# Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests

Saswati Saha, Laurent Perrin, Laurence Röder, Christine Brun, Lionel Spinelli

*Nucleic Acids Research, 2022* **1**
*https://doi.org/10.1093/nar/gkac715*

# Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests

**Saswati Saha [1,*], Laurent Perrin[1,2], Laurence Röder[1], Christine Brun [1,2] and Lionel Spinelli[1,*]**

[1]Aix Marseille Univ, INSERM, TAGC (UMR1090), Turing Centre for Living systems, Marseille, France and [2]CNRS, Marseille, France

Received April 06, 2022; Revised July 29, 2022; Editorial Decision August 01, 2022; Accepted September 12, 2022

## ABSTRACT

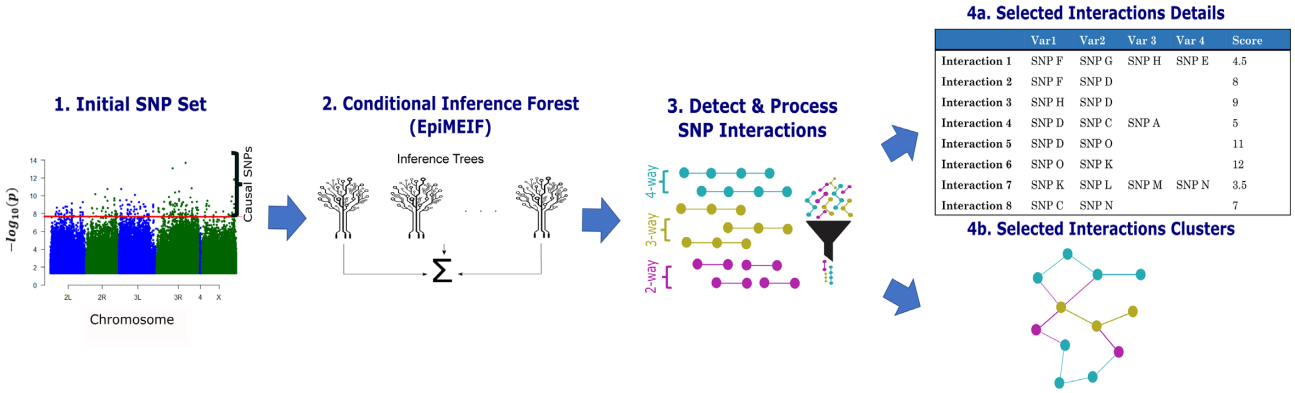
**Understanding the relationship between genetic variations and variations in complex and quantitative phenotypes remains an ongoing challenge. While Genome-wide association studies (GWAS) have become a vital tool for identifying single-locus associations, we lack methods for identifying epistatic interactions. In this article, we propose a novel method for higher-order epistasis detection using mixed effect conditional inference forest (*epiMEIF*). The proposed method is fitted on a group of single nucleotide polymorphisms (SNPs) potentially associated with the phenotype and the tree structure in the forest facilitates the identification of n-way interactions between the SNPs. Additional testing strategies further improve the robustness of the method. We demonstrate its ability to detect true n-way interactions via extensive simulations in both cross-sectional and longitudinal synthetic datasets. This is further illustrated in an application to reveal epistatic interactions from natural variations of cardiac traits in flies (Drosophila). Overall, the method provides a generalized way to identify higher-order interactions from any GWAS data, thereby greatly improving the detection of the genetic architecture underlying complex phenotypes.**

## INTRODUCTION

Over the past few decades, there has been a growing interest in phenotype to genotype association and Genome-Wide Association Studies (GWAS) have proven to be 'the standard tool' for identifying these associations (1,2). GWAS have attained tremendous success in identifying causal variants that exhibit independent, additive, and cumulative effects on the investigated phenotype trait (3). However, testing for associations *via* a single-locus test is an oversimplified approach to tackle the complexity of underlying biological mechanisms (4). Complex phenotypes and their variations within a population are speculated to be caused by multiple genetic variations and their interactive effects, which are referred to as *epistatic interactions* (5–8). However, the exhaustive evaluation of all possible epistatic interactions among millions of single nucleotide polymorphisms (SNPs) raises several issues, otherwise known as the 'curse of dimensionality'(9). Given a dataset with $n$ SNPs, the exhaustive epistasis search with the order of $m$ (number of interactive SNPs) requires $\binom{n}{m}$ combinations of SNPs to be tested, resulting in a complexity of $O(n^m)$ (10). Indeed, due to the exponential complexity involved in the higher-order exhaustive search algorithms, they are not applicable to large datasets. To address the above issues, several parametric modelling approaches (11,12), machine learning algorithms (8,10), and combinatorial optimizations (13,14) have been explored. But they are exclusively designed/used for detecting binary or higher-order interactions in case–control studies. There exist some approaches (15–18) that can detect pairwise interactions in cross-sectional studies with quantitative traits but they are simply not scalable to higher-order interactions. Moreover, current methods for detecting epistasis do not address the complexities and challenges involved in longitudinal datasets that allow studying the natural trajectory of traits and/or disease progression. As a solution, we propose a novel approach of epistasis detection, called *epiMEIF*, using a mixed-effect conditional inference forest (*MEIF*). The goal of our approach is to reveal higher-order interactions of genetic variants for complex quantitative phenotypes.

Recursive partitioning approaches or tree-based algorithms like random forest have already proven to be effective for detecting the genetic loci and their interactions

*To whom correspondence should be addressed. Tel: +33 491828712; Email: saswati.saha@univ-amu.fr.
Correspondence may also be addressed to Lionel Spinelli. Tel: +33 491828735; Email: lionel.spinelli@univ-amu.fr




| | Var1 | Var2 | Var 3 | Var 4 | Score |
|---|---|---|---|---|---|
| Interaction 1 | SNP F | SNP G | SNP H | SNP E | 4.5 |
| Interaction 2 | SNP F | SNP D | | | 8 |
| Interaction 3 | SNP H | SNP D | | | 9 |
| Interaction 4 | SNP D | SNP C | SNP A | | 5 |
| Interaction 5 | SNP D | SNP O | | | 11 |
| Interaction 6 | SNP O | SNP K | | | 12 |
| Interaction 7 | SNP K | SNP L | SNP M | SNP N | 3.5 |
| Interaction 8 | SNP C | SNP N | | | 7 |

**Figure 1.** Global overview of *epiMEIF*: 1) The method begins with a set of SNPs potentially associated with the phenotype obtained from a single-locus association test (via LMM). SNP markers that have a nominally significant effect on the phenotype are selected. 2) The mixed effect conditional inference forest (MEIF) is applied to the phenotype, selected SNPs, and other additional covariates that might explain the variability of the phenotype. 3) Identified SNP interactions from MEIF are subjected to additional statistical testing (ANOVA and Max-T test) that helps in filtering the stable interactions. 4) The final set of interactions can be visualized in two ways: a) the interaction score table that captures the different sets of interaction and their associated score from *epiMEIF* and b) as an interaction network where the nodes denote the different variants, and the edges denote the interactions from *epiMEIF*. Different order interactions are illustrated by different coloured edges in the network.

that impact the phenotypic outcome in case–control studies (7,8,19,20). Nevertheless, *MEIF* extended the application of tree-based algorithms beyond case-controls studies. It is particularly an improvement over existing methods because it (i) demonstrates how tree-based algorithms can be adapted for detecting higher-order interactions from complex phenotype datasets (both cross-sectional and longitudinal) using conditional inference forest (cforest), (ii) simultaneously account for missing/censored genome data using cforest, and other confounding factors like population structure in the GWAS datasets using mixed effects model. Moreover, we propose to substantiate the epistatic interactions obtained from MEIF using two additional statistical testing approaches: Max T-test and ANOVA test (see Figure 1 and Materials and Methods, for more details). Overall, *epiMEIF* provides a generalized way to obtain genetic variants and their higher order interactions from any GWAS data.

To the best of our knowledge, an alternative approach that can detect higher-order interactions in both cross-sectional and longitudinal datasets does not exist, therefore preventing us from evaluating our method *via* benchmarking. Hence, to evaluate our method, we applied it to an extensive simulation study and illustrated the power performance of the method under several practically relevant scenarios. We then analysed its performances on a real dataset: natural variation of heart period in young flies (Drosophila) and heart period during aging in Drosophila, where heart period measures the duration of a complete cardiac contraction/relaxation cycle. We evaluated the method's performance based on its ability to validate in an independent cohort and to recover previously published genetic data associated with cardiac functions. We demonstrate that the obtained networks of statistical interactions provide insightful information with respect to cardiomyocytes' structure and functions. These analyses illustrate the high performance of the method to identify higher-order interactions from both cross-sectional and longitudinal data. The proposed statistical methods are implemented in R and the source codes can be found in the Github repository (https://github.com/TAGC-NetworkBiology/epiMEIF).

The manuscript is organized as follows: a precise overview of the Materials and Methods is presented in Section Materials and Methods and more details pertaining to Materials and Methods can be found in the Supplementary Materials. Section DGRP Cardiac Dataset presents details on the Drosophila Cardiac Dataset on which the *epiMEIF* method is applied. The results obtained from the implementation of the method on real and synthetic datasets are presented in Section Results, where the inferences from cross-sectional data applications are presented in Section Applications on cross-sectional Data, and longitudinal data applications in Section Applications on longitudinal data. The paper concludes with a discussion in section Discussion.

## MATERIALS AND METHODS

We propose an approach for epistasis detection using a mixed effect conditional inference forest to identify the genetic variants and their epistatic interactions responsible for the complex quantitative traits. The novelty of this approach lies not only in the epistasis detection but also in the amalgamation of the mixed effects model and conditional inference forest (cforest). We have divided the explanation of the method into four parts: (i) we explain the MEIF model, (ii) we illustrate how the tree structure in the conditional inference forest (cforest) is utilized to detect high-order SNP interactions that impact the variation of the phenotype, (iii) we explain how MEIF can be adapted to weighted MEIF, this is particularly useful for cross-sectional datasets and (iv) we show how the interactions from the MEIF can be validated using independent statistical tests. Cforests, developed by Torsten Hothorn *et al.* (53), can be considered as an alternative to the random forests where the ensemble algorithms use conditional inference trees as base learners. More details on the usage of cforests can be found in the Supplementary Materials where

we present an extended/detailed version of the Material and Methods.

### Mixed effect conditional inference forest (MEIF)

Our method is primarily inspired by the mixed effect random forest (MERF) proposed by Hajjem *et al.* (21). It is a combination of random effects model and random forest where the application of random forest is extended to clustered data for a continuous outcome. Note that they did not consider any fixed effects outside the random forest in their model. We aim to implement MERF on genomic data where some of the variants may not be properly sequenced for all the samples and often suffer from missing data issues. To fit a random forest on genotype data with missing values one usually needs to impute the genotype data and then apply the above method. To avoid that, we propose MEIF, where we combine the mixed effects model with cforests instead of random forests and we have elaborated in the supplementary materials on how we have extended the MERF model to the generalized MEIF model for our application (section S1.1). MEIF can simultaneously account for missing/censored genome data using cforest, and other additional covariates or confounding factors like population structure in the GWAS datasets using the mixed effects model. The mixed effects conditional inference forest (MEIF) used in our method can be defined as follows:

$$\begin{aligned} &y_i = \sum a_j X_{ij} + Z_i b_i + f(S_{i1}, S_{i2}, \cdots, S_{iN}, T_i) + \epsilon_i, \\ &b_i \sim N_q(0, D), \ \epsilon_i \sim N_{n_i}(0, R_i), \\ &i = 1, \ldots, K \quad j = 1, \cdots, p, \quad \sum n_i = n \end{aligned} \tag{1}$$

where

- $i$ is the cluster index. Assume that there are $K$clusters in the training data (e.g. in the Drosophila cardiac genetic dataset, cluster denotes the Drosophila strain where all flies within a strain have the same genotype data and are assumed to have a same phenotype distribution),
- $y_i$ is the response variable. $y_i = [y_{i1}, y_{i2}, \cdots, y_{in_i}]^T$ is the $n_i \times 1$ vector of responses for the $n_i$ observations in cluster $i$ (e.g. in the Drosophila cardiac genetic dataset, $y_i$ is the phenotype of interest corresponding to all the flies observed within a particular strain $i$),
- $X_{ij} = [X^1_{ij}, X^2_{ij}, \cdots, X^{n_i}_{ij}]^T$ is the $n_i \times 1$ vector of $j$th fixed-effect covariate for cluster $i$, and $\boldsymbol{X_j} = [X_{1j}, X_{2j}, \cdots, X_{Kj}]^T$ is the $n \times 1$ vector of $j$th fixed-effect covariate over all clusters (e.g. in the Drosophila cardiac genetic dataset, $X_{ij}$ comprises the variables like the date on which fly is dissected),
- $Z_i$ is the $n_i \times q$ matrix of random-effect covariates for cluster $i$ that we do not intend to include in the cforest (e.g. in the Drosophila cardiac genetic dataset, $Z_i$ comprises strain covariate to which the fly belongs).
- $S_1, S_2, \cdots S_N$ are the marker variables in our study and $T$ capture the time or the longitudinal component of the data ($T_i$ denotes the time at which $y_i$ is observed). $f(S_{i1}, S_{i2}, \cdots, S_{iN}, T_i)$ is ideally estimated using conditional inference forest (cforest). The covariate time ($T$) can be omitted for cross-sectional data.
- $a_j$ is fixed effect of the $j$th covariate,
- $b_i$ is the $q \times 1$ unknown vector of random effects for cluster $i$. $b_i$ follows normal distribution with mean 0 and variance covariance $D$, for cluster $i = 1 \ldots K$.
- $\epsilon_i$ is independent, identically distributed ($iid$) noise. $\epsilon_i$ follows normal distribution with mean 0 and variance covariance $R_i$, for cluster $i = 1 \ldots K$.

Note that we incorporate the cforest on the markers and time covariate only as we would like to capture the nonlinear effects of the genotype on the phenotype and extract the interacting genotype components that might impact the phenotype. Similar to GWAS methods like GCTA (22), EMMAX (23), MEIF can also incorporate the population structure in GWAS data with the random effects component shown in Equation (1). The model in Equation (1) can be modified as follows to incorporate population structure in GWAS data:

$$y_i = \sum a_j X_{ij} + \boldsymbol{g} + f(S_{i1}, S_{i2}, \cdots, S_{iN}, T_i) + \epsilon_i, \tag{2}$$

Here, $\boldsymbol{g}$ is a $n \times 1$ vector of the total genetic effects of the individuals with $g \sim N(0, A\sigma^2_g)$, and $A$ is interpreted as the genetic relationship matrix (GRM) between individuals. We can therefore estimate $\sigma^2_g$ by the restricted maximum likelihood (REML) approach as in other GWAS.

Pooling data across all the clusters in equation (1), we have the following *MEIF* model for cross-sectional dataset:

$$Y = Zb + a'X + f(S_1, S_2, \cdots, S_N) + \epsilon$$

$$\Rightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_K \end{pmatrix} = \begin{pmatrix} Z_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z_K \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_i \\ . \\ b_K \end{pmatrix} + \langle X_1 |..| X_L \rangle \begin{pmatrix} a_1 \\ \vdots \\ a_j \\ . \\ a_L \end{pmatrix}$$

$$+ f(S_1, S_2, \cdots, S_N) + \epsilon$$

For longitudinal dataset, the above models can be written as follows:

$$\begin{pmatrix} Y^1 \\ \vdots \\ Y^t \end{pmatrix} = \begin{bmatrix} Z^1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & Z^t \end{bmatrix} \begin{pmatrix} b^1 \\ \vdots \\ b^t \end{pmatrix} + \begin{pmatrix} a^{1'} X^1 \\ \vdots \\ a^{t'} X^t \end{pmatrix} + f(S_1, S_2, \cdots, S_N, T) + \epsilon$$

where $t$ denote the total number of time points in the longitudinal data and superscript $t$ for each component denote the corresponding response and predictors at time point $t$. Note that we have adopted the approach of the original article (21) to fit the above model (an expectation-maximization (EM) technique is used to fit the different components of *MEIF*). We have integrated the cforest function in the partykit R package (24) with the MERF R codes provided by Hajjem *et al.* (21) to build MEIF in R. The source codes of MEIF can be found in the Github repository (https://github.com/TAGC-NetworkBiology/epiMEIF). More details on the MEIF model can be found in the Supplementary Materials (section S1.1).

### Epistatic interactions detection with Epi-MEIF

The random forest technique has been predominantly used for classification and prediction analyses (25) and as a feature selection tool (26). Moreover, unlike the single marker

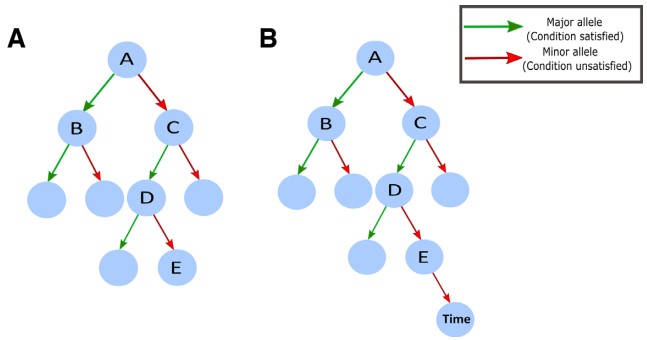



**Figure 2.** Detecting interaction from tree-based methods. (**A**) Example tree generated by the cforest algorithm. SNP pairs A and B, A and C, A and D, A, and E, C and D, C and E, and D and E represent descendant pairs, which may indicate epistatic genetic effects, whereas SNP pairs B and C, B and D, and B and E represent non-descendant pairs, which may indicate independent additive genetic effects. (**B**) Example tree generated by the cforest algorithm when applied to a longitudinal dataset.

GWAS model such as GCTA (22) or FastLMM (27), the phenotype predicted using random forest (RF) or cforest accounts for the combined/interaction effect of various SNPs (8). In this regard, another unique feature of forest-based algorithms is that the tree structure in the forest can be exploited to detect the SNP interactions. The nature of the trees constructed in the forest allows for interaction detection in the sense that each path through a tree corresponds to a particular combination of values taken by certain predictor variables, thus including potential interactions between them. If a particular path (or a sequence of covariates) occurs more often than the others in multiple trees of the random forest, we can claim the group of variables in the path are in interactions. We explain in the subsequent sections how to detect interactions from MEIF in cross-sectional and longitudinal datasets.

It is noteworthy to mention here that MEIF is not scalable to an entire genome-wide setting. Hence, to retrieve the set of SNP inputs for MEIF, we propose to apply a single-locus association analysis (using a linear mixed model (LMM)) that helps to filter the SNP markers that have a significant effect on the phenotype. This can help in alleviating the computational burden of *epiMEIF* (illustrated in Figure 1). Note that after fitting MEIF (using the model in Equation (1)) on the above-mentioned SNPs and the other environmental and population covariates, the cforest ($f$) component is extracted to detect epistatic interactions.

*Detecting epistasis in cross-sectional data.* Figure 2A illustrates the mechanism of epistasis set construction from the random forest/conditional inference forest in the cross-sectional dataset (28). Precisely, if two SNPs, C and E have a large epistatic effect then the combination C and E will appear more often in the same branch of a tree than in other branches or trees (see Figure 2A). This combination will thus form a parent/descendant (child, grandchild and so on) pair. On the contrary, if two SNPs, like B and C, have large but independent main effects on the response variable, they will also appear frequently within the same tree, but in different branches (that are not descendant pairs) (see Figure 2A). Thus, the descendant pairs that occur most frequently in the forest can be recognized as possible pairwise epistatic interactions. Similarly, C, D and E are expected to occur more often in the same branch of the tree if they have a large epistasis effect (see Figure 2A). The tree structure thus allows to identify higher-order (more than binary) interactions. Appendix 1 further elaborates the different steps of the MEIF.

The mechanism illustrated in Figure 2A will finally yield the list of all potential interaction sets from the forest. We compute a SNP Interaction matrix ($I$) that measures the strength of each interaction set based on their frequency of occurrence in the same branch of the forest (see Appendix 1). Note that the above mechanism is relevant for identifying interactions in cross-sectional phenotype-genotype association studies and it needs to be extended for longitudinal case studies.

*Detecting epistasis in longitudinal data.* Ideally, the longitudinal aspect of the dataset is captured with the variable 'Time'. It is worthwhile to point out here that with regards to the data applications discussed in the current article, we are essentially interested in those interactions where the temporal variation of the phenotype changes significantly from any of the alternative genotype combinations to the reference genotype combination (reference: genotype combination arising from the combination of major alleles of multiple SNPs, alternative: all genotype combinations except the reference). Hence, the entire process of extracting interactions from MEIF in the longitudinal data application comprises two steps:

1. **Remove the longitudinal phenotypic trend of the reference population**: This is done using the following approach. The MEIF model is fitted and the phenotype $\hat{Y}$ is predicted for the genotype data where all SNPs have major alleles i.e. $SNP_i = 0,\ \forall i \in \{1, \cdots, N\}$, keeping the other covariates fixed. The predicted phenotype can be denoted as $\hat{Y}_{reference\_population}$ and this effect is removed from the true phenotype ($\tilde{Y} = Y - \hat{Y}_{reference\_population}$). Finally, the MEIF is implemented on the resultant phenotype $\tilde{Y}$ and the interactions are extracted using the mechanism discussed below.
2. **Extract the interactions from the resultant MEIF**: The mechanism to extract interactions from the trees in the longitudinal MEIF application is illustrated with the help of Figure 2B. Since our primary objective is to capture the epistatic interactions that are responsible for the change of the phenotype over 'Time', we focus on those branches of the trees that have 'Time' in at least one of the child nodes (see Figure 2B). If A, C and D appear more often in the same branch of a tree as the ancestor of 'Time', then A, C, and D are expected to have epistatic effect having an impact on the longitudinal variation of the phenotype (see Figure 2B). Each path of SNPs from root to leaf preceding the node having the 'Time' covariate is a possible interaction. Certain combinations of the SNPs that appear more frequently in the same branches are believed to interact with each other. Adopting this intuition, we extract all SNP interactions using the com-

binations of SNPs from the forest that are in the same branch as 'Time' in the forest. Thereafter, we build SNP Interaction matrix ($I$) that measures the strength of each interaction set based on their frequency of occurrence in the same branch.

Combining 1) and 2) will indeed yield the SNP interactions where at least one of the alternative genotype combinations generates a significantly different longitudinal trend from the reference genotype combination. As illustrated in Appendix 1, we fit 10 cforests or MEIFs after removing the longitudinal trend of the reference population and then obtain the interaction sets that are occurring with high scores in 90% of the forests (9/10 forest). Further illustrations on detecting interactions from longitudinal datasets can be found in the extended Material and Methods in the Supplementary Materials (Section S1.2.2 and Supplementary Figure S1).

### Weighted conditional inference forest

A notable shortcoming of tree-based methods is that they are dependent on marginal effects (8). The main association effect of the SNP on the phenotype is denoted as the marginal effect here. To address this, we developed a weighted adaptation of mixed effect conditional inference forest (weighted *epiMEIF*). The Linear Mixed Model (LMM) or any single-locus association test, that we propose to apply prior to MEIF application provides the significance of each variable in the phenotype-genotype association dataset. We utilize this to rank the SNPs based on the size of their marginal effects and the $P$-value from the LMM. The weight assigned to each SNP is inversely proportional to the rank of the SNP (i.e. highly significant SNPs have lower weights). This ensures that SNPs with higher marginal effects are drawn less during the tree construction, thereby increasing the chance of capturing interactions involving SNPs that have low marginal effects (29). More details on the weight design can be found in the extended Material and Methods section in the Supplementary Materials (Section S1.3 and Supplementary Figure S2).

One hurdle for the longitudinal data is to decide on the weight of the 'Time' covariate. Assigning a heavy weight can lead to trees where 'Time' is often selected at the root node, whereas assigning a low weight can end up with trees where is 'Time' is never selected for any node. So, we avoided the weighted cforest or *weighted epiMEIF* application for longitudinal data and have applied it only for cross-sectional simulated and real datasets.

### Epistasis network validation

We propose to perform an independent statistical test on the interaction lists obtained from the MEIF. The objective of this testing step is to further substantiate the interactions obtained from MEIF and help to get rid of false positives if any. Within this statistical validation framework, we intend to address two primary questions: (i) Do the SNP combinations in each high-order interaction (selected from the MEIF) give rise to a significant impact on the variation of the phenotype? (ii) Is the impact of this SNP combination better than any random combination of $N$ SNPs, where $N$ is the size of the high-order interaction (under investigation)? To address this, we perform the ANOVA and Max-T test. For each interaction set of size N, ANOVA method tests if the $N$-SNP interaction is statistically significant. It is analogous to the regression-based-test (30,31) where logistic regression or linear regression are used to assess the impact of SNP interactions on diseases or quantitative traits. They typically compare the saturated model that includes interactions against the reduced model that omits interactions using likelihood ratio tests. Max-T test, on the other hand, tests if the SNP combinations in each higher-order interaction (size $N$) give rise to a significant impact on the variation of the phenotype and if the impact of this SNP combination is better than any random combination of $N$ SNPs (see Supplementary Box 1). More details on the two testing approaches can be found in the extended Material and Methods section in the supplementary materials (Section S1.4 and Supplementary Figure S3). Note that while the Max-T test is scalable to higher order interactions, the ANOVA test may not be always scalable to high order interactions. Since we obtain interaction of maximum order 4 in the real dataset used in this article (DGRP dataset), we applied both ANOVA and Max-T test here.

## DGRP CARDIAC DATASET

The *epiMEIF* method has been developed in the framework of a project aiming to identify the genetic architecture of natural variations associated with cardiac aging in Drosophila. We analysed the cardiac performances in a natural population of young (1 week) and old (4 week) flies from the Drosophila Genetic Reference Panel (DGRP (32)), a population consisting of 205 inbred lines derived by 20 generations of full-sib inbreeding from inseminated wild-type caught female flies from the Raleigh, USA population. Whole-genome sequencing data, along with genotype calls, are available for all 205 lines. Contractility and rhythmicity were measured in females, using ∼2000 flies from 168 DGRP lines at 1 week and 1800 flies from 165 lines at 4 weeks. More details on the phenotype and genotype dataset can be found in Saha et al., (33) and in the Supplementary Materials (Section S3 and Supplementary Figure S4). An additional dataset of 20 DGRP lines was also analysed by us where the cardiac performance was studied following the same approach (with 12 flies observed per line) that was not used in the model development. We treat this dataset as our validation dataset and utilized it to test if the genetic interactions predicted from *epiMEIF* method in the original dataset have an impact on the phenotypic variations in this independent dataset. More details on the DGRP dataset and the pre-processing and quality control of the above datasets can be found in the Supplementary Material (Section S3).

## RESULTS

### Overview of Epi-MEIF

We have proposed a new method for epistasis detection in large-scale association studies with complex genetic traits. The method begins with fitting the MEIF model on a set of

SNPs potentially associated with the phenotype (and additional covariates, if any) and exploits the tree structure in cforest to compute the SNP Interaction matrix ($I$). It measures the strength of each interaction set based on their frequency of occurrence in the same path of the forest. Thereafter, the interactions in $I$ are verified using additional testing strategies (Max-T test or ANOVA test) leading to the final SNP Interaction matrix ($F$). These independent testing strategies increase the reliability of the final interactions and help to get rid of false positives if any. Finally, the SNP Interaction matrix ($F$) helps to build the statistical epistatic network of genetic variants. The complete workflow is illustrated in Figure 1. Seldom, the tree-based algorithms are criticized to be biased toward variants with high marginal effects. To address that caveat, we developed a weighted adaptation of mixed effect conditional inference forest (Weighted *epiMEIF*), where the probability of selecting the variant during the root node construction of the tree depends on the significance of the predictor variable from the single-locus association; higher the significance, lower the sampling weight. Note that the novelty of the proposed *epiMEIF* approach is 2-fold: extension of MERF proposed by Hajjem *et al.* (21) primarily to address the challenges pertaining to GWAS data, and proposing a new method to extract high order epistasis, which makes *epiMEIF* more than an extension of MERF.

We analysed the performance of *epiMEIF* on real datasets: natural variation of heart period in young flies (Drosophila) and heart period during aging in Drosophila. To further evaluate our method, we analysed the power performances of *epiMEIF* under diverse simulation scenarios. We have presented the simulation scenarios and results from the simulations and real data analyses in the upcoming section. More details on the simulation scenarios can be found in the Supplementary Materials (Section S4).

### Applications on cross-sectional data

*Analysis on simulated data.* We evaluated the statistical power of the method with the help of synthetic datasets where the ground truth genetic architecture is known. We used the genotypes of the widely used Drosophila Melanogaster Genetic Reference Panel (DGRP (32)) to design our simulation scenarios. The genotype data in our scenarios comprised 1,000 SNPs with a minor allele frequency (MAF) >0.1 subsampled from 2 456 752 genome-wide SNPs of the DGRP genome dataset. We considered three simulation scenarios where all are additive models containing four SNPs with marginal effects and 2 to 4 SNPs with interaction effects (see Figure 3 and Supplementary Materials Section S4, to better understand the overall schematic of how the simulations are conducted). The scenarios differ only in the number of interaction pairs/sets and the number of SNPs involved in these interactions. Scenario 1 considers the simplest scenario where there exists only 1 two-SNP interaction in the linear additive model (called SN1), Scenario 2 comprises 2 two-SNP interactions (SN2) and Scenario 3 comprises 1 three-SNP interaction (SN3). Note that SN2 is so designed such that the set of SNPs involved in the first binary interaction has a lower marginal effect (5 and 6 in SN2, see Figure 3) and SNPs involved in the second binary interaction have a higher marginal effect (7 and 8 in SN2, see Figure 3). Apart from the 6 to 8 SNPs (1, 2, .., 8, illustrated in Figure 3) with which the model is built, we randomly add 30 or 50 other SNPs (from the set of 1000 SNPs) and conducted 100 simulations for each scenario (see Supplementary Materials Section S4 for more details). Note that MEIF implementations in R take around a minute for data with 200 SNPs. More details related to the computation efficiency of the method can be found in the Supplementary Materials (Section S2 and Supplementary Figure S5).

We evaluated the performance of our approach based on the proportion of simulations (out of 100) where the true interactions (example: two-way interaction 5–6 is the true interaction in SN1, see Figure 3) are captured. We call it 'the overall power in capturing the true epistatic interactions' and explored the overall power for both *epiMEIF* and *weighted epiMEIF* (see Table 1). For both the adaptation, the overall power in detecting the true interactions is consistently satisfactory ranging between 90 and 100% across the scenarios SN1 and SN2. The power to detect the binary interactions in SN2 is higher (2–4% increase) with weighted *epiMEIF* compared to *epiMEIF*. Note that despite the increased complexity in identifying higher order interactions (non-binary) in SN3, we attain a power of 70–90%.

Since *epiMEIF* provides a score for each interaction depending on the strength of the interaction sets from the cforest, we also investigated how often the true interactions are captured as the top-ranking interactions (based on the score) in our simulation scenarios (see Supplementary Figure S6). We observed that the *weighted epiMEIF* can capture the true interaction more efficiently compared to the *epiMEIF* for the cross-sectional data simulations (*weighted epiMEIF* appeared in the top ranks 10–30% more often than *epiMEIF* in Supplementary Figure S6). The power gain is more prominent for the simulation results with 30 additional SNPs. It is noteworthy to mention here that though the 'overall power' of *weighted epiMEIF* (71%-86%) is lower than *epiMEIF* (80–90%) in SN3 (see Table 1), the power to detect the true interactions in the top ranks is much higher in *weighted epiMEIF* (45–80%) compared to *epiMEIF* (37–45%) (see Supplementary Figure S6c). Overall, *weighted epiMEIF* is more effective than *epiMEIF* as it gives comparatively robust results across all the scenarios and more often selects the true interactions as top-ranking.

Furthermore, to demonstrate that the mechanism to explore high order epistasis interactions (elaborated in Appendix 1) is also applicable for other tree-based algorithms like MERF, we have implemented the adapted MERF model in Equation (1), but without replacing the randomForest R function with the cforest R function (R codes are provided in https://github.com/TAGC-NetworkBiology/epiMEIF). We have conducted simulations to compare the performance of MEIF with the MERF model in extracting higher-order interactions. We have compared *epiMEIF* and *weighted epiMEIF* with MERF for simulation scenarios SN1 and SN3 with 50 additional SNPs and found MERF has highly comparable performance with the *weighted epiMEIF* (see Supplementary Figure S7). Note that MERF is easily applicable for the simulated datasets as there was no missing genotype information in the simulated data. However, despite the comparable performance

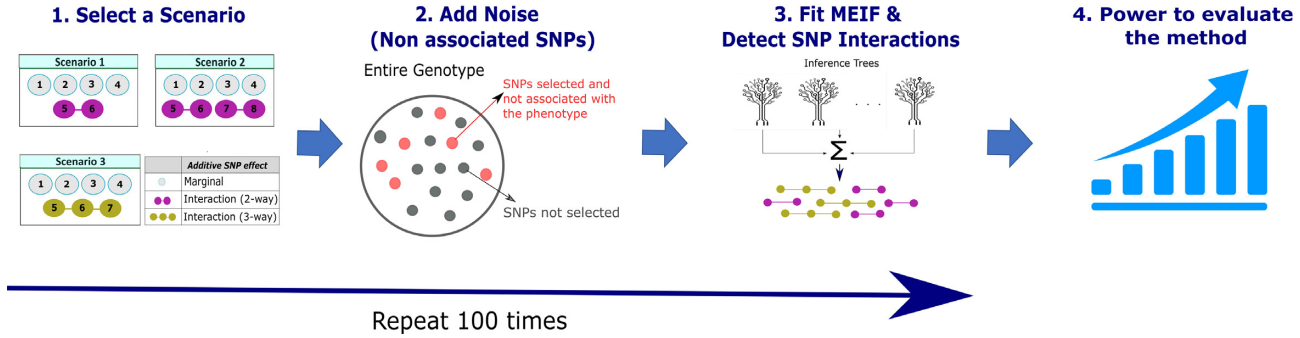



**Figure 3.** Step-by-step schematic diagram illustrating the simulation pipeline: 1) Select a simulation scenario Scenario 1 (SN1), Scenario 2 (SN2), or Scenario 3 (SN3) involving the SNPs 1 to 8. 2) The set of SNPs for fitting *epiMEIF* in the simulated data is prepared using the SNPs with which the simulation scenario is built, and randomly selected SNPs from the set of 1000 SNPs that are not associated with the simulated phenotype (highlighted in red). These additional markers in Step 2 act as noise and help to evaluate the power of the method. 3) Fit *epiMEIF* (or weighted *epiMEIF*) and obtain the set of interactions. Repeat steps 1 to 3, 100 times. 4) Detect the proportion of times *epiMEIF* can capture the true interaction (e.g. Interaction between 5 and 6 in Scenario 1) amongst the 100 iterations.

**Table 1.** Proportion of simulations (out of 100) capturing the true interactions for the different Age 1 simulation scenarios under the two additional sample size cases (30 and 50)

| | #Captured (%)* | | | |
|---|---|---|---|---|
| #Additional SNPs | 30 | | 50 | |
| Method | MEIF | WMEIF | MEIF | WMEIF |
| SN1 | 100 | 100 | 100 | 100 |
| SN2 (SNP_5: SNP_6) | 93 | 95 | 90 | 94 |
| SN2' (SNP_7: SNP_8) | 90 | 94 | 92 | 94 |
| SN3 | 91 | 86 | 80 | 71 |

*#Captured (%): $\frac{\sum_i I(\text{interaction detected in } i^{th} \text{ simulation})}{100}$, where $I$ is an indicator variable.

between MEIF and MERF we prefer MEIF because it can handle well missing genotype data and data imputation is not necessary. Hence, MEIF is more suitable for the data applications (GWAS data) discussed in this article.

*Analysis on real data:* epiMEIF *network construction for natural variations of heart period in flies.* Testing *epiMEIF* on simulated data showed its power for capturing both binary and higher order interactions among SNPs potentially associated with the phenotype. We, therefore, tested the method's ability to detect statistical interactions from the GWAS data on cardiac performance traits in Drosophila. Drosophila is an ideal model to study heart development and adult cardiac function (34–36). We recently investigated the genetic architecture of cardiac performance in young (1 week) flies from the DGRP population and gained insights as to the molecular and cellular processes affected (33). Importantly, correlations observed between identified genes and cardiac dysfunction suggested a conserved genetic architecture of cardiac function in both flies and humans (33). Leveraging this dataset, we investigated how *weighted epiMEIF* would detect epistasis interactions among variants associated with natural variations of heart period (HP). To accommodate computational requirements of *weighted epiMEIF*, interactions were tested on variants that show significant association with the quantitative trait from the single GWAS LMM, with a nominal significance threshold of $10^{-5}$ (3484 *SNPs*, see Supplementary Materials). Finally, *weighted epiMEIF* application on HP dataset led to a dense network of 35 interacting SNPs (Figures 4 and 5A). We have first illustrated the statistical interaction network obtained from *weighted epiMEIF* (Figure 4) and then illustrated the biological significance of the corresponding network with the most connected nodes (Figure 5A). Binary and higher-order statistical interactions were combined to illustrate the interaction network obtained from the *weighted epiMEIF* (Figure 4A) and some of the four-way and three-way interactions are highlighted in Figure 4B and C. We have also illustrated how different allelic combinations of SNPs involved in a selected higher-order interaction are associated with significant HP variations (see Figure 4D). On the contrary, the allelic combinations of SNPs randomly selected in the network (and not participating in a detected interaction) have no effect on the phenotype (see Figure 4E).

Remarkably, many genes in the network have known cardiac functions, in flies and/or in mammals, and the network connects variants in genes whose function and/or subcellular localization are highly correlated (Figure 5A, Supplementary Table S1). In particular, it includes several genes encoding proteins known to interact with actin - either sarcomeric (*cheerio* (*cher*, human orthologue *FLNC*)- denoted by *cher/FLNC* henceforth-, *Zormin/MYPN*) or non sarcomeric (*Src oncogene at 42A (Src42A)/FRK*, *cher/FLNC*, *CG32264/PHACTR1-2*, *Paxillin(Pax)/PXN*, *Rho GTPase activating protein at 71E (RhoGAP71E)/ARHGAP20*)—and with myosin (*smog/GPR158*). In tight interaction with the extracellular matrix (ECM), cytoskeletal proteins play a central role in the mechanical and signalling properties of the cardiomyocytes (37). From this perspective, it is remarkable that two genes encoding ECM components—*multiplexin* (*mp)/Col15-18* and *tenectin* (*tnc)/* AKAP12—have a central place in the network and share many interactions. A third collagen containing ECM constituent, *mspondin (mspo)/SPON2*, also participates in the network. Importantly, *mspo*—whose mammalian orthologue has a cardioprotective activity (38)—, is known to interact

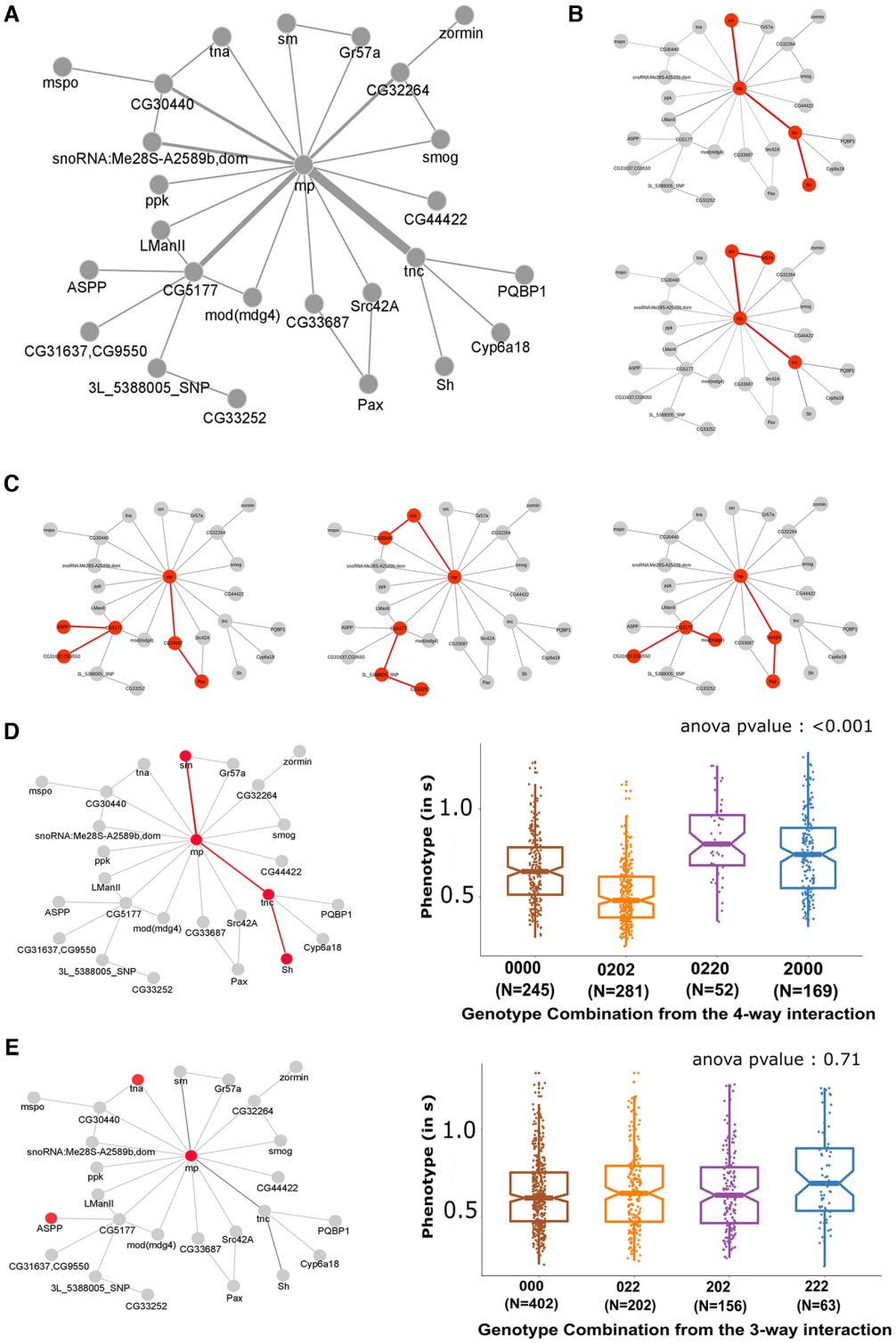



**Figure 4.** Weighted *epiMEIF* high-order statistical interactions detected on natural variation of heart period in flies: (**A**) illustrates the network obtained by accumulating the detected high-level interactions obtained when weighted *epiMEIF* is fitted on the cardiac heart period of the DGRP population at 1 week. The different nodes denote the variants from the DGRP genotype data that are detected with the *epiMEIF*. Whenever the variants are mapped to a gene, the nodes are annotated with the corresponding gene name (instead of the variant name). The edge joining two nodes denotes an interaction and the thickness of the edge quantifies the number of common interactions shared by the two nodes. (**B**) and (**C**) respectively highlight (in red) some of the four-way and three-way interactions detected by *epiMEIF*. (**D**) The boxplot shows how the phenotype distribution varies across the different genotype combinations arising from the selected 4-way interaction: sm-mp-tnc-Sh (in red). Max-T test/ANOVA test is performed to verify the selected high-order interaction obtained from MEIF. The *P*-value for the selected interaction in both tests is <0.001. (**E**) shows how the three-way interaction generated from a group of SNPs selected randomly from the *epiMEIF* network (not sharing edges/interactions) does not have any effect on the variation of the phenotype. The *P*-value from the anova test is 0.71, reporting an insignificant effect of the genotype on the phenotype.

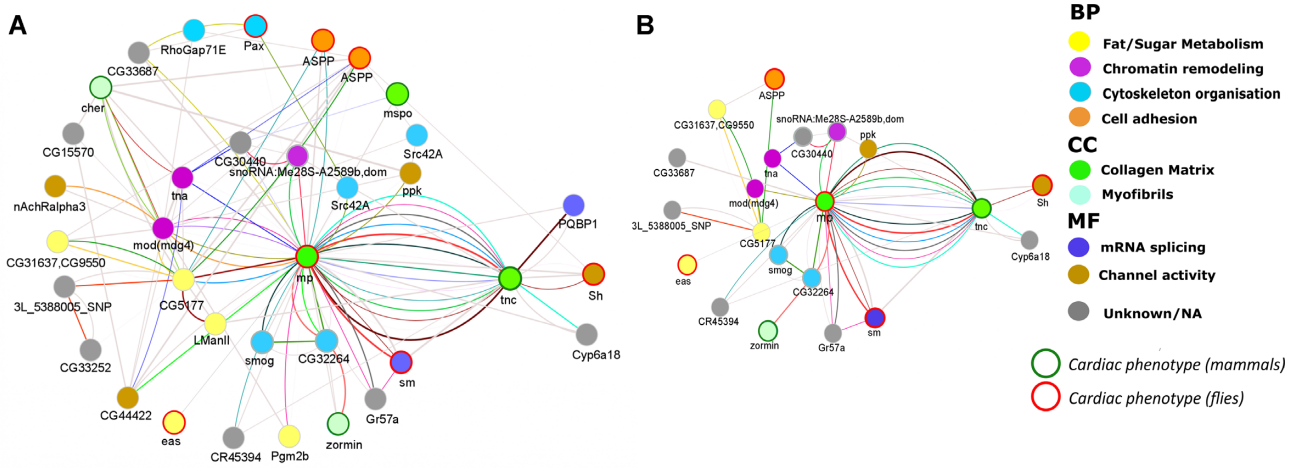



**Figure 5.** Weighted *epiMEIF* high-order statistical interactions detected on natural variation of heart period in flies with biological annotations: (**A**) shows the network obtained when weighted *epiMEIF* is fitted on the cardiac heart period of the DGRP population at 1 week. The different nodes denote the different variants in the weighted *epiMEIF* network, and wherever possible, they are annotated based on the genes to which the variants are mapped. The nodes are coloured according to their cellular and molecular functions, and the colour legend on the right denotes the different cellular and molecular processes (BP: biological process, CC: cellular component, MF: molecular function) expressed in the *epiMEIF* network. The coloured boundary of the node denotes if the annotated genes have mammal orthologs associated with the cardiac phenotype. (**B**) Part of the network in (A) that can be validated with the validation cohort using ANOVA/Max-T test.

genetically with *Pax*. *Pax* encodes an adaptor protein that couples integrins to the actin cytoskelon (39). *epiMEIF* interactions may thus point to a central role of the collagen-containing ECM in the mechanisms leading to the variability of HP phenotype, probably by impacting cytoskeleton dynamics. In addition, the activity of the non-receptor tyrosine kinase *Src42A*, which is involved in several cellular processes including cell adhesion and cytoskeleton organization, was shown to be regulated by *Ankyrin-repeat, SH3-domain* and *Proline-rich-region containing protein (ASPP)/PPP1R13B* (40). Of note, several SNPs in both *ASPP* and *Src42A* are present in the network. Moreover, the network highlights other important features of natural variations of cardiac function, such as variants in *easily shocked (eas)/ETNK1*, an ethanolamine kinase with essential cardiac function (41), in *Shaker* (*Sh*)/*KCNA1*, a voltage-gated potassium channel encoding gene involved in setting the cardiac rhythm (42). Finally, it is worth noting that variants in *tnc* and *mp* interact with variants within *smooth* (*sm*)/*HNRNPL* and *poly-glutamine tract binding protein 1 (PQBP1)*, two genes involved in mRNA splicing, suggesting that regulation of ECM components through mRNA splicing impinges on natural variation of HP.

*Validation of the results.* We then tested if the *epiMEIF* network for heart period was replicated in an independent cohort. Twenty DGRP lines—not included in the first dataset and therefore not used in building the statistical interaction network—were analysed for cardiac phenotypes at 1 week. We used Max-T and ANOVA tests to test whether *epiMEIF* interactions were replicated in this independent population. Note that the interaction network comprised 45 binary interactions that are not part of any higher order interactions, 33 three-way interactions that are not part of any four-way interactions, and 2 four-way interactions (see Supplementary Figure S8a and Supplementary Table S2). We retained those interactions which performed well either with the ANOVA test or Max-T test. Note that approximately 18% (8/45) two-way and 5% (2/33) 3-way interactions could not be tested with the validation cohort due to a lack of observations. Amongst the one tested 14% (5/37) two-way interactions, 42% (13/31) 3-way interactions, and 100% (2/2) four-way interactions could be validated (see Supplementary Figure S8b and Supplementary Table S2). The biological significance of the part of the *epiMEIF* network that can be validated with the validation cohort is represented in Figure 5B. Note that only a part of the network could be validated because the reduced number of lines in the validation cohort made it less powered to test all the interactions. Nevertheless, the main characteristics of the network (in Figure 5A) are found reproduced in this independent population, including the centrality of *multiplexin (mp)* and its interactions with *tenectin* (*tnc*) and *Shaker (Sh)* (see Figure 5B). More details on the interactions in the network can be found in Supplementary Table S3.

## Applications on longitudinal data

We have shown the efficacy of our approach with cross-sectional datasets. Noticeably, the *epiMEIF* method is also applicable for detecting genetic interactions from longitudinal datasets. We have rarely encountered genetic interactions being tested on longitudinal datasets in literature. Malzahn *et al.*, (43) tested for gene-gene interaction using a longitudinal non-parametric association test on Framingham Heart Study (FHS) cohorts, but they did not test for higher-order interactions. Our approach is particularly interesting because it also presents a way to explore higher-order interactions in longitudinal data. Similar to the cross-section data above, we have tested the efficacy of our ap-

proach in longitudinal datasets *via* simulations and real data applications.

*Analysis on simulated data.* We evaluated the statistical power of the method on longitudinal datasets using simulation scenarios similar to those used with cross-section data. Here as well, we have three simulation scenarios—AS1, AS2 and AS3. While AS1 considers the simplest scenario with 1 binary interaction, AS2 considers a comparatively complicated scenario with 2 binary SNP interactions. AS2 is designed as earlier; one set of SNPs involved in the interaction has a lower marginal effect and the other set of SNPs has a higher marginal effect. AS3 comprises 1 three-SNP interaction (see Supplementary Materials for more details). For longitudinal data, we performed two sets of simulations. The first set of simulations comprises aging data for two-time points (1 week and 4 weeks) and the second set comprises aging data for three-time points (1 week, 4 weeks and 7 weeks). The latter is added to demonstrate that the method is applicable for complex longitudinal datasets as well (with more than two time points). We have conducted our simulations 100 times for each scenario, and we run each scenario with 30 or 50 additional variants, randomly chosen from the genome data (with 1000 SNPs), as performed earlier for cross-sectional data simulations. Note that we have conducted the longitudinal simulations only with *epiMEIF*. Calibrating *weighted epiMEIF* for the longitudinal dataset can be tricky as deciding a weight for the 'Time' covariate is challenging, as discussed earlier.

As in the previous simulations, we evaluated the performance of *epiMEIF* on longitudinal data based on 'the overall power in capturing the true epistatic interactions' (see Table 2) and the power to capture the true interactions as top-ranking (see Supplementary Figure S9). The 'overall power' is quite high (95–98%) across all the scenarios (see Table 2). Both the 'overall power' and the 'power to capture the true interactions in the top ranks' are similar across the two situations- with 30 and 50 additional variants (see Table 2, Supplementary Figure S9). However, *epiMEIF* is less efficient in capturing the interactions involving SNPs with lower marginal effect (power to capture the true interactions in top ranks is as low as 10–20% for the interaction pair 1–2 in scenario AS2 in Supplementary Figure S9). For the longitudinal simulations with three-time points, we evaluated the performance of *epiMEIF* for scenarios AS1 and AS3 based on the power to capture the true interactions as top-ranking (see Supplementary Figure S10). Here as well, the 'overall power' and the 'power to capture' the top ranks are similar across the two situations–with 30 and 50 additional variants. The 'overall' power is ∼90–93% for AS1 and 60% for AS3.

Overall, the power to capture true interactions is quite satisfactory across all the scenarios and exploring higher-order interactions with synthetic longitudinal data presented here is novel in GWAS.

*Analysis on real data:* epiMEIF *network construction for natural variations of heart period aging in flies.* To further evaluate our method for identifying epistatic interactions from longitudinal data, we analyzed the aging of the heart function in the DGRP population. Several cardiac aging studies in Drosophila have revealed striking similarities with mammals, both in terms of heart physiology and transcriptional changes (34,35,44,45). This highlights the conserved nature of cardiac aging across organisms. We analysed 165 of the DGRP fly lines included in the previous dataset for the heart period at 4 weeks of age (12 individuals per line). Overall, there is a marked increase in heart period in the DGRP from 1 week to 4 weeks, in agreement with the previous observation (see Supplementary Figure S4). Both 1-week and 4-weeks data on the 165 DGRP lines were used to identify epistatic interactions between variants associated with natural variations of heart period during aging. Similar to the previous 1-week study, *epiMEIF* interactions were identified on variants that show significant association with the aging of HP from the single GWAS longitudinal LMM, with a nominal significance threshold of $10^{-5}$ (1682 *SNPs*). Finally, *epiMEIF* application on HP aging dataset led to a dense network of 26 interacting SNPs; comprising 47 two-way that are not part of any three-way interactions and 12 3-way interactions (Figure 6A, Supplementary Table S4). Strikingly, the *epiMEIF* network comprises tightly interconnected variants within genes involved in diverse biological processes, many of which have characterized cardiac function, either in flies or in mammals. In particular, the network includes variants in 3 genes encoding signalling pathway components, namely *Allatostatin C receptor 2 (Astc-R2)*, the somatostatine receptor orthologue; *red dog mine (Rdog)*, an ATPase-coupled transmembrane transporter orthologous to *ABCC4*; and *CG15611,* the Rho guanine nucleotide exchange factor *ARHGEF25* orthologue. Noticeably, *rdog* and *CG15611* orthologues have known autocrine or paracrine cardiac functions either in humans or mice (46–48). This suggests a primary role of natural variations of cardiac signalling properties in heart senescence. In addition, the numerous interactions that these SNPs engage in within the network suggest their involvement in relation to the other variants associated with the aging of cardiac function. Among these, one SNP into *Dpr-interacting protein κ (DIP-kappa)*—encoding a cell adhesion protein orthologous to *LSAMP*- interacts with 13 variants into 12 genes. One SNP into *CG4704*, which encodes the fly orthologue of the human *MCU1* (Mitochondrial Calcium Uptake 1)—interacts with 10 variants into 9 genes. *MCU1* is a regulatory subunit of the Mitochondrial Calcium Uniporter, which plays a central role in calcium import into the mitochondrion and in mitochondrial calcium

**Table 2.** Proportion of simulations (out of 100) capturing the true interactions for the different Aging simulation scenarios under the two additional sample size cases (30 and 50)

| | #Captured (%) | |
|---|---|---|
| #Additional SNPs<br>Method | 30<br>MEIF | 50<br>MEIF |
| AS 1 | 97 | 97 |
| AS 2 (SNP_1:SNP_2) | 98 | 98 |
| AS 2 (SNP_5:SNP_6) | 97 | 98 |
| AS 3 | 96 | 94 |

*#Captured (%): $\frac{\sum_i I(\text{interaction detected in } i^{th} \text{ simulation})}{100}$, where $I$ is an indicator variable.

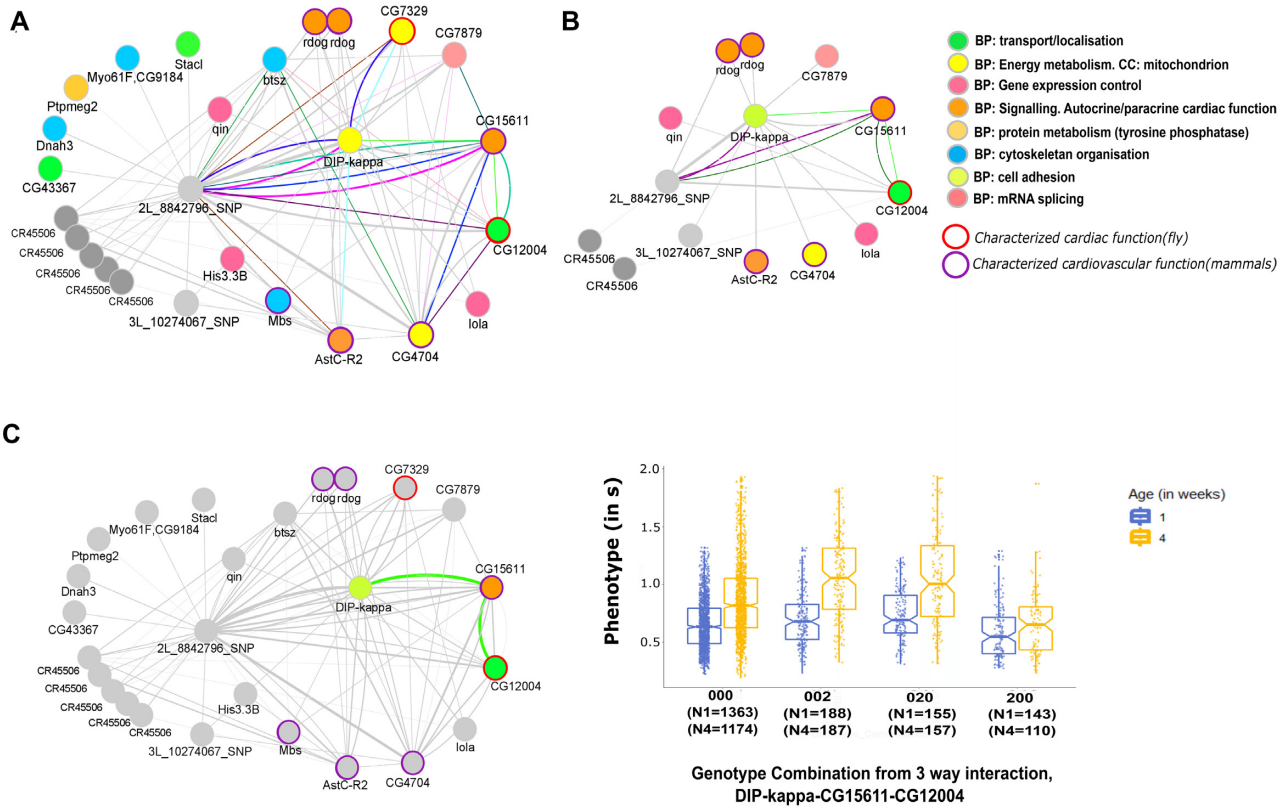



**Figure 6.** *epiMEIF* high-order statistical interactions detected on natural variation of heart period aging in flies: (**A**) network obtained when *epiMEIF* is fitted on the cardiac heartperiod aging data of the DGRP population (flies at 1 and 4 weeks). The different nodes denote the different variants in the *epiMEIF* network, annotated based on the genes to which the variants are mapped. The nodes are coloured according to their cellular and molecular functions. The colour legend on the right denotes the different cellular and molecular processes (BP: biological processes, CC: cellular component). The coloured boundary of the node denotes if the annotated genes have mammal orthologs associated with cardiac phenotype. (**B**) Part of the *epiMEIF* network in (A) that can be validated with the validation cohort using ANOVA/Max-T test. (**C**) A 3-way interaction between DIP-kappa-CG15611-CG12004 is highlighted here. The boxplot distribution of the phenotype at 1 and 4 weeks against the different genotype combinations shows the aging effect of the three-way interaction on the cardiac phenotype. The *P*-value from the Max-T test is <0.001.

ion homeostasis (49). This suggests a central function of these components in the heart period aging. The most connected node corresponds to a SNP that is more than 10 kb away from any gene, precluding its annotation. The network additionally includes SNPs in genes involved in cytoskeleton organization (*Myosin binding subunit (Mbs)*, *bitesize (btsz)*), carbohydrate transport (*CG12004/TMEM184*), energy metabolism (*CG7329/LIPM*) and mRNA splicing (*CG7879/RBM12*). Strikingly, five variants within 150 bp upstream of the lncRNA *CR45506* are retrieved in the network, suggesting a major involvement for this lncRNA in the process, throughout interactions with several other members of the network. Taken together, these SNPs and their interactions allow us to draw some characteristics of the genetic architecture of natural variations in the aging of cardiac function. More details on the interactions in the network can be found in Supplementary Table S5.

*Validation of the results.* The twenty DGRP lines not included in the first dataset were also analyzed at 4 weeks of age, thus providing a validation set for cardiac aging which was used to test for replication of the interaction using Max-T/ANOVA test in this independent cohort. 75% (35/47) two-way interactions and 58% (7/12) three-way interactions could be analysed in this validation cohort (see Supplementary Table S6). Of them, 43% two-way (15/35) and 43% three-way (3/7) were replicated (Figure 6B). Several features of the network were replicated in the independent cohort, including the involvement of *rdog*, *CG15611* and *Astc-R2*. This also confirmed the central positioning of *DIP-kappa*, whose implication in natural variations of cardiac aging, therefore, warrants further investigations.

## DISCUSSION

A new method is proposed here for epistasis detection in large-scale association studies with complex genetic traits using mixed effect conditional inference forest (*epiMEIF*). This method captures higher-order SNP interactions based on the tree structure in the cforest, and combined with mixed models, can handle a wide range of complex GWA studies. The effectiveness of *epiMEIF* is verified in extensive simulation scenarios reflecting a wide spectrum of complex models and with real datasets, illustrating its power for epistasis detection from both cross-sectional and longitudinal data. The additional testing strategies applied a posteriori to the conditional inference forest in *epiMEIF*

not only safeguards the method from detecting false positive interactions but also increases the reliability of the final selected interactions. The ability of the approach to validate part of the higher-order interactions in an independent cohort also supports its' competency in detecting higher-order interactions. Additionally, for the cross-sectional datasets, we proposed an adaptation of *epiMEIF*, named *weighted epiMEIF* that allows identifying genes associated with weak marginal effect variants. This is an important addition to the 'traditional' epistasis approaches that are biased towards variants with strong marginal effects (7,8,50). Similar to LMM, a major advantage of *epiMEIF* is its' ability to effectively account for unwanted correlation between samples, thereby correcting for confounding factors such as population structure (51) or hidden covariates (52) and making it easily applicable for human GWAS studies as well. However, unlike the standard LMM, *epiMEIF* can jointly model the genetic effects of multiple loci or markers on the readout. This is particularly important because recent works have revealed that often, the single-locus association models are insufficient to explain the heritable component of complex traits (53,54). The proposed approach has proven to overcome the caveats of existing GWAS approaches and though we have shown its applicability for identifying associations and epistasis detection, it can be also used for prediction (55,56) and feature selection (50,57). We however acknowledge that the method is not scalable to the entire genome dataset and does not provide an exhaustive list of interactions from the entire list of variants. Interestingly, the *epiMEIF* generates dense genetic interaction networks that are creating hubs around some focal genes. Part of this distinctive topology may be attributable to the nature of tree formation in the cforest algorithm. It is difficult to determine if this topology is also due to the nature of the genetic interactions that may underlie this statistical network. The analysis of the networks obtained from the study of natural variation in cardiac function undoubtedly sheds light and should indicate that the statistical interactions reveal an underlying biological functionality. Indeed, the vast majority of the interactions identified affect genes whose products are involved in the cytoskeleton (sarcomeric and non-sarcomeric) and its dynamics and, in interaction with the extracellular matrix, must participate in the mechanical and signalling properties of cardiomyocytes.

In systems biology, there is an emerging interest in understanding the genetic mechanisms underlying the study of longitudinally measured phenotypes (58,59). Additionally, it has been often discovered that complex diseases are the results of interaction between a large number of units, and therefore, are more likely to be associated with genes that are well connected in the network (60). *epiMEIF* can comprehensively model the dynamics of longitudinal data and cross-sectional data and offer to reveal the relationships between multiple variants, revealing the extensive networks of genetic interactions that are causing the complex diseases. The possibilities offered by *epiMEIF* will allow approaching questions that were previously difficult to address, as tools for the identification of epistatic interactions in complex datasets were lacking. In particular, we do not know when biomolecular interactions produce patterns of statistical epistasis, nor do we know how to biologically interpret statistical evidence of epistasis, since statistical interaction likely does not automatically entail interaction at a biological or mechanistic level. Progress towards these important questions will provide a framework for using genetic information to improve our ability to diagnose, prevent and treat common human diseases (61). Hence, our approach provides the foundation for extracting higher-order statistical interactions flexibly from any type of dataset, that will guide the biologists in formulating their hypothesis. Eventually, network analysis tools (62) may be useful to biologically interpret the statistical evidence of epistasis and bridge the gap between statistical and biological epistasis networks.

## DATA AVAILABILITY

The proposed statistical methods are implemented in R and the source codes and datasets can be found in the Github repository (https://github.com/TAGC-NetworkBiology/epiMEIF).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Fabiana Rossi for her assistance on data analysis. We would like to acknowledge Centre de Calcul Intensif d'Aix-Marseille for granting access to its high-performance computing resources.

## FUNDING

The project leading to this publication has received funding from the ≪ Investissements d'Avenir ≫ French Government program managed by the French National Research Agency [ANR-16-CONV-0001], from Excellence Initiative of Aix-Marseille University - A*MIDEX and from Fondation de France [00071034]; Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high-performance computing resources. Funding for open access charge: ≪ Investissements d'Avenir ≫ French Government program managed by the French National Research Agency [ANR-16-CONV-0001]; Excellence Initiative of Aix-Marseille University - A*MIDEX.
*Conflict of interest statement*. None declared.

## REFERENCES

1. Korte,A. and Farlow,A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, **9**, 29.
2. Uffelmann,E., Huang,Q.Q., Munung,N.S., de Vries,J., Okada,Y., Martin,A.R., Martin,H.C., Lappalainen,T. and Posthuma,D. (2021) Genome-wide association studies. *Nat. Rev. Methods Primers*, **1**, 59.
3. Tam,V., Patel,N., Turcotte,M., Bossé,Y., Paré,G. and Meyre,D. (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*, **20**, 467–484.
4. Niel,C., Sinoquet,C., Dina,C. and Rocheleau,G. (2015) A survey about methods dedicated to epistasis detection. *Front. Genet.*, **6**, 285.
5. Lander,E. and Kruglyak,L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.*, **11**, 241–247.
6. Glazier,A.M. (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.

7. Jiang,R., Tang,W., Wu,X. and Fu,W. (2009) A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinf.*, **10**, S65.
8. Yoshida,M. and Koike,A. (2011) SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinf.*, **12**, 469.
9. Chattopadhyay,A. and Lu,T.-P. (2019) Gene-gene interaction: the curse of dimensionality. *Ann. Transl. Med.*, **7**, 813–813.
10. Wan,X., Yang,C., Yang,Q., Xue,H., Tang,N.L.S. and Yu,W. (2010) Predictive rule inference for epistatic interaction detection in genome-wide association studies. *Bioinformatics*, **26**, 30–37.
11. Schüpbach,T., Xenarios,I., Bergmann,S. and Kapur,K. (2010) FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, **26**, 1468–1469.
12. Wan,X., Yang,C., Yang,Q., Xue,H., Fan,X., Tang,N.L.S. and Yu,W. (2010) BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.
13. Bayat,A., Hosking,B., Jain,Y., Hosking,C., Kodikara,M., Reti,D., Twine,N.A. and Bauer,D.C. (2021) Fast and accurate exhaustive higher-order epistasis search with BitEpi. *Sci. Rep.*, **11**, 15923.
14. Yang,C., He,Z., Wan,X., Yang,Q., Xue,H. and Yu,W. (2009) SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, **25**, 504–511.
15. Hemani,G., Theocharidis,A., Wei,W. and Haley,C. (2011) EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics*, **27**, 1462–1465.
16. Calle,M.L., Urrea Gales,V., Malats i Riera,N. and Van Steen,K. (2008) MB-MDR: model-based multifactor dimensionality reduction for detecting interactions in high-dimensional genomic data. ResearchGate doi: https://www.researchgate.net/publication/36731394, January 2007, preprint: not peer reviewed.
17. Cattaert,T., Calle,M.L., Dudek,S.M., Mahachie John,J.M., Van Lishout,F., Urrea,V., Ritchie,M.D. and Van Steen,K. (2011) Model-Based multifactor dimensionality reduction for detecting epistasis in case-control data in the presence of noise: MB-MDR for case-control data with errors. *Ann. Hum. Genet.*, **75**, 78–89.
18. Zhang,X., Huang,S., Zou,F. and Wang,W. (2010) TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, **26**, i217–i227.
19. Culverhouse,R., Klein,T. and Shannon,W. (2004) Detecting epistatic interactions contributing to quantitative traits. *Genet. Epidemiol.*, **27**, 141–152.
20. Schwarz,D.F., König,I.R. and Ziegler,A. (2010) On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, **26**, 1752–1758.
21. Hajjem,A., Bellavance,F. and Larocque,D. (2014) Mixed-effects random forest for clustered data. *J. Stat. Comput. Simul.*, **84**, 1313–1328.
22. Yang,J., Lee,S.H., Goddard,M.E. and Visscher,P.M. (2011) GCTA: a tool for Genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
23. Zhou,X. and Stephens,M. (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.*, **44**, 821–824.
24. Hothorn,T. and Zeileis,A. (2015) partykit: a modular toolkit for recursive partytioning in R. *J. Mach. Learn. Res.*, **16**, 3905–3909.
25. Breiman,L. (2001) Random forest. *Mach. Learn.*, **45**, 5–32.
26. Genuer,R., Poggi,J.-M. and Tuleau-Malot,C. (2010) Variable selection using random forests. *Pattern Recognit. Lett.*, **31**, 2225–2236.
27. Lippert,C., Listgarten,J., Liu,Y., Kadie,C.M., Davidson,R.I. and Heckerman,D. (2011) FaST linear mixed models for genome-wide association studies. *Nat. Methods*, **8**, 833–835.
28. Yao,C., Spurlock,D.M., Armentano,L.E., Page,C.D., VandeHaar,M.J., Bickhart,D.M. and Weigel,K.A. (2013) Random forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *J. Dairy Sci.*, **96**, 6716–6729.
29. Saha,S. and Brannath,W. (2020) Point and interval estimation of the target dose using weighted regression modelling. arXiv doi: https://doi.org/10.48550/arXiv.2007.05974, 12 July 2020, preprint: not peer reviewed.
30. Cordell,H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468.
31. Cordell,H.J. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
32. Mackay,T.F.C., Richards,S., Stone,E.A., Barbadilla,A., Ayroles,J.F., Zhu,D., Casillas,S., Han,Y., Magwire,M.M., Cridland,J.M. *et al.* (2012) The Drosophila melanogaster genetic reference panel. *Nature*, **482**, 173–178.
33. Saha,S., Spinelli,L., Castro-Mondragon,J.A., Kervadec,A., Kremmer,L., Roder,L., Krifa,S., Torres,M., Brun,C., Vogler,G. *et al.* (2021) Genetic architecture of natural variation of cardiac performance in flies genetics. bioRxiv doi: https://doi.org/10.1101/2021.06.08.447524, 07 February 2022, preprint: not peer reviewed.
34. Ocorr,K., Perrin,L., Lim,H.-Y., Qian,L., Wu,X. and Bodmer,R. (2007) Genetic control of heart function and aging in Drosophila. *Trends Cardiovasc. Med.*, **17**, 177–182.
35. Blice-Baum,A.C., Guida,M.C., Hartley,P.S., Adams,P.D., Bodmer,R. and Cammarato,A. (2019) As time flies by: investigating cardiac aging in the short-lived Drosophila model. *Biochim. Biophys. Acta (BBA) - Mol. Basis Dis.*, **1865**, 1831–1844.
36. Seyres,D., Roder,L. and Perrin,L. (2012) Genes and networks regulating cardiac development and function in flies: genetic and functional genomic approaches. *Brief. Funct. Genomics*, **11**, 366–374.
37. Sequeira,V., Nijenkamp,L.L.A.M., Regan,J.A. and van der Velden,J. (2014) The physiological role of cardiac cytoskeleton and its alterations in heart failure. *Biochim. Biophys. Acta*, **1838**, 700–722.
38. Yan,L., Wei,X., Tang,Q.-Z., Feng,J., Zhang,Y., Liu,C., Bian,Z.-Y., Zhang,L.-F., Chen,M., Bai,X. *et al.* (2011) Cardiac-specific mindin overexpression attenuates cardiac hypertrophy via blocking AKT/GSK3β and TGF-β1-Smad signalling. *Cardiovasc. Res.*, **92**, 85–94.
39. Zervas,C.G., Psarra,E., Williams,V., Solomon,E., Vakaloglou,K.M. and Brown,N.H. (2011) A central multifunctional role of integrin-linked kinase at muscle attachment sites. *J. Cell Sci.*, **124**, 1316–1327.
40. Langton,P.F., Colombani,J., Aerne,B.L. and Tapon,N. (2007) Drosophila ASPP regulates C-terminal Src kinase activity. *Dev. Cell*, **13**, 773–782.
41. Lim,H.-Y., Wang,W., Wessells,R.J., Ocorr,K. and Bodmer,R. (2011) Phospholipid homeostasis regulates lipid metabolism and cardiac function through SREBP signaling in Drosophila. *Genes Dev.*, **25**, 189–200.
42. Johnson,E., Ringo,J., Bray,N. and Dowse,H. (1998) Genetic and pharmacological identification of ion channels central to the *Drosophila* cardiac pacemaker. *J. Neurogenet.*, **12**, 1–24.
43. Malzahn,D., Balavarca,Y., Lozano,J.P. and Bickeböller,H. (2009) Tests for candidate-gene interaction for longitudinal quantitative traits measured in a large cohort. *BMC Proc*, **3**, S80.
44. Monnier,V., Iché-Torres,M., Rera,M., Contremoulins,V., Guichard,C., Lalevée,N., Tricoire,H. and Perrin,L. (2012) dJun and Vri/dNFIL3 are major regulators of cardiac aging in Drosophila. *PLos Genet.*, **8**, e1003081.
45. Cannon,L., Zambon,A.C., Cammarato,A., Zhang,Z., Vogler,G., Munoz,M., Taylor,E., Cartry,J., Bernstein,S.I., Melov,S. *et al.* (2017) Expression patterns of cardiac aging in *Drosophila*. *Aging Cell*, **16**, 82–92.
46. Sassi,Y., Abi-Gerges,A., Fauconnier,J., Mougenot,N., Reiken,S., Haghighi,K., Kranias,E.G., Marks,A.R., Lacampagne,A., Engelhardt,S. *et al.* (2012) Regulation of cAMP homeostasis by the efflux protein MRP4 in cardiac myocytes. *FASEB J.*, **26**, 1009–1017.
47. Sassi,Y., Ahles,A., Truong,D.-J.J., Baqi,Y., Lee,S.-Y., Husse,B., Hulot,J.-S., Foinquinos,A., Thum,T., Müller,C.E. *et al.* (2014) Cardiac myocyte–secreted cAMP exerts paracrine action via adenosine receptor activation. *J. Clin. Invest.*, **124**, 5385–5397.
48. Ongherth,A., Pasch,S., Wuertz,C.M., Nowak,K., Kittana,N., Weis,C.A., Jatho,A., Vettel,C., Tiburcy,M., Toischer,K. *et al.* (2015) p63RhoGEF regulates auto- and paracrine signaling in cardiac fibroblasts. *J. Mol. Cell Cardiol.*, **88**, 39–54.
49. Garbincius,J.F., Luongo,T.S. and Elrod,J.W. (2020) The debate continues – what is the role of MCU and mitochondrial calcium uptake in the heart?*J. Mol. Cell Cardiol.*, **143**, 163–174.

50. Bureau,A., Dupuis,J., Falls,K., Lunetta,K.L., Hayward,B., Keith,T.P. and Van Eerdewegh,P. (2005) Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.*, **28**, 171–182.
51. Jamrozik,J. and Schaeffer,L.R. (1997) Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation holsteins. *J. Dairy Sci.*, **80**, 762–770.
52. Fusi,N., Stegle,O. and Lawrence,N.D. (2012) Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comput. Biol.*, **8**, e1002330.
53. Bloom,J.S., Ehrenreich,I.M., Loo,W.T., Lite,T.-L.V. and Kruglyak,L. (2013) Finding the sources of missing heritability in a yeast cross. *Nature*, **494**, 234–237.
54. Pickrell,J., Clerget-Darpoux,F. and Bourgain,C. (2007) Power of genome-wide association studies in the presence of interacting loci. *Genet. Epidemiol.*, **31**, 748–762.
55. Stephan,J., Stegle,O. and Beyer,A. (2015) A random forest approach to capture genetic effects in the presence of population structure. *Nat. Commun.*, **6**, 7432.
56. Botta,V., Louppe,G., Geurts,P. and Wehenkel,L. (2014) Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS One*, **9**, e93379.
57. Szymczak,S., Holzinger,E., Dasgupta,A., Malley,J.D., Molloy,A.M., Mills,J.L., Brody,L.C., Stambolian,D. and Bailey-Wilson,J.E. (2016) r2VIM: a new variable selection method for random forests in genome-wide association studies. *BioData Mining*, **9**, 7.
58. Lugo-Martinez,J., Ruiz-Perez,D., Narasimhan,G. and Bar-Joseph,Z. (2019) Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, **7**, 54.
59. Wang,H., Nie,F., Huang,H., Yan,J., Kim,S., Nho,K., Risacher,S.L., Saykin,A.J., Shen,L. and for the Alzheimer's Disease Neuroimaging Initiativefor the Alzheimer's Disease Neuroimaging Initiative (2012) From phenotype to genotype: an association study of longitudinal phenotypic markers to Alzheimer's disease relevant SNPs. *Bioinformatics*, **28**, i619–i625.
60. Liu,X., Maiorino,E., Halu,A., Glass,K., Prasad,R.B., Loscalzo,J., Gao,J. and Sharma,A. (2020) Robustness and lethality in multilayer biological molecular networks. *Nat. Commun.*, **11**, 6043.
61. Moore,J.H. and Williams,S.M. (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays*, **27**, 637–646.
62. Battiston,F., Nicosia,V. and Latora,V. (2014) Structural measures for multiplex networks. *Phys. Rev. E*, **89**, 032804.


## APPENDIX 1. MEIF ALGORITHM

- Step 1: Fit the MEIF as shown in equation (1).
- Step 2: Extract the cforest component $\hat{f}(S_1, S_2, \cdots, S_N)$ from the fitted MEIF model, where in each forest the decision is taken based on the cumulative decision from n trees. Following the mechanism in Figure 2A each tree gives rise to a potential interaction set $\{Q_t, t \in 1, 2, \cdots, n\}$. Compiling all the interactions sets from the $n$ trees $\{Q_t, t \in 1, 2, \cdots, n\}$, SNP Interaction Matrix ($I$) is computed, that measures the strength of each interaction set based on their frequency of occurrence in the cforests.
- Step 3: Since cforest is a bagging algorithm where the individual trees are built from bootstrap samples, different iteration of cforest may give rise to different SNP Interaction Matrix ($I$). Hence, to enhance the stability of the final interactions sets, we propose to fit the MEIF/cforest 10 times and then obtained the interaction sets that are occurring with high scores in 90% of the forests (9/10 forest).
- Step 4: The interaction sets along with their pooled interaction score from the 10 forests are utilized to construct statistical epistatic clusters (see Figure 1).