keyword or name of a pathway to find pathway(s) and URI
pathwaysource	name of pathway database which should be used, only databases part of Pathway Commons (e.g. Reactome)
type	describes type of information returned, default: 'Pathway'
organism	type of organism researched, default: 'homo sapiens'

Value

data frame with 3 columns: name, uri, biopaxClass

Author(s)

Bernadette Wendel

References

For more details on Pathway Commons database

- Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, Anwar N, Schultz N, Bader GD, Sander C Pathway Commons, a web resource for biological pathway data. Nucleic Acids Research 2010, 39(Database):D685-D690.
- Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, Franz M, Siper MC, Cheung M, Wrana M, Mistry H, Mosier L, Dlin J, Wen Q, O'Callaghan C, Li W, Elder G, Smith PT, Dallago C, Cerami E, Gross B, Dogrusoz U, Demir E, Bader GD, Sander C Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. Nucleic Acids Research 2019

Examples

```
PathwaysURI<-searchPathway("neuronal system","reactome")
head(PathwaysURI);dim(PathwaysURI)
```

SNPBimFile *Example of SNP information data*

Description

The data set `SNPBimFile` is a data frame with simulated information on 2000 SNPs.

Usage

```
SNPBimFile
```

Format

A data frame with 2000 rows and 3 columns:

CHR chromosome of the SNP

SNP name of the SNP

BP base pair position of the SNP

Source

Simulated data set

summaryKMR *Summary function*

Description

summaryKMR is the summarizing function to return an Outcome object

Usage

```
summaryKMR(object)
```

Arguments

object	an object of class Outcome

Author(s)

Bernadette Wendel

WidePhenotype *Example quantitative longitudinal phenotype for 100 individuals*

Description

The data set WidePhenotype contains a simulated normally distributed phenotype for 100 individuals and three measurement points. The data are saved in wide format (one individual per row). The phenotype is complete (no missing values at any measurement point). In addition to the phenotype, we simulated two types of covariables (a binary and a continuous variable).

Usage

```
WidePhenotype
```

Format

A data frame with 100 rows and 6 columns:

IID Identifier of each individual, pattern: IID*n* with *n*=1,...,100

Age normally distributed variable, representing the age of an individual

Gender a binary (0 or 1) variable, representing the gender of an individual

pheno.t1 quantitative phenotype at time point 1

pheno.t2 quantitative phenotype at time point 2

pheno.t3 quantitative phenotype at time point 3

Source

Simulated data set

Index