Published in final edited form as:

Mol Biosyst. 2012 April ; 8(5): 1544–1552. doi:10.1039/c2mb05333k.

Symmetric Epistasis Estimation (SEE) and its application to dissecting interaction map of *Plasmodium falciparum*

Yang Huang¹, Geoffrey Siwo², Stefan Wuchty¹, Michael T. Ferdig², and Teresa M. Przytycka^{1,*}

¹National Center for Biotechnology Information, NLM, NIH, 8600 Rockville Pike, Building 38A, Bethesda, MD 20894, USA

²Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, 100 Galvin Life Sciences, Notre Dame, IN 46556, USA

Abstract

It is being increasingly recognized that many important phenotypic traits, including various diseases, are governed by a combination of weak genetic effects and their interactions. While the detection of epistatic interactions that involve a non-additive effect of two loci on a quantitative trait is particularly challenging this interaction type is fundamental for the understanding of genome organization and gene regulation. However, current methods that detect epistatic interactions typically rely on the existence of a strong primary effect, considerably limiting the sensitivity of the search. To fill this gap, we developed a new method, SEE (Symmetric Epistasis Estimation), allowing the genome-wide detection of epistatic interactions without the need for a strong primary effect. We applied our approach to progeny crosses of the human malaria parasite *P. falciparum* and *S. cerevisiae*. We found an abundance of epistatic interactions in the parasite and a much smaller number of such interactions in yeast. The genome of *P. falciparum* also harboured several epistatic interaction hotspots that putatively play a role in drug resistance mechanisms. The abundance of observed epistatic interactions might suggest a mechanism of compensation for the extremely limited repertoire of transcription factors. Interestingly, epistatic interactions hotspots were associated with elevated levels of linkage disequilibrium, an observation that suggests selection pressure acting on *P. falciparum*, potentially reflecting host-pathogen interactions or drug-induced selection.

Introduction

Complex traits, including human diseases, often involve epistatic interactions [1,2], a property defined as a non-additive/non-independent effect of two loci on a quantitative trait [3,4,5]. The detection of interactions multi-locus linkage in general and epistatic interaction in particular is a very challenging task since the large number of traits causes a massive multiple testing issue and demands considerable computational resources. Therefore, exhaustive search methods for detecting loci interactions are usually cumbersome [6]. To deal with such constraints, current methods often pre-filter locus pairs and zoom in on a candidate set of interactions that are expected to be enriched with such interacting locus pairs. For example, Storey *et al.* proposed a step-wise strategy for multiple loci linkage analysis [7], requiring that one locus had a significant, marginal effect on the underlying trait. In a recent study, Hannum *et al.* designed a bi-clustering procedure for the

*corresponding author: przytyck@mail.nih.gov, phone 301-402-1723, fax 301-480-4637 .

Additional Files Additional File 1: Supplementary online materials (including Fig. S1 and Table S2, S3).

Table S1: List of 1,713 epistatic interactions in *P. falciparum* detected by SEE.

identification of interacting locus pairs and applied their approach to uncover functional links between protein complexes in yeast [8]. Litvin *et al.* developed a three-stage method, which additionally utilized gene expression clustering to detect interactions in yeast [9]. While these methods differ in the type of detected interactions they commonly relied on the step-wise identification of primary and secondary loci. However, interactions of gene products do not necessarily depend on an outstanding influence of a single factor. In fact, transcriptional regulation of most genes often involves multiple regulators that have small individual effects [10]. Furthermore, such interactions are difficult to detect by the stepwise method unless the set of progenies is extremely large, prompting the need for approaches that are independent of strong primary locus effects.

Generally, methods to detect epistatic interactions feature two main characteristics: (i) To mitigate a potential multiple testing issue methods limit the number of upcoming statistical tests, for example by identifying strong primary loci [7] or leveraging modularity [8,11]. (ii) As a second key characteristic approaches employ different statistical models to detect certain types of interactions. For example, Hannum *et al.* [8] detected pairs of interacting loci without focusing on epistasis. Storey *et al.* [7] and Zhang *et al.* [11] first uncovered loci that jointly controlled a trait without demanding that their impact was non-additive/non-independent [9]. Subsequently, epistatic locus pairs were identified by testing the significance of the interaction term.

Complementing existing methods, we designed a novel method, SEE (Symmetric Epistasis Estimation), allowing us to detect epistatic interactions in relatively small sets of progenies without the identification of primary loci. Our approach relied on a graph-based filtering step specifically designed to retain a large fraction of symmetric, candidate epistatic interactions. Compared to previous methods we also adopted the classical, more rigid, definition of epistasis, allowing us to rigorously check the significance of candidate interactions.

We applied our approach to uncover epistatic interaction in the malaria parasite *P. falciparum*. In this organism, genetic crosses have yielded few recombinant clones, contributing to statistical challenges. While the first eQTL study of *P. falciparum* by Gonzales *et al.* [12] suggested a potential role of epistatic regulation of gene expression [13], a broader understanding of transcriptional regulatory mechanisms in *P. falciparum* remained unclear. Applying our approach to a HB3 $\times$ Dd2 parasite cross [12] we found more than 1,500 putative epistatic interactions between locus pairs on different chromosomes and identified several epistatic interaction hotspots of biological significance. Interestingly, we found that the level of linkage disequilibrium (LD) between locus pairs was correlated with the number of genes whose expression was influenced by the corresponding epistatic interaction. Such disequilibrium provides an additional level of interactions between the loci.

We also applied our method to an eQTL dataset of *S. cerevisiae* [14]. Surprisingly, we found much fewer epistatic interactions and no epistatic interaction hotspots. After ruling out that our results were statistical artefacts caused by the small number of *P. falciparum* progenies, we hypothesized that selection pressure acting on *P. falciparum* contributed to observed epistatic interactions and elevated LD, potentially reflecting host-pathogen interactions or drug induced selection.

Results

New method for uncovering epistatic interactions

Adopting the classical perspective on epistasis [5] we defined that two loci I' and I'' have an epistatic interaction effect on gene g , if the expression y of gene g is significantly better explained by a widely used synergistic/epistatic interaction model:

$$y = b_0 + b_1 x' + b_2 x'' + b_3 x' x'' + \varepsilon \quad (\text{Eq.1})$$

than the additive/independent model:

$$y = b_0 + b_1 x' + b_2 x'' + \varepsilon, \quad (\text{Eq.2})$$

where x' and x'' are the genotypes of loci I' and I'' , and ε is a noise term. We designed an efficient algorithm, SEE (Symmetric Epistasis Estimation), which allows us to uncover epistatic interactions without enumerating all possible combinations of locus pairs and genes or relying on a strong primary locus (Fig. 1a). In a filtering step, we assumed that the genotype of each locus was either 0 or 1, representing the corresponding parent. Since each differentially expressed gene was either represented as up or down regulated we obtained 16 possible expression phenotype-genotype configurations (Figure 1b). Specifically, we identified eight patterns that suggested a synergistic, (*i.e.* epistatic) effect, referred to as interacting configurations. Subsequently, we identified triplets, containing one gene and two loci, which were compatible with interacting configurations using a graph theoretical algorithm (see Methods for details). Using the formal definition of epistasis we separately fitted the interaction (eq.1) and additive (eq.2) models to the data. Subsequently, we applied a statistical test to check if the interaction model provided a significantly better fit than the additive model (see Methods for details). Finally, we corrected obtained p -values with an empirical null distribution, using the q -value approach [16] to select triplets with a given FDR threshold. Additionally, we applied our statistical framework to randomly sampled triplets. Comparing to such a random set, we confirmed that SEE detected candidate triplets that were significantly enriched with epistatic interactions (Fig. S1 in Additional File 1). In the algorithm, we introduced a graph based measure of support for each candidate triplet. Increasing the support threshold resulted in a smaller number of predicted interactions at lower FDR.

General patterns of epistatic interactions in *P. falciparum*

To avoid biases related to linkage disequilibrium (LD) in *P. falciparum*, we only considered epistatic interactions between two loci from different chromosomes. In the eQTL data set, encompassing 5,150 genes and 329 marker loci in 34 progeny strains of *P. falciparum*, we identified 3,796 triplets with a corrected p -value threshold < 0.001 (FDR < 0.003). Since linkage disequilibrium in the *P. falciparum* cross was high, gene expression traits were often mapped to multiple locus pairs such as (l'_1, l''_1) and (l'_2, l''_2) , where l'_1, l'_2 and l''_1, l''_2 were close on different chromosomes. Therefore, we corrected p -values in an additional way that accounted for the effects of LD (see Methods for details). In particular, we obtained 1,713 epistatic interactions between 1,111 locus pairs and 323 probes (315 genes) (Table S1), suggesting that up to 6% of all parasite genes might be regulated by epistatic interactions.

Interestingly, we found that epistatic interactions were enriched in triplets that were best described by configurations 7 and 10. Commonly, these two configurations involved the expression of the target gene on one level when the interacting loci are inherited from the same parental strain (genotype 00 or 11) while, at the same time there is the opposite

expression level when loci were inherited from different parental strains (genotype 01 or 10). In Boolean terms such configurations are best described by the ‘XOR’ function.

To assess the advantage of our approach, we compared the performance of SEE to the results obtained by a step-wise approach. Since the original step-wise method [7] used a less rigorous epistasis model, we implemented a modified version. While we applied the same filtering step to find strong primary loci, we checked for the presence of epistatic interactions with secondary loci using our statistical framework. We found a relatively small overlap with our original results. Given the small number of progenies such an observation is rather expected since our method has difficulties to uncover interacting loci with a strong primary effect. In turn, the step-wise method is known to miss interactions in the absence of such effects.

Epistatic interactions versus one-locus eQTLs

Since our method emphasized synergistic effects of locus pairs on gene expression, we observed large differences in the genome-wide distributions of one-locus eQTLs [15] and epistatic interacting locus pairs (Fig. 2a). In Fig. 2b, we showed the genome-wide profile of the number of partner loci for any given locus. In addition, we observed that the number of a locus’ target genes is significantly correlated to the number of its interacting partner loci (Pearson correlation $r_p = 0.82$, $p < 10^{-10}$, Fig. 2c).

Epistatic interaction hotspots

We defined a pair of loci as an epistatic interaction hotspot if the two loci synergistically co-regulated at least 10 target genes. Applying this criterion, we found 14 such epistatic interaction hotspots with p -value $< 10^{-10}$ (see Methods for details). Since some hotspots were close to each other on the genome, and their sets of target genes often significantly overlapped, we grouped hotspots into four hotspot regions (see Methods). We found that the epistatic interaction hotspots contained loci from subtelomeric regions on chromosome 3, 7, 9, 12 and 14 (Table S2 in Additional File 1, Fig. 3a), which also harboured many previously detected one-locus hotspots [15]. In addition, the majority of epistatically interacting locus pairs had a relatively small number of target genes while a minority accumulated a large number of targets (Fig. 3b).

Increased LD is associated with epistatic interaction hotspots

We wondered if the pairs of genetic regions that were involved in epistatic interaction hotspots had any linkage disequilibrium pattern. While we found a total of 49,638 locus pairs that were on different chromosomes, only 388 locus pairs had a LD measure $r^2 > 0.2$, where r^2 was calculated using genotype data of progeny strains. Two out of 14 epistatic interaction hotspots had $r^2 > 0.2$ ($p < 0.006$). Computing Pearson correlation r_p between the numbers of target genes of two epistatically interacting loci and r^2 between the two loci, we found a weak but significant correlation of $r_p \approx 0.15$ ($p < 10^{-6}$, Fig. 3c). This observation indicated that high LD is associated with a large number of target genes, suggesting that high LD between epistatically interacting loci might have been maintained for regulatory functions.

Target functions influenced by epistatic interaction hotspots

First, we focused our functional analysis on hotspot (3_8.6, 7_2.9) that had highest LD among all epistatic interaction hotspots ($r^2 = 0.24$). Utilizing GO-specific functional gene annotations from GeneDB [17] we observed that the set of target genes was enriched with genes carrying a methyl transferase domain (The list of target genes of this hotspot is provided in Table S2). Two out of 56 genes that appeared in the “methyl-transfer pathway”

also were among the target genes of this hotspot ($p < 0.01$, hypergeometric test). Both of these genes use the same methyl donor S-adenosylmethionine (SAM) and are homologous to rRNA-methylating enzymes, suggesting complex and tight regulation of this pathway. SAM is a ubiquitous methyl donor in many biochemical pathways, ranging from methylation of proteins, lipids and nucleic acids to serving as a precursor in polyamine biosynthesis. Therefore, perturbation of SAM levels would indeed be expected to have a wide range of consequences that could be consistent with the dramatically different drug selection histories of the parents of the HB3 $\times$ Dd2 cross. For example, a recent study [13] contrasted expression divergence following drug pressure in a laboratory line, demonstrating the differential expression of the SAM dependent methyltransferase gene *PFE1115c*, located in a chromosome 5 locus that was previously linked to multidrug resistance. Furthermore, an earlier study [18] reported a higher sensitivity of the chloroquine/pyrimethamine resistant strain K1 to polyamine analogs (*i.e.* pathway inhibitors) as compared to the drug sensitive strain NF54.

As defined in GeneDB, our set of target genes also was enriched with components of the proteasome degradation machinery and their timed transcription, with 3 out of 85 such genes appearing in the set of 14 genes ($p < 0.002$). Other interesting target genes in this epistatic interaction hotspot included *PF07_0038*, a gene located close to the chloroquine resistance transporter (*Pfcr*) [19], and a general transcription factor *PFB0290c* that could be widely involved in trans-regulation. Expression levels of this transcription factor in the eQTL experiment showed a significant correlation to *Pfcr* in both chloroquine-resistant and sensitive progenies (average correlation $\bar{r}_p = 0.72$), suggesting a possible collaborative role in the endogenous function of the transporter *Pfcr*. Considering results of previously mentioned studies such observations indicate a potential importance of this hotspot in the context of drug resistance.

Focusing on another hotspot (7_20.2, 8_48.9) that harbours 10 target genes, we found 3 erythrocyte-targeting genes. While *PF11_0486* is a merozoite adhesive erythrocyte binding protein and *PFE1625c* is a pseudo gene for erythrocyte membrane protein 1 (*PIEMP1*), *PFL2575c* is a Plasmodium exported protein that carries an erythrocyte trafficking motif on the second exon. Considering that more than 400 erythrocyte-targeting genes [20] exist, the probability of randomly detecting three such genes in this hotspot combination was $p < 0.03$, indicating a potential regulatory signal emerging from this hotspot locus pair.

Interestingly, 3 of 4 members of the pyruvate dehydrogenase complex that plays an important role in redox reactions appeared to be epistatically regulated. Redox reactions are essential when *P. falciparum* degrades hemoglobin in human red blood cells. Specifically, we found *PF11_0256*, pyruvate dehydrogenase E1 alpha subunit, regulated by hotspot (7_20.2, 8_48.9). While *PF08_0066* is a lipoamide dehydrogenase, regulated by hotspot (3_2.9, 7_20.2) we noticed *PF10_0407*, dihydrolipoamide acyltransferase component E2, among target genes of hotspot (12_28.6, 14_0).

Potential mechanisms underpinning the epistasis hotspot between loci 3_8.6 and 7_2.9

We considered possible mechanisms for the regulatory basis of the transcription of the target genes for the epistatic hotspot between loci 3_8.6 and 7_2.9. We confined the search for candidate regulators in each of these loci by considering only genes located within the regions physically bounded by each marker (3_8.6 or 7_2.9) and their flanking markers. In the simplest case, epistatic transcriptional regulatory loci would contain transcription factors or their regulators. Although we did not identify direct transcription regulators in either locus, it may be relevant that locus 3_8.6 contains 3 ribosomal protein genes (*PFC0290w*, *PFC0295c*, *PFC0300c*) while 7_2.9 contains a SAM-dependent tRNA m5C-

methyltransferase (PF07_0015); consequently, both loci carry critical components of translation. Furthermore, the target genes for the (3_8.6, 7_2.9) epistatic interactions are enriched with SAM-dependent methyltransferases, consistent with this regulatory theme. Recent studies specify that tRNAs can be cleaved into small RNAs [21] and that tRNAs are protected from this cleavage by cytosine methylation [22]. Moreover, methylation of tRNA serves a regulatory role in stem cell differentiation. Yet, there is no evidence as to whether this acts directly through control of mRNA translation or indirectly through small RNAs derived from tRNAs [23]. Accordingly, our study suggests that epistasis between these two loci indirectly mediates methylation of tRNA.

Epistatic loci harbor transcription regulators

We identified several cases where two direct transcription regulators were within epistatically interacting locus pairs. In one case, the genomic region of epistatically interacting loci 3_0 and 7_11.5 contained RNA polymerase subunits *ABC23* (*PFC0115c*) and *ABC10* (*PF07_0027*). In the other case, the epistatically interacting loci 3_0 and 5_88.9 coded for transcription factor II H p52 subunit (*PFL2125c*) and a homolog of the transcription elongation factor *SPT6* (*PF14_0059*), respectively. The p52 subunit is essential for the phosphorylation of the carboxy terminal domain (CTD) of RNA polymerase II while *SPT6* plays an important role in this reaction for the transition from transcription initiation to elongation.

We also observed an epistatically interacting locus pair (14_77.5, 4_66) that contained members of the recently identified ApiAP2 family of putative transcriptional regulators (*PF14_0271* and *PFD0985w*), where the target genes were two heat shock proteins *PF11_0216* and *PF13_0021*. Since position specific matrices were available for *PFD0985w* only [24], we scanned the upstream regions of the two target genes with Matrix-scan [25]. Indeed, we found potential binding sites within 1,000 bp upstream of genes *PF11_0216* and *PF13_0021* for domain D2 of ApiAP2 protein *PFD0985w* ($p < 10^{-5}$), providing additional evidence for the potential role of the newly discovered ApiAP2 family in regulating gene expression.

Comparison of epistatic interactions in *P. falciparum* and *S. cerevisiae*

To capture properties that might be specific to parasite crosses, we applied SEE to yeast progenies. Using yeast eQTL data [14] from 112 progenies, we observed 110 epistatic interactions (Table S3). Even with FDR < 0.7 this set of epistatic interactions in yeast is considerably smaller than the corresponding set in *P. falciparum*, ruling out that the observed difference between the two species was an artefact of the small number of strains in the parasite crosses (see Methods). In addition, such a small number was comparable to the set of 170 two-locus interactions (including non-epistatic interactions) obtained by the original step-wise method in [7].

To gain further insight into the differences between Plasmodium and yeast crosses, we computed LD patterns in the yeast cross and in a recently generated *P. falciparum* cross 7G8 × GB4 that has 281 genetic markers and 32 progeny clones [26]. High LD was also a feature of this *P. falciparum* cross although we observed a different LD pattern than in the HB3 × Dd2 cross (Table 1). In comparison, LD in yeast was much lower than in *P. falciparum*, suggesting that both epistatic interaction hotspots and elevated linkage disequilibrium could reflect selection in the generation and expansion of the *P. falciparum* progeny clones after isolation from the chimp blood.

Discussion

Understanding the scale and the role of multiple loci linkage in general and epistasis in particular is an important challenge of today's association studies and eQTL analysis. The large number of possible expression traits and pairs of loci causes a considerable multiple testing problems. Therefore, a key component of computational approaches for the detection of interactions is a filtering strategy to reduce the number of performed test. Previously reported step-wise methods limit the number of candidate interactions by requiring the presence of a primary locus that has a significant marginal effect on the underlying trait. However, such approaches fail to detect interactions with small individual effects. In contrast, our new method SEE determines candidate interactions in a way that increases the likelihood to detect epistatic interactions between loci without requiring a significant contribution of any individual locus. Especially in small sets of progenies, however, the design of the initial filtering step may limit SEE's ability to detect epistatic interactions that involve a locus with a relatively large primary effect. In a recently proposed, alternative approach Zhang *et al.* built on expression modularity to detect locus interactions [11]. Defining modules as a set of genes and markers, this approach searches for modules where genes are linked to loci. Specifically, modules that are linked to two loci were considered candidates for possible epistatic effects, allowing the detection of three epistatic interaction modules that contained a total of only 39 genes in yeast. Such a module centric approach can uncover interactions that might be too weak to be detected on gene-by-gene basis but might miss interactions that do not involve a coherent module.

In comparison, SEE differs in the way the search space is reduced as well as in the statistical way the significance of an epistatic interaction is assessed. In particular, Zhang *et al.* used t-statistics to test if the interaction term in eq. 1 is non-zero [11], where the trait is modelled as the first principle component of a module's gene expression matrix. Recently, however, Wang *et al.* suggested that the non-epistatic interaction model (a statistical framework without the interaction term in eq. 1) might be sufficient to explain most biologically relevant interactions in human [27]. We therefore adopted a more stringent statistical test of epistasis than previous methods in which we assessed the impact of the epistatic interaction term we fitted the expression of individual genes by both the epistatic (eq. 1) and additive (eq. 2) model, allowing us to check if the epistatic model provided a significantly better fit than the additive model. This statistical test provides a higher confidence that uncovered interactions were indeed non-additive than a test for the significance of the interaction term. Only five out of 29 interactions defined by two epistatic interaction modules discovered by Zhang *et al.* [11] and only 19 out of 170 interactions reported in Storey et al. [7] passed this test. Notably, we generally found that all methods provided small sets of epistatic interactions in yeast, while overlaps between sets of predicted interactions were small as well. Such a result is hardly surprising given the difference in the statistical models, and the small number of epistatic interactions that were detected by each method.

Providing the first view on the structure of epistatic interactions in *P. falciparum* we revealed extensive epistasis and epistatic interaction hotspots. Since one of the Dd2 parental strain in the HB3 $\times$ Dd2 cross is multi-drug resistant while HB3 is generally drug sensitive, we conclude our findings may be valuable leads for studies of the parasite's drug resistance. Interestingly, epistatic interactions hotspots were associated with elevated levels of linkage disequilibrium, suggesting selection pressure acting on *P. falciparum*. Currently, we cannot fully ascertain if this observed selection pressure is independent of the host or occurs in the context of a specific chimpanzee host. A new *P. falciparum* cross 7G8 $\times$ GB4 also showed high LD with a pattern that differed from the HB3 $\times$ Dd2 cross, indicating that the observed effect might be strain dependent. Indeed, Dd2 has been a subject of intensive Chloroquine and Sulfadoxine-Pyrimethamine related pressure in nature and in the lab. Since GB4 and

7G8 also carry different origins of Chloroquine resistance difference in drug selection might impact the pattern of linkage disequilibrium in this cross.

Specific mechanisms of interactions between multiple loci in the parasite are yet to be discovered, but various mechanisms could be envisioned. For example, interactions may exist among loci, harbouring interacting members of the transcriptional machinery, and we indeed found two cases that supported such a scenario. Epistatic regulatory effects can also arise when two loci encode transcription factors that regulate expression of one or more genes in an indirect way. Consistent with such a mechanism, we found an interacting locus pair encoding potential ApiAP2 proteins. Members of this family of putative transcription factors have been demonstrated to bind to very closely related DNA motifs in the genome of *P. falciparum* [28]. Such proteins may bind similar regulatory motifs, potentially contributing to competitive binding, leading to the observed epistasis. Indeed, we found potential binding sites in the promoter region of both target genes of the corresponding interacting locus pair, suggesting that such transcriptional regulation at interacting loci could allow the pathogen to fine tuning gene regulation.

Conclusion

We have developed a new computational approach, SEE (Symmetric Epistasis Estimation), to identify epistatic interactions. Mitigating the multiple testing problem our method significantly reduces the number of tests using a novel filtering approach. Unlike most of previous methods, SEE treats both interacting loci in a symmetric way, complementing existing approaches. Thanks to this important property of our approach, we performed the first study of epistatic interactions in *P. falciparum* eQTL, allowing us to obtain more than 1,700 epistatic interactions. Illustrating a genome wide epistatic interaction landscape in the malaria pathogen we pinpoint epistatic interaction hotspots whose properties suggest importance for drug resistance adaptations. Of particular interest is a weak but significant correlation between epistatic interaction hotspots and patterns of linkage disequilibrium, emerging from our analysis. Such linkage disequilibrium suggests a selection pressure acting on *P. falciparum* that potentially reflects host-pathogen interactions or drug-induced selection and thus provides additional line of evidence of supporting relation of hotspots to drug resistance.

Experimental

Genomic and expression data

We used the data of the single locus eQTL study by Gonzales *et al.* which utilized 34 progeny strains with 329 markers obtained from a HB3 $\times$ Dd2 cross in *P. falciparum* [12]. Genome-wide expression levels were measured 18 hours after the parasite invaded human erythrocytes (RBCs), using 7,665 probes that represented 5,150 ORFs. For the yeast eQTL analysis, we used data of a BY $\times$ RM cross [14], covering 112 progenies, 5,747 genes and 2,956 genetic markers. We binned adjacent markers if the Hamming distance between their genotype data (without considering missing data) was less than 8 [29], reducing the set of markers to 558, close to the number of genetic markers in *P. falciparum*.

Representing eQTL data with a graph-theoretic framework and detection of eQTL association cliques

Building on our previous work [15] we first constructed an eQTL association graph with three partitions of nodes, corresponding to strains (*S*), genes (*G*), and loci (*L*) and then search for tripartite cliques in this graph. This step is similar as in [15] and here we summarize it very briefly. The way we utilize these cliques, described in the subsequent

sections, is novel. Each gene was represented by two nodes g_u and g_d , corresponding to g either being up- or down-regulated. Specifically, we considered the expression value of g in a strain up(down)-regulated if it was equal or larger(smaller) than the mean expression value of g across all strains plus (minus) one standard deviation. An edge between g_u/g_d and a progeny strain s indicated that g 's expression is up-regulated or down-regulated in strain s . An edge between I_0/I_1 and a progeny strain s indicated that the genotype of I was 0/1 in strain s (Fig. 1a). In the eQTL association graph we detected eQTL maximal association cliques used a previously described approach [15]. Association clique is defined as a maximal clique that contains a subset of loci L' , a subset of progeny strains P' , and a subset of genes G' . In an association clique, each locus in L' has the same genotype (0 or 1) among all progeny strains in P' , and each gene is up-/down-regulated among all progeny strains in P' . Vertices in L' are fully connected to vertices in P' , and vertices in G' are fully connected to vertices in P' as well. In other words, each eQTL association clique is a maximal subgraph that cannot be extended further, maintaining full connectivity (Fig. 1a).

Interacting configurations

We scanned such maximum eQTL association cliques in order to find candidate triplets. We defined 16 possible combinations of joint genotypes and expression states that can be assigned to a triplet that consisted of two loci I' , I'' and gene g (Fig 1b). Possible joint genotype combinations for I' and I'' are 00, 01, 10, and 11. For the purpose of defining configurations, we considered two possible states (up or down) for the expression of a gene g . Accordingly we defined $2^4 = 16$ possible configurations. Out of these 16 configurations, we identified eight interacting configurations, while the remaining configurations were considered non-interacting. Intuitively, an interacting configuration can only be explained by the synergistic interaction model, while a configuration is defined non-interacting otherwise. For example, configuration 7 is an interacting configuration (Fig. 1b), where the joint genotypes 00 and 11 are accompanied by a down-regulated gene g , while 01 and 10 coincides with an up-regulated g . If we simply encode down-/up-regulation of gene expression as 0/1, configuration 7 can be considered a XOR function that can be captured by our synergistic model using $y = x' + x'' - 2x'x'' + \varepsilon$. In contrast, configuration 4 is non-interacting where the underlying gene g is up-regulated when locus $I' = 1$ and *vice versa*. The absence of a clear impact of I'' on the expression of g suggests that the additive model is sufficient to explain this configuration.

Configurations in a given triplet might not be always consistent for all strains. Therefore, while selecting candidate triplets, we demanded that each triplet had an interacting configuration that was consistent across a large enough number of strains. Specifically, we computed the support for the consistency of the triplet I', I'' with a configuration c , $sp_c^{I', I''}$ follows. We defined the supporting subgraph as an association clique where the underlying triplet appeared. Next, we denoted by S_h ($h \in [1, \dots, 4]$) the progeny set in such supporting subgraph. In Fig. 1a, we show a triplet represented by four subgraphs that fit configuration 7 where $S_1 = (s_2, s_6)$, $S_2 = (s_3, s_4)$, $S_3 = (s_1)$ and $S_4 = (s_5)$. We defined

$S_c^{I', I''} = |S_1| + |S_2| + |S_3| + |S_4|$. Since a triplet may also fit non-interacting configurations, we defined $sp_{IC}^{I', I''} = \max_{c \in IC} (sp_c^{I', I''})$ and $sp_{NIC}^{I', I''} = \max_{c \in NIC} (sp_c^{I', I''})$, measuring the support for interaction (IC) and non-interaction (NIC) configurations of a given triplet. To qualify as a candidate triplet, we demanded that the corresponding support for interaction configurations (IC) reached a certain level and was larger than for non-interaction configurations (NIC). In particular, we defined $sp_{\Delta}^{I', I''} = sp_{IC}^{I', I''} - sp_{NIC}^{I', I''}$ for each triplet. For *P. falciparum*, we required that each association clique contained at least five strains, $sp_{\Delta}^{I', I''} \geq 5$ and $sp_{IC}^{I', I''} \geq 13$. For

yeast, we required that each association clique had at least 12 strains, $sp_{\Delta}^{\prime\prime} \geq 2$ and $sp_{ic}^{\prime\prime} \geq 35$. Intuitively, the requirement that association cliques contained more strains improved sensitivity defined as the percentage of the number of significant interactions over the number of candidate triplets but decreased the number of candidate triplets (See Additional File 1).

Testing candidate triplets for epistatic interactions

After obtaining candidate triplets, we estimated their significance using residual empirical thresholds (RET) [30,31]. For each candidate triplet I', I'', g , we first obtained a residual value by removing the main effects of two loci from the expression value. Instead of discretized values we used the original expression values to compute the residual and evaluated the probability that the synergistic interaction model fitted the data better than the additive model. Computing the residual for the H^{th} expression value of g in the strain where the genotype of I_i and I_j was 00, we applied the following equation

$$r_h^{00} = y_h^{00} - \frac{-0\bullet}{y} - \frac{-0\bullet}{y} + \frac{-\bullet\bullet}{y},$$

where y_h^{00} was the H^{th} expression value of gene g in the strain where the genotypes of I' and I'' were 00. $\frac{-0\bullet}{y}$ ($\frac{-\bullet 0}{y}$) was the mean expression of g in strains where the genotype of $I'(I'')$ was 0, and $\frac{-\bullet\bullet}{y}$ was the mean expression of gene g across all strains. Similarly, we computed the residual for g 's expression values in strains where the genotype of I' and I'' were 01, 10 and 11. We fitted the residual with the synergistic model, where the p -value for the interaction term in the regression was taken as the nominal p -value for comparing the synergistic to the additive model.

Correcting p-values and estimating FDR

To correct nominal p -values obtained in the previous step we generated a null p -value distribution from randomized eQTL data. To control for linkage disequilibrium and any skewed inheritance patterns we kept the genotype data intact in the randomization procedure. Specifically, we randomly permuted strain labels and kept gene specific expression profiles unchanged. We applied SEE to such randomized data utilizing exactly the same thresholds and parameters as used for the real data. We repeated this process 10 times and calculated a nominal p -value for each triplet as described previously. Forming a 'global' null distribution by pooling all nominal p -values from random data we defined a corrected p -value as the fraction of random p -values at or below the p -value of the given candidate triplet. After we confirmed that the distribution of corrected p -values of random triplets was flat (data not shown), we calculated a FDR for candidate triplets [16]. Applying a corrected p -value threshold = 0.001 (FDR < 0.003) we obtained a set of 3,796 triplets.

Validation of FDR estimation

To estimate the number of false positives in an alternative way, we computed how many "epistatic interactions" were produced by the previously described random generation of eQTL data. Specifically, we applied SEE to 10 random eQTL datasets and computed p -values in exactly the same way as for the real-data. Averaging over all 10 random runs, we obtained 12 "epistatic interactions", suggesting that our FDR estimate was indeed accurate ($3,796 \times 0.003 \approx 11$).

Accounting for the potential impact of LD

Our procedure of correcting nominal p -values using the global null distribution dis-accounted for the possibility that a gene expression trait might be mapped to multiple, genomically close locus pairs. Since linkage disequilibrium in *P. falciparum* is high, two neighboring loci on the same chromosome tend to have similar genotypes. To further select the most significant epistatic interactions, we corrected nominal p -values in an alternative way. We generated a null distribution for a group of candidate triplets sharing the same gene, where locus pairs of triplets in a group were close to each other. In particular, locus pairs l'_1, l''_1 and l'_2, l''_2 were considered close to each other if there was at most one locus between l'_1 and l'_2 . For each candidate triplet in a group of size n , we randomly permuted the residual and fitted the permuted residual with the synergistic (*i.e.* epistatic interaction) model to obtain a null p -value. Repeating this process m/n times for each candidate triplet we obtained a null distribution of m p -values for each group and defined a corrected p -value as the fraction of values in the distribution that was less than the given p . Setting $m = 1,000$ and requiring that selected epistatic interactions had a corrected p -value < 0.001 we found 1,713 out of 3,796 initially obtained epistatic interactions, involving 1,111 locus pairs and 323 probes (315 genes).

Detection of epistatic interaction hotspots

We defined a pair of loci as an epistatic interaction hotspot if the two loci synergistically co-regulated at least 10 target genes. We grouped overlapping hotspots into hotspot regions by binning two epistatic interactions with locus pairs l'_1, l''_1 and l'_2, l''_2 if there was at most one locus between l'_1 and l'_2 , and one locus between l''_1 and l''_2 . Recently Breitling *et al.* [32] proposed a permutation method to accurately estimate the clustered distribution of false eQTLs (*i.e.* the size of false eQTL hotspots) on the genome. Their method randomly permuted columns in the genotype matrix, where rows corresponded to loci, and columns represented progeny strains. In addition, the gene expression matrix was kept unchanged and the integrity of the correlation structure of gene expression matrix was maintained, a major source of false eQTL hotspots. To detect eQTL hotspots, eQTL mapping was subsequently applied to randomized data. Following this procedure and subsequently applying SEE with the same parameters used for real data, we found that the average maximum number of target genes was < 2 ($p = 2 \times 10^{-11}$, one-sided t -test), indicating that 10 was a reasonable threshold for defining real hotspots.

Testing the effect of small progeny number using yeast eQTL

Since the number of progenies of *P. falciparum* was relatively small, we used yeast eQTL data [14] to investigate the impact of different numbers of progenies on the performance of SEE. We randomly selected five sets of 34 progenies together with their associated genotype and gene expression profiles. Using the parasite specific parameter setting we applied SEE to the five yeast progeny sets. On average we found less than one candidate triplet with corrected p -values < 0.05 , suggesting that no spurious epistatic interactions were detected in yeast when only 34 progenies were used. Considering limited numbers of detected epistatic interactions and their high FDR obtained from the full yeast eQTL dataset, we ruled out the possibility that a large number of epistatic interactions in *P. falciparum* was an artefact of a small set of progenies, producing a large number of false positives.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet.* 2009; 10:392–404. [PubMed: 19434077]
2. Musani SK, Shriner D, Liu N, Feng R, Coffey CS, et al. Detection of gene x gene interactions in genome-wide association studies of human population data. *Hum Hered.* 2007; 63:67–84. [PubMed: 17283436]
3. Fisher RA. The correlation between relatives on the supposition of mendelian inheritance. *Philosophical Transactions of the Royal Society of Edinburgh.* 1918; 52:399–433.
4. Phillips PC. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* 2008; 9:855–867. [PubMed: 18852697]
5. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet.* 2002; 11:2463–2468. [PubMed: 12351582]
6. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet.* 2005; 37:413–417. [PubMed: 15793588]
7. Storey JD, Akey JM, Kruglyak L. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol.* 2005; 3:e267. [PubMed: 16035920]
8. Hannum G, Srivas R, Guenole A, van Attikum H, Krogan NJ, et al. Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet.* 2009; 5:e1000782. [PubMed: 20041197]
9. Litvin O, Causton HC, Chen BJ, Pe'er D. Modularity and interactions in the genetics of gene expression. *Proc Natl Acad Sci U S A.* 2009
10. van Noort V, Huynen MA. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.* 2006; 22:73–78. [PubMed: 16380193]
11. Zhang W, Zhu J, Schadt EE, Liu JS. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput Biol.* 2010; 6:e1000642. [PubMed: 20090830]
12. Gonzales JM, Patel JJ, Ponmee N, Jiang L, Tan A, et al. Regulatory hotspots in the malaria parasite genome dictate transcriptional variation. *PLoS Biol.* 2008; 6:e238. [PubMed: 18828674]
13. Jiang H, Patel JJ, Yi M, Mu J, Ding J, et al. Genome-wide compensatory changes accompany drug-selected mutations in the *Plasmodium falciparum* crt gene. *PLoS One.* 2008; 3:e2484. [PubMed: 18575593]
14. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science.* 2002; 296:752–755. [PubMed: 11923494]
15. Huang Y, Wuchty S, Ferdig MT, Przytycka TM. Graph theoretical approach to study eQTL: a case study of *Plasmodium falciparum*. *Bioinformatics.* 2009; 25:i15–20. [PubMed: 19477981]
16. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A.* 2003; 100:9440–9445. [PubMed: 12883005]
17. Hertz-Fowler C, Peacock CS, Wood V, Aslett M, Kerhornou A, et al. GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res.* 2004; 32:D339–343. [PubMed: 14681429]
18. Klenke B, Barrett MP, Brun R, Gilbert IH. Antiplasmodial activity of a series of 1,3,5-triazine-substituted polyamines. *J Antimicrob Chemother.* 2003; 52:290–293. [PubMed: 12837742]
19. Su X, Kirkman LA, Fujioka H, Wellems TE. Complex polymorphisms in an approximately 330 kDa protein are linked to chloroquine-resistant *P. falciparum* in Southeast Asia and Africa. *Cell.* 1997; 91:593–603. [PubMed: 9393853]
20. Marti M, Good RT, Rug M, Knuepfer E, Cowman AF. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science.* 2004; 306:1930–1933. [PubMed: 15591202]
21. Lee YS, Shibata Y, Malhotra A, Dutta A. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.* 2009; 23:2639–2649. [PubMed: 19933153]
22. Schaefer M, Pollex T, Hanna K, Tuorto F, Meusburger M, et al. RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. *Genes Dev.* 2010; 24:1590–1595. [PubMed: 20679393]

23. Blanco S, Kurowski A, Nichols J, Watt FM, Benitah SA, et al. The RNA-Methyltransferase Misu (NSun2) Poises Epidermal Stem Cells to Differentiate. *PLoS Genet.* 2011; 7:e1002403. [PubMed: 22144916]
24. Campbell TL, De Silva EK, Olszewski KL, Elemento O, Llinas M. Identification and Genome-Wide Prediction of DNA Binding Specificities for the ApiAP2 Family of Regulators from the Malaria Parasite. *PLoS Pathog.* 2010; 6:e1001165. [PubMed: 21060817]
25. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc.* 2008; 3:1578–1588. [PubMed: 18802439]
26. Hayton K, Gaur D, Liu A, Takahashi J, Henschen B, et al. Erythrocyte binding protein PfRH5 polymorphisms determine species-specific pathways of *Plasmodium falciparum* invasion. *Cell Host Microbe.* 2008; 4:40–51. [PubMed: 18621009]
27. Wang X, Elston RC, Zhu X. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction? *Nat Rev Genet.* 2011; 12:74. [PubMed: 21102529]
28. De Silva EK, Gehrke AR, Olszewski K, Leon I, Chahal JS, et al. Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci U S A.* 2008; 105:8393–8398. [PubMed: 18541913]
29. Lee SI, Pe'er D, Dudley AM, Church GM, Koller D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc Natl Acad Sci U S A.* 2006; 103:14062–14067. [PubMed: 16968785]
30. Doerge RW, Churchill GA. Permutation tests for multiple loci affecting a quantitative character. *Genetics.* 1996; 142:285–294. [PubMed: 8770605]
31. Anderson MJ, Braak CJFT. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation.* 2003; 73:85–113.
32. Breitling R, Li Y, Tesson BM, Fu J, Wu C, et al. Genetical genomics: spotlight on QTL hotspots. *PLoS Genet.* 2008; 4:e1000232. [PubMed: 18949031]

(a) In the first step, we constructed a tripartite eQTL association graph where nodes are partitioned, representing the expression status of genes, genotypes of loci and specific strains. Subsequently, we identified eQTL association cliques and determined candidate triplets that were consistent with interaction configurations. Comparing the additive and the synergistic models, we statistically tested each candidate triplet for the presence of an epistatic interaction. (b) Combinations of joint binary locus genotypes and expression state of a gene make up 16 different configurations. Bright coloured configurations correspond to combinations consistent with synergistic effects, defined as interacting configurations (2, 3, 5, 7, 10, 12, 14, 15). Dim coloured patterns are defined as non-interacting configurations (1, 4, 6, 8, 9, 11, 13, 16) since the expression of the underlying gene was explained best by individual or additive effects.

Figure 2. Characteristics of epistatic interactions

In the upper panel of (a) we show the genome-wide distribution of the number of target genes as a function of the genomic positions of given epistatically interacting locus pairs. While we found some overlaps with one-locus eQTLs as detected by the GeD method [15] (lower panel), distributions appear to be largely different. (b) Genome-wide distribution of the number of partner loci, which co-appeared with a given locus in an epistatic interaction. In (c) we found that the number of partner loci and the number of target genes of a given locus are strongly correlated (Pearson's $r = 0.83$, $p < 10^{-10}$).

Figure 3. Landscape of epistatic hotspots

(a) In a 2D plot of genome-wide distribution of epistatically interacting locus pairs we grouped neighbouring loci in bins. In particular, we demanded that the LD measure r^2 between any two loci in one bin was at least 0.75. Selecting the middle locus as the representative of each group we found 4 hotspots where two loci synergistically co-regulated at least 10 target genes. (b) In a distribution of the number of targets of interacting locus pairs the majority of epistatically interacting pairs had a small number of target genes while a small number of pairs regulated a large number of genes. In (c) we logarithmically grouped the number of target genes and calculating mean LD in each bin. Using untreated data we found a significant correlation between the number of target genes and LD measure, r^2 , of two epistatically interacting loci (Pearson's $r = 0.15$, $p < 10^{-6}$).

Table 1

Mean LD between two loci on two different chromosomes in HB3 × Dd2 and 7G8 × GB4 crosses of *P. falciparum* and BY × RM cross of *S. cerevisiae*.

	<i>P. falciparum</i> HB3 × Dd2	<i>P. falciparum</i> 7G8 × GB4	<i>S. cerevisiae</i> BY × RM
Mean of top 1% of LD values	0.23	0.30	0.09
Mean of top 10% of LD values	0.13	0.16	0.04
Mean of all LD values	0.03	0.04	0.009