BIOINFORMATICS

Vol. 00 no. 00 2014
Pages 1–7

EPIQ - Efficient detection of SNP-SNP epistatic interactions for quantitative traits

Arkin Ya'ara [1], Rahmani Elior [1], Kleber E. Marcus [4], Laaksonen Reijo [5,6], Maerz Winfried [4,7,8], Halperin Eran [1,2,3,*]

[1]The Blavatnik School of Computer Science, Tel Aviv University, Tel-Aviv 69978, Israel
[2]Department of Molecular Microbiology and Biotechnology, George Wise Faculty of Life Science, Tel-Aviv University, Tel-Aviv 69978, Israel
[3]International Computer Science Institute, Berkeley, CA 94704, USA
[4]V[th] Department of Medicine (Nephrology, Hypertensiology, Endocrinology, Diabetology, Rheumatology), Medical Faculty of Mannheim, University of Heidelberg, Mannheim D-68167, Germany
[5]Zora Biosciences Oy, Espoo 02150, Finland
[6]University of Tampere, Tampere 33104, Finland
[7]Clinical Institute of Medical and Chemical Laboratory Diagnostics, Medical University of Graz, Graz A-8036, Austria
[8]Synlab Academy, Synlab Services GmbH, Mannheim D-68165, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Gene-gene interactions are of potential biological and medical interest, as they can shed light on both the inheritance mechanism of a trait and on the underlying biological mechanisms. Evidence of epistatic interactions has been reported in both humans and other organisms. Unlike single-locus genome wide association studies (GWAS), which proved efficient in detecting numerous genetic loci related with various traits, interaction-based GWAS have so far produced very few reproducible discoveries. Such studies introduce a great computational and statistical burden by necessitating a large number of hypotheses to be tested including all pairs of SNPs. Thus, many software tools have been developed for interaction-based case-control studies, some leading to reliable discoveries. For quantitative data, on the other hand, only a handful of tools exist, and the computational burden is still substantial.

Results: We present an efficient algorithm for detecting epistasis in quantitative GWAS, achieving a substantial runtime speedup by avoiding the need to exhaustively test all SNP pairs using metric embedding and random projections. Unlike previous metric embedding methods for case-control studies, we introduce a new embedding, where each SNP is mapped to two Euclidean spaces. We implemented our method in a tool named EPIQ (EPIstasis detection for Quantitative GWAS), and we show by simulations that EPIQ requires hours of processing time where other methods require days and sometimes weeks. Applying our method to a dataset from the Ludwigshafen Risk and Cardiovascular Health study discovered a pair of SNPs with a near-significant interaction (p=2.2×10^{-13}), in only 1.5 hours on 10 processors.

Availability: XXXX

*to whom correspondence should be addressed

Contact: heran@post.tau.ac.il

1 INTRODUCTION

Genome wide association studies (GWAS) have so far detected thousands of single nucleotide polymorphism (SNP) loci that are associated with various traits (Hindorff *et al.*, 2009). Unfortunately, for most complex traits the discovered SNPs explain only a small fraction of the estimated heritability, a phenomena often referred to as the 'missing heritability' (Maher, 2008). One plausible explanation suggested for this problem is the existence of an epistatic effect, where two or more loci have a synergetic influence on the phenotype, also referred to as gene-gene interactions (Maher, 2008). The discovery of interacting SNP loci has an additional benefit, as it may shed light on the underlying biological mechanism or involved pathways.

Despite evidences of gene-gene interactions reported both in human and in other organisms (Evans *et al.*, 2006), very few reproducible discoveries were reported by GWAS (Prabhu and Pe'er 2012; Liu *et al.* 2011 for example). The amount of data produced in a single study is a possible cause: When searching for groups of k SNPs with an epistatic effect, the number of possible k-sized groups is $\Theta(m^k)$, where m is the number of SNP loci. With current GWAS typically including hundreds of thousands of SNPs, this implies both a computational and statistical burden even for groups sizes as small as $k = 2$: The numerous tests takes days and even weeks to compute and require a substantial correction for multiple hypothesis, leading in some cases to a loss of power (Evans *et al.*, 2006). One common approach is the reduction of the search space, usually by filtering candidate loci pairs: Marchini

et al. (2005) suggested selecting a subset of SNPs with a moderate marginal effect and testing for interaction in pairs where at least one locus is included in the subset. Reduction of the search space can also be done by manipulating contingency tables (Wan *et al.*, 2010; Zhang *et al.*, 2010) or searching for a linkage-disequilibrium (LD) contrast between cases and controls (Prabhu and Pe'er, 2012; Brinza *et al.*, 2010). A more straightforward approach is increasing the computational power, either by multi-threaded implementations or by utilizing special hardware (Hu *et al.*, 2010; Yung *et al.*, 2011). Binary operations are used in some cases to speedup performance (Prabhu and Pe'er, 2012; Wan *et al.*, 2010).

All of these tools, and many others, are designed for case-control studies, whereas for the quantitative case, where the tested phenotypes are physiological measurements of some sort, the selection of available software is limited. Since the phenotype tested is not dichotomous, testing for quantitative associations can be more challenging compared to case-control studies, as methods utilizing contingency tables, LD-contrast or binary operations are usually inapplicable. Methods tailored for case-control studies can be applied on quantitative traits after dichotomizing the phenotype (as in Bhattacharya *et al.* 2011), however, the resulting statistical test is different than the original, thus a loss of power is inevitable and would be difficult to quantify.

In this study we present EPIQ (EPIstasis detection for Quantitative GWAS) - an efficient algorithm for detecting pairs of SNP loci that have an epistatic effect on quantitative phenotypes. EPIQ achieves a substantial runtime speedup by avoiding the need to exhaustively test all SNP pairs: It applies a carefully chosen transformation that maps each genotyped SNP to a vector in a Euclidean space. This transformation has the property that SNP pairs with an epistatic effect are converted to vector pairs with a large inner product. A random projections method is subsequently applied to efficiently recover these SNPs. A novelty of our method is that each SNP is projected to two different points, for a more efficient detection of interacting SNPs. We show on simulated data that in just over 3 hours our algorithm was able to process a dataset that would take days or weeks using state of the art software, and present the results of running EPIQ on data from the Ludwigshafen Risk and Cardiovascular (LURIC, Winkelmann *et al.* (2001)) health study.

2 METHODS

Outline

EPIQ is designed to efficiently discover SNPs that have a significant epistatic effect over a quantitative phenotype, without exhaustively testing all pairs of SNPs in a dataset. This goal is achieved in two steps: a filtering stage - generating a list of candidate SNP pairs, and a validation stage - fitting a linear regression model to these pairs. By shortening the list of pairs for to be tested during the filtering stage, running time for the linear regression step is reduced substantially. Filtering is performed by assigning a score to each SNP; this score is stochastically generated so that for each pair of SNPs, the expected value for the product of their scores is proportional to the generalized likelihood ratio (GLR) test statistic of their interaction. This means epistatic pairs are expected to have a high score product. By performing multiple iterations and collecting pairs that pass a given threshold, we assure with high probability that if an interacting pair exists, it is included in the candidates list and will be reported during validation stage. To do so we present a new test-statistic τ^2 which is roughly proportional to the GLR test score, and apply a random projection algorithm that discovers pairs with exceptionally high τ^2 scores.

Model description

Model input EPIQ receives as input a vector $\mathbf{y} \in \mathbb{R}^n$, representing the phenotypic values of all n individuals in the cohort, and a matrix $\mathbf{X}_{n\times m} \in \{0,1\}^{n\times m}$ representing the cohort at m polymorphic loci. The algorithm is adjusted for binary SNPs, therefore genotypes should be converted to a binary representation according to the expected type of interaction. For example, converting AA to 0 and aA, aa to 1 states a dominant model of interactions. The phenotype vector $\mathbf{y}$ is centered so that it has zero mean and standard deviation of 1. $\mathbf{x} \in \{0,1\}^n$ denotes the column vector of allelic values measured for all n samples at a certain locus. x_i is the allele value of this locus for person number i and y_i is the phenotype value of person i. We denote $p = \Pr(x_i = 1)$, and estimate it with the maximum likelihood estimator $\hat{p} = \text{mean}(\mathbf{x})$. Denote $\mathbf{x}^2 = (x_1^2, x_2^2, \ldots, x_n^2)^T$, $\mathbf{x}\mathbf{x}' = (x_1x_1', x_2x_2', \ldots, x_nx_n')^T$ and $\mathbf{y}\mathbf{x}\mathbf{x}' = (y_1x_1x_1', y_2x_2x_2', \ldots, y_nx_nx_n')^T$

Linear model When testing for an epistatic interaction between a pair of SNPs, the linear model can be defined as follows (Cordell, 2009):

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i' + \alpha_3 x_i x_i' + \epsilon_i \tag{1a}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{1b}$$

$$H_0 : \alpha_3 = 0,\ H_1 : \alpha_3 \neq 0 \tag{1c}$$

Since tests for interaction are usually performed after testing for a main effect for each of the SNPs, it is reasonable to zero the main effects from the model. By altering the model so that $\alpha_1 = \alpha_2 = 0$ and α_0, α_3 are replaced with β_0, β_1 respectively, a new, simpler model is obtained:

$$y_i = \beta_0 + \beta_1 x_i x_i' + \epsilon_i \tag{2a}$$

$$\epsilon_i \sim N(0, \sigma^2) \tag{2b}$$

$$H_0 : \beta_1 = 0,\ H_1 : \beta_1 \neq 0 \tag{2c}$$

In this case, using ordinary least squares (OLS), the GLR tests statistic is:

$$2\ln \text{GLR} = -n \ln\Big(\frac{\sum_i (y_i - \hat{\beta}_0 - \hat{\beta_1} x_i x_i')^2}{\sum_i (y_i - \overline{\mathbf{y}})^2}\Big) \tag{3}$$

Where $\bar{\mathbf{y}} = \text{mean}(\mathbf{y})$. This simplification allows us to define an alternative test statistic, τ^2, which is an approximation of the GLR test-statistic and can be very useful for our filtering stage. Disregarding the main effect can in fact lead to false positive results, but these will only be a fraction of the total number of SNP pairs and will all be discarded during the validation stage, where pairs are tested against the full linear model (eq. 1).

To achieve simplicity and efficiency, the model does not include covariates - the residuals from the phenotype adjusted for other parameters should be used as the response variable. Population stratification can be addressed by applying an adjustment method such as EIGENSTRAT (Price *et al.*, 2006) and using the first axes of variation as covariates while adjusting the phenotype. The model assumes linkage equilibrium between SNPs, an assumption that does not hold for GWAS, where proximal SNPs are in linkage disequilibrium (LD). As a result, the distribution of the τ^2 score for proximal SNPs deviates from what is expected under the null assumption, which results in an excess of pairs passing the filtering stage. This problem can be addressed by dismissing proximal pairs during the filtering stage, and exhaustively testing them later during post-processing time. As the number of proximal pairs in LD is $O(m)$, the cost of this correction is minor.

Generalized likelihood ratio test and the new test-statistic τ^2

In the following section we introduce our new test statistic τ^2 and show that for large sample sizes, $2\ln$ GLR $\approx \tau^2$. The new tests statistic is presented

not as a means to achieve more power, rather as a means for reducing runtime by serving as a proxy to the GLR test statistic: We show in the next section how random projections methods can efficiently detect pairs with a high τ^2 score, as a filtering stage for detecting statistically significant interactions.

Since $\mathbf{y}$ is standardized, the denominator of equation 3 equals n. Replacing $\hat{\beta}_0, \hat{\beta}_1$ in equation 3 with their OLS estimators $\overline{\mathbf{y}} - \hat{\beta}_1\overline{\mathbf{xx'}}$, $\frac{\overline{\mathbf{yxx'}}}{\overline{\mathbf{x^2x'^2}} - \overline{\mathbf{xx'}}^2}$ respectively, it is easy to verify that:

$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i x_i')^2 = \sum_i (y_i - \frac{\overline{\mathbf{yxx'}}}{\overline{\mathbf{xx'}} - \overline{\mathbf{xx'}}^2}(x_i x_i' - \overline{\mathbf{xx'}}))^2 \tag{4a}$$

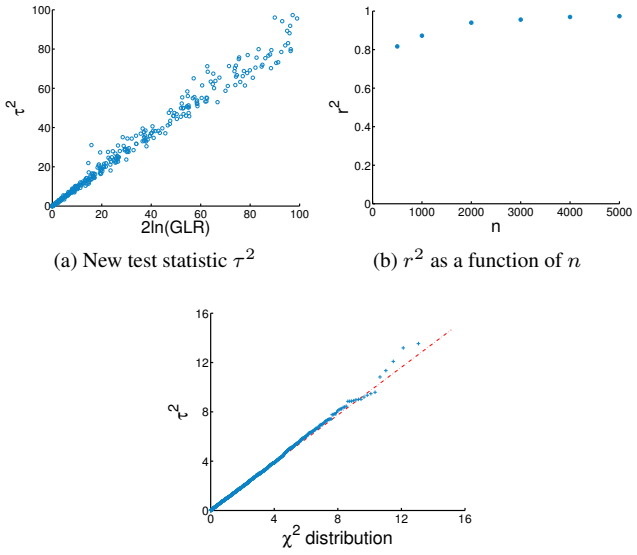
$$= n(1 - \frac{\overline{\mathbf{yxx'}}^2}{\overline{\mathbf{xx'}} - \overline{\mathbf{xx'}}^2}) \tag{4b}$$

$$2\ln \text{GLR} = -n \ln\left(1 - \frac{\overline{\mathbf{yxx'}}^2}{\overline{\mathbf{xx'}} - \overline{\mathbf{xx'}}^2}\right) \tag{5}$$

Under the linkage equilibrium assumption, $\overline{\mathbf{xx'}} \xrightarrow{p} pp'$. Using first order Taylor expansion, after neglecting $\overline{\mathbf{xx'}}^2$, we conclude that for a large sample size:

$$\tau^2 \equiv n\frac{\overline{\mathbf{yxx'}}^2}{\hat{p}\hat{p}'} \approx 2\ln \text{GLR} \tag{6}$$

Figure 1a displays $2\ln$ GLR vs. τ^2. With r^2 of 0.99, τ^2 is a good approximation of the GLR score. As seen on Figure 1c, under the null assumption of no interaction, the distribution of τ^2 is very close to a chi-square distribution with 1 degree of freedom, similar to the GLR test statistic. (chi-square goodness-of-fit test p-value = 0.396). As a result, the task of finding an interacting SNP pair can now be replaced with the task of finding a pair with significantly high τ^2 score. We show on the next section that this task can be done efficiently without testing all pairs.



(a) New test statistic τ^2 (b) r^2 as a function of n

(c) QQ Plot of τ^2 vs χ^2 distribution

Fig. 1: **The new test-statistic**: (a) $2\ln$ GLR vs. τ^2. Data was generated with $n = 5000$, MAF $\in [0.01, 0.5]$; marginal and epistatic effects were sampled uniformly from the range $(0, 1)$; $r^2 = 0.99$. (b) r^2 of the linear correlation between $2\ln(\text{GLR})$ and τ^2, as a function of n: τ^2 is highly correlated with the original test statistic for all tested sample sizes. (c) τ^2 distribution is proportional to the chi-square distribution with 1 degree of freedom. Passed a chi-square goodness-of-fit test with p-value of 0.396.

Efficient discovery of interacting SNPs

We describe an algorithm for finding pairs of SNP where τ^2 is larger than a given threshold. For each binary SNP $\mathbf{x}$ we define a vector $\mathbf{v} = (v_1, \ldots, v_n)$ where $v_i = \sqrt{\frac{|y_i|}{\hat{p}\sqrt{n}}}x_i$ and a vector $\mathbf{u} = (u_1, \ldots, u_n)$ where $u_i = \text{sign}(y_i)v_i$. For example, if $\mathbf{y} = (-0.1, 0.2, -0.3, -0.4, -0.5, 0.6)$ and $\mathbf{x} = (1, 1, 0, 1, 0, 0)$ then $\mathbf{v} = \frac{1}{\sqrt{0.5\sqrt{6}}}(\sqrt{0.1}, \sqrt{0.2}, 0, \sqrt{0.4}, 0, 0)$ and $\mathbf{u} = \frac{1}{\sqrt{0.5\sqrt{6}}}(-\sqrt{0.1}, \sqrt{0.2}, 0, -\sqrt{0.4}, 0, 0)$. It is easy to see that

$$\forall \mathbf{x}, \mathbf{x}' : \mathbf{v} \cdot \mathbf{u}' = \frac{1}{\sqrt{n\hat{p}\hat{p}'}} \sum_{i=1}^{n} y_i x_i x_i' = \tau \tag{7}$$

So instead of searching for pairs with an exceptional τ score, we are now looking for an exceptional inner product size. To do so we apply a random projections method: we perform multiple iterations; in each iteration we sample a random vector $\mathbf{r} = (r_1, r_2, ..., r_n)$, where $r_i \sim N(0, 1)$. For each SNP $\mathbf{x}$ we calculate two scores: $a = \mathbf{v} \cdot \mathbf{r}$ and $b = \mathbf{u} \cdot \mathbf{r}$. Since r_i are sampled i.i.d with mean 0 and variance 1, the expected value of the two scores' product is τ:

$$\forall \mathbf{x}, \mathbf{x}' : \text{E}_r[ab'] = \text{E}_r[\sum_i r_i v_i \sum_j r_j u_j'] = \mathbf{v} \cdot \mathbf{u}' = \tau \tag{8}$$

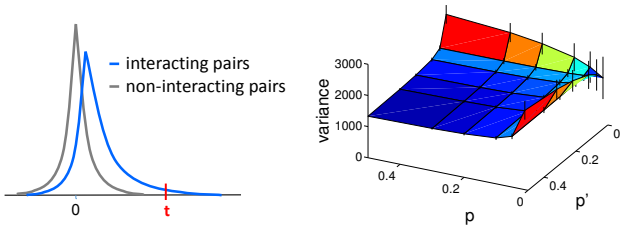
Note that while non-interacting SNPs have a zero expected value for ab', pairs with a significant p-value after a Bonferroni correction of $\binom{10^6}{2}$ are expected to have τ^2 of over 55. It can also be shown that $\text{Var}[ab'] = \tau^2 + \|\mathbf{v}\|^2\|\mathbf{u}'\|^2$. As a result, the distribution of ab' has a longer tail under the alternative assumption, so for any positive threshold t, the probability of $|ab'| \geq t$ is always greater for interacting pairs (see Figure 2a). We utilize this fact to distinguish between interacting and non-interacting pairs: we perform several iterations where a vector $\mathbf{r}$ is sampled, and the scores a and b are calculated for all SNPs. In each iteration we collect the pairs of SNPs whose scores product pass a given threshold t. The last part can easily be done without testing all pairs: We define a vector $\vec{\mathbf{a}} = (a_1^2 \ldots a_m^2)$ and a vector $\vec{\mathbf{b}} = (b_1^2 \ldots b_m^2)$. Both vectors are first sorted in descending order and then scanned in linear time to find pairs $\mathbf{x}, \mathbf{x}'$ such that $a^2 b'^2 > t^2$.

Since, as seen on Figure 2b, the variance of ab' is affected by the combination of minor allele frequencies (MAFs) of both SNPs, different t thresholds are used for different minor allele frequency combinations. This is done by assigning SNPs to bins of similar MAF: Each bin B has two score vectors, $\vec{\mathbf{a}}^B, \vec{\mathbf{b}}^B$, sorted by their score value. For each pair of bins, B and B', we report all SNP pairs $\mathbf{x} \in B, \mathbf{x}' \in B'$ where $a^2 b'^2 \geq t^2_{BB'} \wedge a'^2 b^2 \geq t^2_{BB'}$, when $t^2_{BB'}$ is the appropriate threshold. Reported SNP pairs are validated against the linear model. Optimal t thresholds for each pair of bins were empirically calculated, as described in the following section. See algorithm pseudo-code 1.

Runtime analysis The improvement in runtime achieved by EPIQ is due to the fact that only a fraction of the SNP pairs is tested. The algorithm performs L iterations, each iteration has $\text{O}(nm)$ operations for calculating a, b scores and $\text{O}(m \log m)$ operations for sorting score vectors. If we denote ψ as the average fraction of SNP pairs that pass the threshold t at each iteration, then scanning the vectors for interaction candidates would take $\text{O}(\binom{m}{2}\psi)$ and the total runtime including validations is $\text{O}\big(L(nm + mlog(m) + \binom{m}{2}\psi n)\big)$. As exhaustive testing af all pairs take $\text{O}\big(\binom{m}{2}n\big)$, speedup is achieved when $L << \psi$ and also $L << m$. To speedup performance, EPIQ keeps all SNP data in memory, therefore space complexity is $\text{O}(nm)$.

for $l = 1$ *to* L **do**
  sample a vector $\mathbf{r} \in \mathbf{R}^n$ s.t. $r_i \sim N(0,1)$;
  foreach *SNP* $\mathbf{x}$ **do**
    $B :=$ bin number of $\mathbf{x}$;
    append $a^2 := (\mathbf{v} \cdot \mathbf{r})^2$ to $\vec{\mathbf{a}}^B$;
    append $b^2 := (\mathbf{u} \cdot \mathbf{r})^2$ to $\vec{\mathbf{b}}^B$;
  end
  foreach *bin* B **do**
    sort vector $\vec{\mathbf{a}}^B$;
    sort vector $\vec{\mathbf{b}}^B$;
  end
  foreach *pair of bins* B, B' **do**
    scan vectors $\vec{\mathbf{a}}^B, \vec{\mathbf{b}}^{B'}$ and validate all pairs $\mathbf{x}, \mathbf{x}'$ where $\min(a^2 b'^2, a'^2 b^2) \geq t^2_{BB'}$;
    report pairs that pass validation as interacting;
  end
end

Algorithm 1: EPIQ pseudo-code



(a) ab' score distribution (b) variance of ab' vs MAF

Fig. 2: (a) An illustration of the ab' score distribution for interacting pairs (blue) and non-interacting pairs (gray): interacting pairs have a higher probability of passing a threshold t during the filtering stage. (b) Variance of ab' as a function of minor allele frequencies, for an interacting pair with corrected p-value of 0.05.

Choosing the parameters L, t to assure the requested power with minimal runtime Given the stochastic nature of the algorithm, there is always a possibility that interacting pairs will be missed. This creates a trade-off between power and runtime, controlled by a *success rate* parameter. Setting this parameter to 90%, for example, would mean that the probability of missing a SNP pair with a significant GLR score is at most 10%, and the overall power achieved is at least 90% compared to an all-pairs scan. Strongly interacting pairs have an even larger chance of being detected, as the probability of passing the filtering stage is a function of the GLR score.

According to the success rate requested by the user, optimal values for L and t can be set. The two parameters are strongly linked with runtime: Higher t values reduce probability of success, which means more iterations are required in order to provide the requested success rate. This elongates the filtering stage, but also might shorten the validation stage by reducing false positive rate. To calculate optimal parameters one must first calculate $f \equiv \Pr[\min(a^2 b'^2, a'^2 b^2) \geq t^2]$ for both interacting and non-interacting pairs. One can show that $(a, b') \sim N_2\left([0,0], \begin{bmatrix} \|\mathbf{v}\|^2 & \tau \\ \tau & \|\mathbf{u}'\|^2 \end{bmatrix}\right)$, so the probability of the event $a^2 b'^2 \geq t^2$ can be easily calculated. The value of f, on the other hand, is not as simple to calculate analytically. As a result, the choice of the parameters was done empirically: A sample dataset was randomly generated, using minor allele frequencies taken from the 1000 genomes project (Abecasis *et al.*, 2012), as explained in the results section. SNPs were distributed among bins of similar MAFs, and each pair of bins was assigned with the maximal threshold value that enabled the required success rate, given the current number of iterations. As a final step, the number of iterations that led to the shortest runtime was chosen.

Simulated datasets

In order to test our algorithm we generated several datasets of diploid genotypes, with cohort sizes varying between 1000 and 5000, and the number of SNP loci between $10,000$ and 1 million. While generating the SNPs we used the minor allele frequency distribution found on the 1000 genomes project (Abecasis *et al.*, 2012) and assumed Hardy-Weinberg equilibrium. We later converted the datasets to a binary representation using a dominant coding, where AA was translated to 0 and aA, aa to 1. We used these datasets to demonstrate the runtime and power of EPIQ under different conditions.

The Ludwigshafen Risk and Cardiovascular Health (LURIC) study

We applied our method to measurements of lipid concentration in cells (Cer(d18:0/24:1)), taken from the LURIC study. The LURIC study consists of 3,316 white patients hospitalized for coronary angiography between 1997 and 2000 at a tertiary care center in Southwestern Germany (Winkelmann *et al.*, 2001). To limit clinical heterogeneity, individuals suffering from acute illnesses other than acute coronary syndrome (ACS), chronic non-cardiac diseases and a history of malignancy within the five past years were excluded.

Laboratory Procedures Fasting blood samples were obtained by venipuncture in the early morning. Genomic DNA was prepared from EDTA anticoagulated peripheral blood by using a common salting-out procedure. Genotyping was done using the Affymetrix Human SNP Array 6.0 at the Synlab Center of Laboratory Diagnostics Heidelberg and the Mannheim Institute of Public Health of Heidelberg University.

Quality control We used PLINK (Purcell *et al.*, 2007) for quality control, excluding SNPs with call rate $< 95\%$. We excluded individuals with call rate $< 97\%$, ambiguous on genetic sex test or showing high estimated identity by decent (IBD) scores (PI_HAT ≥ 0.1875), controlling for cryptic relatedness. For the population stratification part we used the POPRES dataset (Nelson *et al.*, 2008) as a reference population. We considered the first four components of a multidimensional scaling (MDS) on both LURIC and POPRES individuals for determining and removing outliers. Finally, we had 687,253 SNPs and 859 individuals remaining for the analysis, of which 826 had lipid cell concentration measurements.

3 RESULTS

In this section we show that in just a few hours EPIQ can process amounts of data that would take weeks and even years on common existing software. We demonstrate how the power of EPIQ is affected by the underlying model of interaction and present the results of applying EPIQ to a dataset from the Ludwigshafen Risk and Cardiovascular Health (LURIC) study.

Runtime improvement

While epistasis detection tools for case-control studies are relatively common, not many quantitative pairwise-epistasis tools were found. We chose to compare EPIQ against PLINK (Purcell *et al.*, 2007), FastEpistasis (Schüpbach *et al.*, 2010), EpiGPU (Hemani *et al.*, 2011) and EpiGPUHSIC (Kam-Thong *et al.*, 2011). All four tools perform an exhaustive search, using different hardware and various statistical tests. PLINK is a commonly used

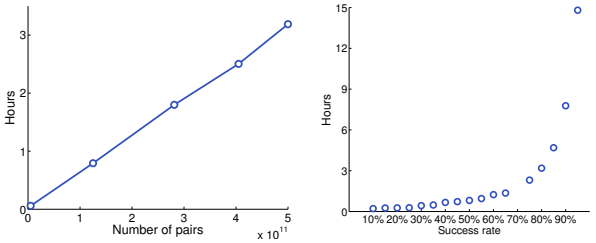
Table 1. Runtime of the C++ implementation of EPIQ, compared to other programs available. EPIQ was run with the parameter *success rate* set to 80%, therefore runtime is compared against testing 80% of the pairs in the exhaustive search algorithms ($n = 1000, m = 10^6$).

Tool	Computational method	Statistical test	Cores	Runtime
PLINK (Purcell *et al.*, 2007)[a]	Exhaustive search	OLS	1	$\sim$ 10 years
FastEpistasis (Schüpbach *et al.*, 2010)[b]	Exhaustive search	OLS	8	381 hours
EpiGPU (Hemani *et al.*, 2011)[b]	Exhaustive search	F-test	-	9.3-90 hours[c]
EpiGPUHSIC (Hemani *et al.*, 2011)[b]	Exhaustive search	HSIC	-	194 hours
EPIQ (Kam-Thong *et al.*, 2011)[b]	Random projections	OLS on binary SNPs	8	3.2 hours

[a] Times were extrapolated according to a test of 1000 SNPs performed on the same 2.5 GHz processor, scaling linearly with the number of SNP pairs.
[b] Times were extrapolated according to self-reported performance.
[c] Runtime varies with the chosen GPU.

whole-genome association analysis toolset. Its *epistasis* option performs linear regression tests on all SNP pairs. FastEpistasis is an efficient parallel extension of the PLINK epistasis module. While the first two tools run on regular processors, EpiGPU and EpiGPUHSIC run on graphical processing units (GPU), which are specialized electronic circuits that provide up to $\times 100$ speedup in performance. The former two tools utilize different statistical tests as well: EpiGPU performs an F-test, while EpiGPUHSIC is a quantitative extension of HSIC (Gretton *et al.*, 2005), which uses the correlation coefficient difference between cases and controls, as an approximation to the significance of the interaction term. Since EPIQ was run with the parameter *success rate* set to 80%, we compared its runtime against testing 80% of the pairs in the exhaustive search algorithms. As seen on table 1, EPIQ shows a great improvement in runtime, compared to the exhaustive tools.

We ran EPIQ using different inputs in order to test how the program scales with changes in the number of SNPs, cohort size or requested power. As seen on Figure 3a, EPIQ scales linearly with the number of SNP pairs in the dataset. Figure 3b shows that gaining more power becomes increasingly time-consuming when approaching 100% power, as can be expected in stochastic algorithms of this sort. However, one can achieve almost 100% power in a matter of hours. Scaling in the number of samples is above linear as well: while testing 5×10^{11} pairs takes 3.2 hours for 1000 individuals, it takes 7 times longer for 3000 samples, and 30 times longer for 5000 samples. Nevertheless, for moderate cohort sizes EPIQ remains an efficient choice. (All benchmark tests were performed on a Ubuntu Linux server with 2.5 GHz processor.)



(a) Runtime vs number of SNP pairs (b) Runtime vs power

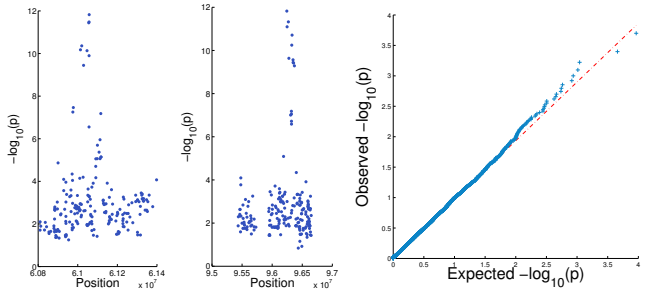
Fig. 3: **Runtime of EPIQ for different settings**: (a) Runtime for various numbers of SNP pairs, $n = 1000$; EPIQ scales linearly with the number of pairs. (b) Runtime of EPIQ for different power thresholds; Nearly 100% power can be achieved in a matter of hours ($n = 1000, m = 10^6$).

Power analysis

To evaluate the power of our algorithm, we compared it against two commonly used baseline methods suggested by Marchini *et al.* (2005). The first is a simple exhaustive all-pairs test, where all SNP pairs are tested for interaction. Although this method is not feasible for large datasets, the power achieved by an all-pairs test is of relevance, as this is the upper bound for the power of our algorithm. The second baseline we compared against is a method in which the top K marginal predictors are identified, and then tested for all pairwise iterations between them. When choosing $K = \sqrt{2m}$, for example, the number of tests performed is $\binom{\sqrt{2m}}{2} \approx m$. We refer to this method as the 'two-step' algorithm. In all our tests we apply the conservative Bonferroni correction, in order to address the issue of multiple hypothesis. Since EPIQ implicitly evaluates all SNP pairs, the number of tests for a multiple testing correction is $\binom{m}{2}$, as in the all-pairs algorithm. For each test the program generated a quantitative phenotype according to the linear model described earlier, $y_i = \beta_0 + \beta_1 x_i x_i' + \epsilon_i$, where $\mathbf{x}, \mathbf{x}'$ are two SNPs that were randomly chosen as the interacting pair. The phenotype was later standardized so that $\bar{\mathbf{y}} = 0,\ \text{stdev}(\mathbf{y}) = 1$.

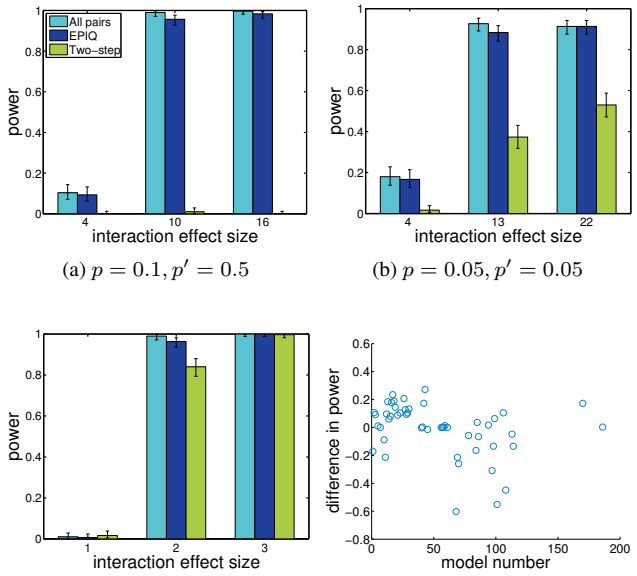
We compared EPIQ against the two methods, using different MAFs for the interacting SNPs and *success rate*=80% (Figures 5a-5c). Note that although the requested success rate was 80%, the actual power of EPIQ (shown in dark blue) is consistently more than 80% of the power achieved by the all-pairs algorithm (light blue), as this parameter states the minimal relative power. Another conclusion drawn from these figures is that in some cases there is a substantial difference in power between the all-pairs test and the two-stage test (green), in favor of the all-pairs test. The opposite is true for large MAFs, as in this case the marginal effect is easy to detect, and the multiple testing correction is less stringent for a two-stage approach (not shown). Similar results where described by Evans *et al.* (2006), which showed that for various models of interaction, an exhaustive all-pairs search is more powerful compared to the two-step strategy, despite the harsher multiple testing correction ($O(m^2)$ compared to $O(m)$). In these cases, using EPIQ can yield a substantial improvement in power.

Comparison with PLINK In order to further investigate the power achieved by EPIQ, we carried 50 experiments comparing our method to the linear regression performed by PLINK and FastEpistasis, using the 50 distinct models of interaction from Li and Reich (2000), which were adapted for quantitative traits. These models assume that there are two phenotypic means in the population: 0 and 1, and each model of interaction determines a different partitioning of the population to either mean. For example, model M1 states that only the individuals that are homozygous with the minor allele



(a) chr5 (b) chr6 (c) QQplot of p-values distribution

Fig. 4: **Results on the LURIC dataset**: (a)+(b) Manhattan plots of 100 SNPs up and down-stream of rs436969, rs9385393. The --epistasis option of PLINK was used to test for interactions in all 40,401 pairs and the smallest p-value for each SNP was recorded. Note that the p-value for the top scoring pair is slightly higher than the one calculated by EPIQ, as EPIQ was run on the binary representation of the SNPs. (c) A QQ-plot of the p-values distribution shows a negligible inflation. p-values were calculated for a sample of 10,000 SNP pairs.



(a) $p = 0.1, p' = 0.5$ (b) $p = 0.05, p' = 0.05$

(c) $p = 0.2, p' = 0.2$ (d) Compared to PLINK

Fig. 5: **Power of the different algorithms**: (a)-(c) EPIQ, all-pairs search and the two-stage search, to discover the true interacting SNP pair. p and p' are the minor allele frequencies of the interacting pair in each test. Under all settings, the relative power of EPIQ compared to the exhaustive search exceeds the requested success rate of 80%. (d) The power of EPIQ compared to the full linear model PLINK: Each dot represent a distinct model of interaction from Li and Reich (2000), the x axis is the model number, the y axis is the average power of EPIQ minus the power of PLINK.

on both SNPs have the higher phenotypic mean (see Li and Reich (2000) for more details). We generated 2000 samples, where the two SNPs account for 10% of the trait variance. Before running EPIQ, we converted the genotypes to two binary representations, a dominant one and a recessive one, and applied EPIQ to both encodings (as in Brinza *et al.* (2010) and Prabhu and Pe'er (2012)). We compared EPIQ's results with the power achieved by applying the full linear model of PLINK on the original genotypes. Figure 5d shows the results for all models, when the x axis is the model number and the z axis is the power of EPIQ minus the power of PLINK, averaged over all MAF combinations. 23 of the 50 models showed greater power when using EPIQ, 15 showed greater power with PLINK, and the remaining 12 result in a similar power when using either method. Several of the 15 models where PLINK shows higher power describe either a complex and biologically unintuitive pattern of interaction (such as M101), or have a large marginal effect, which makes them easy to discover using Marchini's two stage algorithm Marchini *et al.* (2005).

Results from the LURIC study

We applied EPIQ to measurements of lipid concentration in cells (Cer(d18:0/24:1)), taken from the LURIC study, setting the *success rate* parameter to 90%. Lipid concentration in cells was converted to the log scale, standardized and corrected for BMI, sex, age and statins usage, using the residuals as the input for EPIQ. Processing of 826 individuals and 687,253 SNPs took 1.5 hours on 10 processors, identifying a single pair of SNPs (rs436969 (chr5, HWE p=0.005), rs9385393 (chr6, HWE p=1)) with a p-value of 2.2×10^{-13}, which is near-significant after applying a Bonferroni correction. No genes exist within 100KBp up and down-stream of the SNPs. Figure 4 shows a Manhattan plot of the SNPs surrounding the pair of SNPs.

4 DISCUSSION

In this paper we demonstrated how random projections methods can be applied on quantitative GWAS, achieving in most cases at least an order of magnitude speedup compared to other existing tools, scaling linearly with the number of SNP pairs. We showed that EPIQ required only 1.5 hours on 10 processors for a real dataset of 687,253 SNPs and 826 individuals, identifying a pair of SNPs with a possible epistatic interaction, demonstrating that the model's assumptions do not hinder an efficient discovery of interacting pairs. This speedup is gained in exchange of a minor loss of power. As mentioned before, a search for interacting SNP pairs in current GWAS suffers from an inherent multiple testing problem, where p-values must be as small as 10^{-13} or less in order to be considered significant. Like RAPID (Brinza *et al.*, 2010) and SIXPAC (Prabhu and Pe'er, 2012) have done, EPIQ turns this limitation into an advantage: For a given sample size, smaller p-values are a result of larger effect sizes. This in turn makes interacting pairs more distinct from the rest of the SNPs and consequently easier to detect by EPIQ. The drawback is that as sample size increases, a smaller effect size is required for achieving the same significance level. In this case EPIQ is required to perform more iterations in order to distinguish between interacting and non-interacting pairs and therefore does not scale linearly with sample size. Thus, as GWAS expand to include increasingly larger cohorts, further adjustments in the algorithm would be required.

We wish to state that the settings used in this article are only a portion of a wide range of options. The approach we described can be extended to fit other statistical tests, and the binary coding of the genotypes can be performed differently than described, to match other underlying models of interaction. EPIQ can also be used in

conjunction with methods such as GRAT (Kostem and Eskin, 2013), which utilize LD between SNPs for choosing a subset of proxy SNPs, thus reducing the number of tests and further improving runtime. Moreover, the runtime reported relates to the current code implementation of the algorithm; Different implementations, such as GPU based code, are likely to achieve even better results. With the decrease in runtime, permutation tests for significance become a feasible option, resulting in increased power compared to stringent methods for multiple hypothesis correction.

ACKNOWLEDGEMENT

The collections and methods for the Population Reference Sample (POPRES) are described by Nelson *et al.* (2008). The datasets used for the analyses described in this manuscript were obtained from dbGaP at `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v4.p2` through dbGaP accession number phs000145.v4.p2.

Funding: This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. E.H is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. E.H and Y.A were supported by the Israel Science Foundation grant no. 1425/13. E.H. was also partially supported by National Science Foundation grant III-1217615. The genotype dataset was obtained from the LURIC study. The LURIC study was supported by the 6th Framework Program (integrated project Bloodomics, grant LSHM-CT-2004-503485), by the 7th Framework Program (integrated project AtheroRemo, grant agreement number 201668 and RiskyCAD, grant agreement number 305739) of the European Union and by the INTERREG IV Oberrhein Program (Project A28, Genetic mechanisms of cardiovascular diseases) with support from the European Regional Development Fund (ERDF) and the Wissenschaftsoffensive TMO.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**(7422), 56–65.

Bhattacharya, K., McCarthy, M. I., and Morris, A. P. (2011). Rapid testing of gene-gene interactions in genome-wide association studies of binary and quantitative phenotypes. *Genetic epidemiology*, **35**(8), 800–8.

Brinza, D., Schultz, M., Tesler, G., and Bafna, V. (2010). RAPID detection of gene-gene interactions in genome-wide association studies. *Bioinformatics (Oxford, England)*, **26**(22), 2856–62.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature reviews. Genetics*, **10**(6), 392–404.

Evans, D. M., Marchini, J., Morris, A. P., and Cardon, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLoS genetics*, **2**(9), e157.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer.

Hemani, G., Theocharidis, A., Wei, W., and Haley, C. (2011). EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics (Oxford, England)*, **27**(11), 1462–5.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(23), 9362–7.

Hu, X., Liu, Q., Zhang, Z., Li, Z., Wang, S., He, L., and Shi, Y. (2010). SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell research*, **20**(7), 854–7.

Kam-Thong, T., Pütz, B., Karbalai, N., Müller-Myhsok, B., and Borgwardt, K. (2011). Epistasis detection on quantitative phenotypes by exhaustive enumeration using GPUs. *Bioinformatics (Oxford, England)*, **27**(13), i214–21.

Kostem, E. and Eskin, E. (2013). Efficiently identifying significant associations in genome-wide association studies. *Research in Computational Molecular Biology*, pages 118–131.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human heredity*, **50**(6), 334–49.

Liu, Y., Xu, H., Chen, S., Chen, X., Zhang, Z., Zhu, Z., Qin, X., Hu, L., Zhu, J., Zhao, G.-P., and Kong, X. (2011). Genome-wide interaction-based association analysis identified multiple new susceptibility Loci for common diseases. *PLoS genetics*, **7**(3), e1001338.

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, **456**(7218), 18–21.

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature genetics*, **37**(4), 413–7.

Nelson, M. R., Bryc, K., King, K. S., Indap, A., Boyko, A. R., Novembre, J., Briley, L. P., Maruyama, Y., Waterworth, D. M., Waeber, G., Vollenweider, P., Oksenberg, J. R., Hauser, S. L., Stirnadel, H. A., Kooner, J. S., Chambers, J. C., Jones, B., Mooser, V., Bustamante, C. D., Roses, A. D., Burns, D. K., Ehm, M. G., and Lai, E. H. (2008). The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research. *American journal of human genetics*, **83**(3), 347–58.

Prabhu, S. and Pe'er, I. (2012). Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome research*, **22**(11), 2230–40.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, **38**(8), 904–9.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, **81**(3), 559–75.

Schüpbach, T., Xenarios, I., Bergmann, S., and Kapur, K. (2010). FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics (Oxford, England)*, **26**(11), 1468–9.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., and Yu, W. (2010). BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies. *The American Journal of Human Genetics*, **87**(3), 325–340.

Winkelmann, B. R., März, W., Boehm, B. O., Zotz, R., Hager, J., Hellstern, P., and Senges, J. (2001). Rationale and design of the LURIC study–a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease. *Pharmacogenomics*, **2**(1 Suppl 1), S1–73.

Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). GBOOST: a GPU-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics (Oxford, England)*, **27**(9), 1309–10.

Zhang, X., Huang, S., Zou, F., and Wang, W. (2010). TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics (Oxford, England)*, **26**(12), i217–27.