

METHODOLOGY ARTICLE

Open Access

Data-driven assessment of eQTL mapping methods

Jacob J Michaelson¹, Rudi Alberts², Klaus Schughart², Andreas Beyer^{1*}

Abstract

Background: The analysis of expression quantitative trait loci (eQTL) is a potentially powerful way to detect transcriptional regulatory relationships at the genomic scale. However, eQTL data sets often go underexploited because legacy QTL methods are used to map the relationship between the expression trait and genotype. Often these methods are inappropriate for complex traits such as gene expression, particularly in the case of epistasis.

Results: Here we compare legacy QTL mapping methods with several modern multi-locus methods and evaluate their ability to produce eQTL that agree with independent external data in a systematic way. We found that the modern multi-locus methods (Random Forests, sparse partial least squares, lasso, and elastic net) clearly outperformed the legacy QTL methods (Haley-Knott regression and composite interval mapping) in terms of biological relevance of the mapped eQTL. In particular, we found that our new approach, based on Random Forests, showed superior performance among the multi-locus methods.

Conclusions: Benchmarks based on the recapitulation of experimental findings provide valuable insight when selecting the appropriate eQTL mapping method. Our battery of tests suggests that Random Forests map eQTL that are more likely to be validated by independent data, when compared to competing multi-locus and legacy eQTL mapping methods.

Background

For decades scientists have used a variety of analytical techniques to relate allelic inheritance patterns in the genome to variation in continuous physical traits of interest. The goal of such analyses is often to locate quantitative trait loci (QTL), or genomic locations that exert an influence on the manifested trait. Understanding the genomic location of these genetic control points may provide insight into the genetic and molecular framework responsible for enabling the trait.

In the past decade, the advent of the DNA microarray and other high-throughput molecular technologies has updated the paradigm of the QTL. A QTL where mRNA expression is the complex trait of interest is generally referred to as an expression QTL or eQTL [1]. By using DNA microarrays eQTL can be measured for basically all genes in the genome, rendering eQTL data information rich and potentially very powerful. eQTL have been

studied in yeast, mouse, rat, human, and plants [2-6] and eQTL have proven to be useful for elucidating the molecular mechanisms of human diseases [7-10].

Although complex traits are by definition controlled by the coordination of multiple genes, the prevailing techniques for mapping them have been deeply rooted in univariate thinking - testing for genetic association to a trait one locus at a time, ignoring combinatorial effects and interactions. In contrast, Broman and Speed [11] defined the QTL problem as one of multivariate variable selection, where ideally all loci and their combinations are allowed to enter and exit the model as the data dictate. Viewing eQTL mapping as a variable selection problem opens the door to using a host of machine learning algorithms which have rarely, if at all, been applied to QTL and eQTL studies [12-15]. Such a fresh look at the QTL problem may help to uncover latent and meaningful information in otherwise underexploited data.

A systematic comparison of eQTL mapping approaches is necessary to inform the research community which methods work best and in which contexts. Toward that goal, the purpose of this work is twofold.

* Correspondence: andreas.beyer@biotec.tu-dresden.de

¹Cellular Networks and Systems Biology, Biotechnology Center - TU Dresden, Dresden, Germany

Full list of author information is available at the end of the article

First, we establish a framework for comparing available eQTL mapping methods based on the tendency of each method to map eQTL that are systematically supported by external biological data. This is important because methods papers proposing new (e)QTL mapping techniques often draw their conclusions either solely or largely on the basis of simulated data [11,12,14,16-21]. This is perhaps understandable in the case of earlier work with QTL, where only a limited number of phenotypes were available and external knowledge about their context and probable genetic regulators was not available in a systematic form, making biology-based benchmarking difficult. However, this is not the case in the era of eQTL. Although some genes remain uncharacterized, there are rich sources of data for many genes that give insight about their role and context within the cell. Such knowledge is often contained in databases like the Kyoto Encyclopedia of Genes and Genomes (KEGG) [22], which makes using it as a basis for a benchmark easier. Our battery of knowledge-driven benchmarks consists of 1) assessing the proportion of *cis*-eQTL recovered by each method, 2) testing each method's high-scoring eQTL for enrichment of loci related to the target by KEGG pathway information, and 3) agreement of each method's high-scoring eQTL with systematic loss-of-function studies. In this framework we tested three variable importance measures from Random Forests (RF) [23] as well as sparse partial least squares (SPLS) [12], the lasso [24], the elastic net [25], Haley-Knott regression (HK) [20], and composite interval mapping (CIM) [19]. We also performed simulations to complement the findings of the knowledge-driven benchmarking framework. We show that multi-locus methods in general (Random Forests, SPLS, lasso, elastic net) are better at recovering biologically meaningful loci than traditional QTL mapping methods such as HK and CIM. Second, we demonstrate that based on both simulations and the knowledge-driven benchmarks, RF shows superior performance as an eQTL mapping method. RF has previously been applied to genome-wide association studies (GWAS) and QTL studies [14,15,17,26,27]. The contribution of our work, however, lies in the discovery that the most naive measure of variable importance in RF, the variable selection frequency (RFSF), actually performs much better than the more popular permutation importance (RFPI) in this context. Since RFSF has been ignored in all previous works using RF in the QTL or GWAS context, its use here represents a novel eQTL mapping method with demonstrated superior performance.

Results

In order to evaluate the performance of the eQTL mapping methods in a comprehensive way, we used both

simulated data and a variety of published and previously unpublished experimental data from mouse and yeast. The mouse data sets include gene expression data from four tissues of recombinant inbred (RI) BXD mouse strains: regulatory T-cell (H. Chen, RA, and KS, unpublished data), lung (RA, L. Lu, R. Williams, and KS, unpublished data), hematopoietic stem cells [28], and hippocampus [29]. The yeast data were taken from [30]. Further details are available in the methods section.

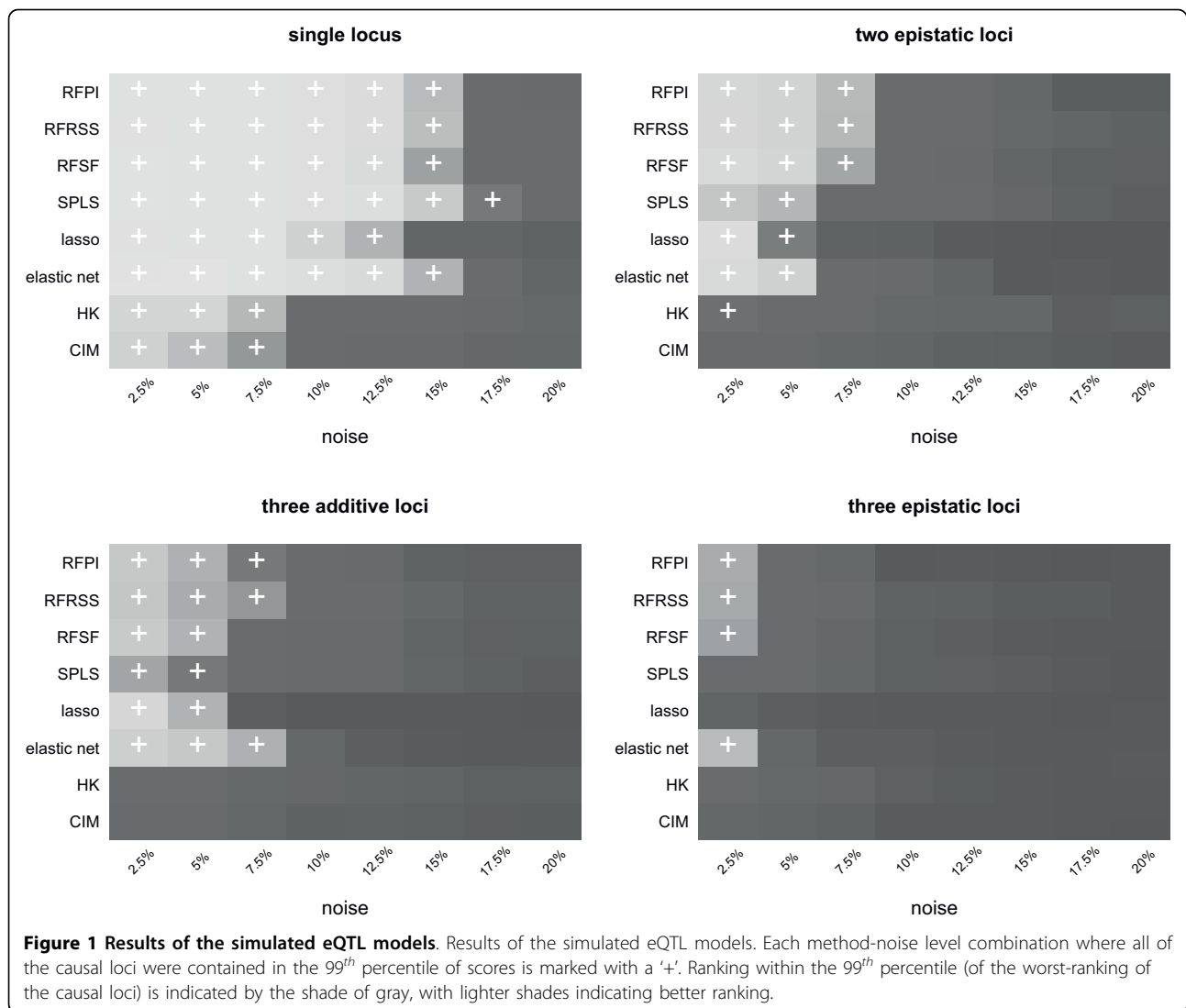
We note here that one of the goals of this comparison is to determine how susceptible each method is to the effects of linkage disequilibrium. In light of this goal we used all genotype data as-is, without prefiltering or fusing markers, or assigning surrogate eQTL post-hoc. This enables a straightforward comparison across all mapping methods.

Simulations

We first set out to examine the performance of each method when the underlying model generating the data was known completely. We used the actual BXD genotypes and generated traits based on four models: single causal locus, two epistatic causal loci, three additive causal loci, and three epistatic loci. These configurations were sufficient to clearly distinguish the performance of the methods. Further details of the construction of the simulated data are given in the methods section. The goal of this investigation was to determine how well each method performed at placing *all* causal loci in the 99th percentile of scores, over a range of increasing Gaussian noise in the trait. The results are given in Figure 1. In the single locus scenario, the performance gap between the newer multi-locus methods (RF, SPLS, the lasso, and the elastic net) and the legacy methods (HK and CIM) is quite apparent. In the single locus case, HK and CIM are unable to correctly identify causal loci in traits with more than 7.5% noise, and fail almost completely at pinpointing causal loci in the more complex two and three locus models. The elastic net and RF deliver comparable performance in the more complex models, with RF performing better in epistatic scenarios and the elastic net performing slightly better in the three-locus additive model. It should be noted that while SPLS, the lasso, and the elastic net do not explicitly search for interactions, they may still find loci participating in epistasis due to small but detectable marginal effects of the interaction.

cis-eQTL counts

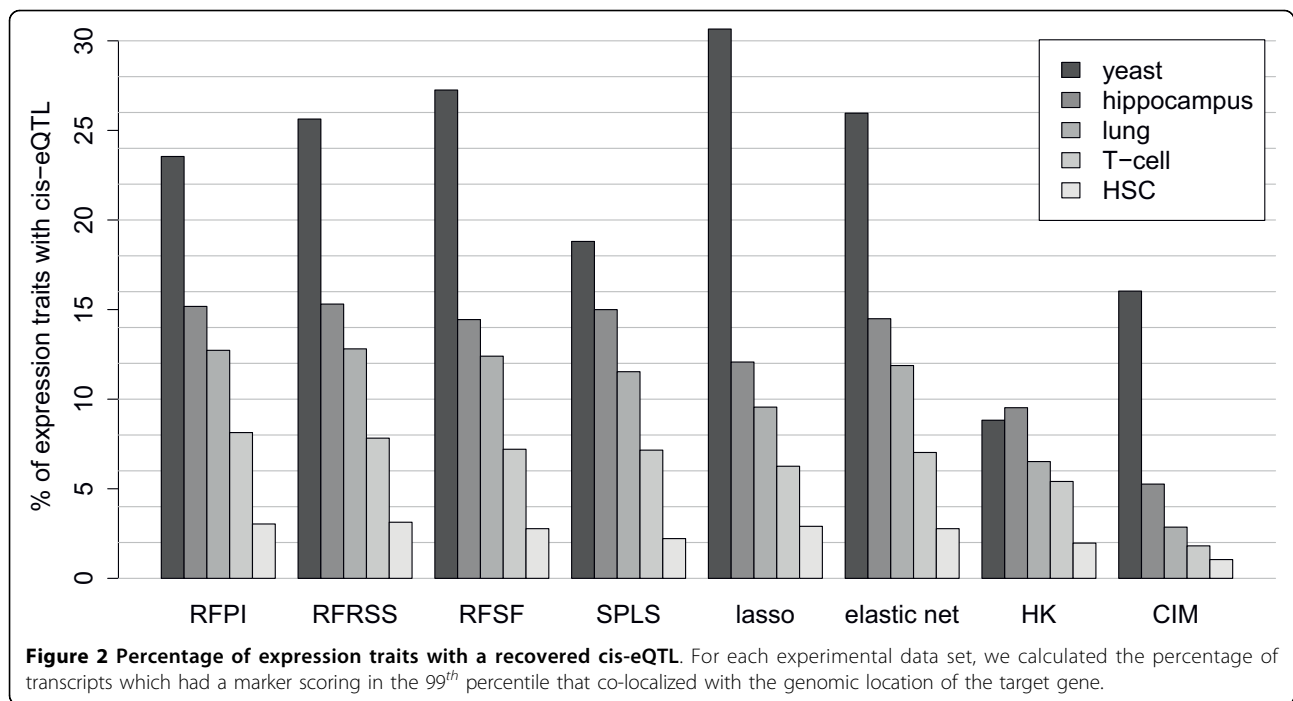
A "back of the envelope" approach for gauging the practical performance of a mapping method is the proportion of *cis*-eQTL found among all target transcripts in experimental data. Since promoter regions are often polymorphic, one would expect under optimal



conditions to be able to recover an eQTL at the genomic location of many of the examined target transcripts. In this sense, the “external information” used in the benchmark is the knowledge of the genomic location of the gene – which, when compared to QTL in general, is information unique to eQTL. The results of this assessment are shown in Figure 2. Taken individually, no single method dominated the others. However, the legacy methods (HK and CIM) again showed poor performance when compared to their more modern counterparts. A relationship between study size and proportion of recovered *cis*-eQTL is also uncovered, with the larger studies (yeast, mouse hippocampus and mouse lung with 114, 67, and 44 observations, respectively) generally yielding higher proportions of *cis*-eQTL than smaller studies (mouse regulatory T-cell and mouse hematopoietic stem cell with 33 and 22 observations, respectively).

KEGG enrichment

We used the pathway information available in the KEGG database to establish relationships between target genes and potential regulators. KEGG was chosen because of its position as a standard in pathway information and because it is generally a better reflection of the molecular relationships between genes (compared to GO for instance). However, in principle other sources of pathway information could be used. One would not expect to recover an entire pathway in every eQTL map, but on a large scale there should be some overlap between the eQTL and the relationships contained in KEGG. We assert that methods that show higher agreement with the information in KEGG are more desirable for eQTL mapping. We formalize this by assessing the enrichment of high-scoring eQTL for loci near genes known to participate in the same pathways as the gene whose expression trait is being mapped. A graphical



depiction of this idea is given in Figure 3 and further details on the enrichment test are given in the methods section.

We tested pathway enrichment in yeast and mouse eQTL separately. For yeast, we included an additional enrichment test, which connected target genes not to pathways in which they participate, but to pathways in which the target's known transcription factors participate. We used the distributional properties of the enrichment P values to compare the eQTL mapping methods, with results for the yeast data shown in Figure 4. It should be noted that HK did not deviate significantly from the uniform distribution in either the pathway member or the TF-centric enrichment tests ($P = 0.72$ and $P = 0.07$, respectively, by the Kolmogorov-Smirnov test). In contrast, RFSF showed superior performance on the yeast data ($P = 1.56 \times 10^{-133}$ and $P < 10^{-324}$ for the pathway member and TF-centric KEGG enrichment tests, respectively).

The mouse data showed more modest enrichment across all tissues and with all methods, suggesting perhaps that larger studies are needed to better recover the complex regulatory systems present in higher eukaryotes (Figure 5). All methods yielded significant deviation from the uniform distribution in each tissue ($P < 0.05$ by the KS test). Again, RFSF yielded the greatest degree of enrichment in all tissues. SPLS, the lasso, and the elastic net produce sparse models, which means that not all loci are assigned a coefficient as a score. This had the effect that for a small minority of expression traits,

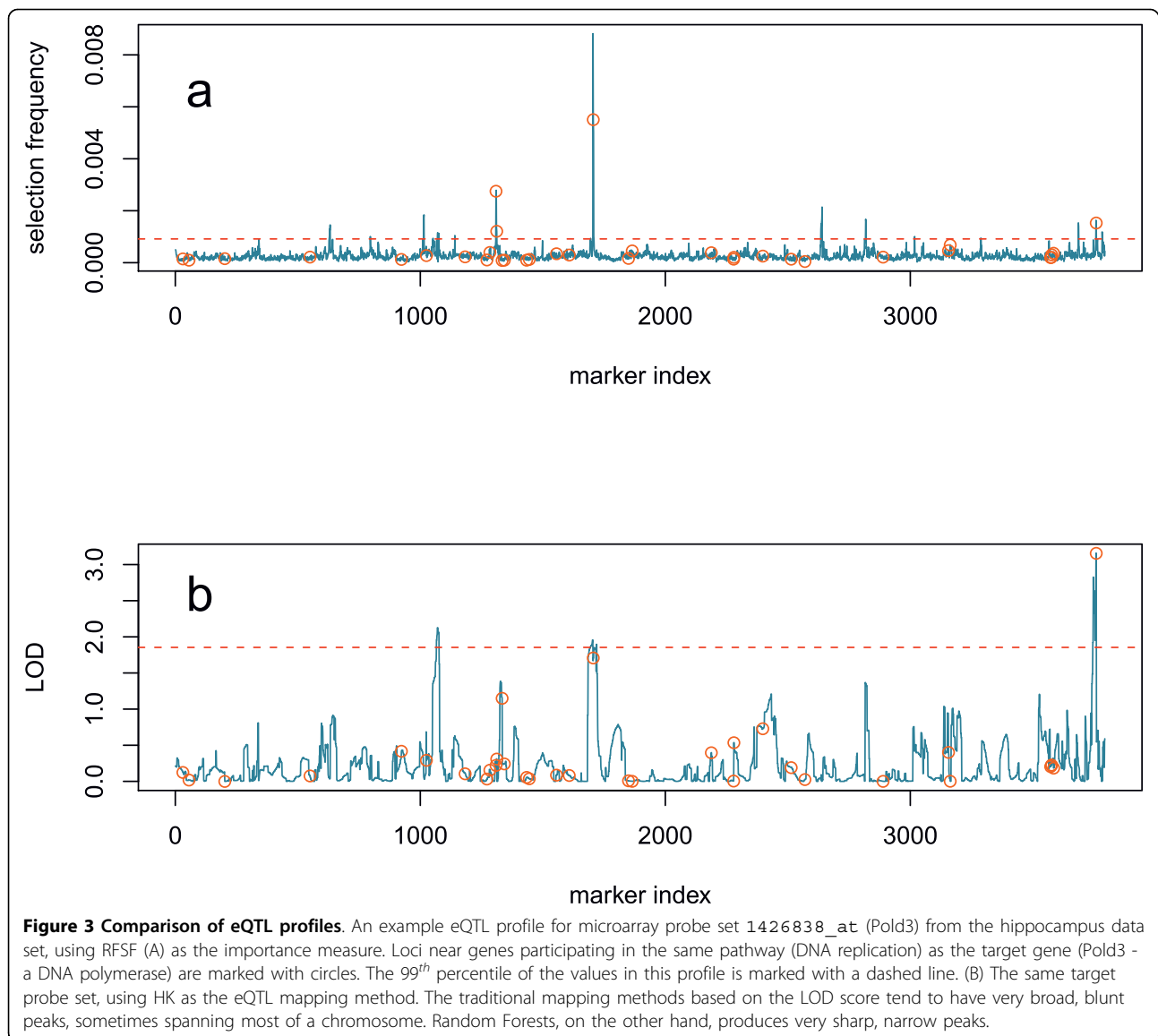
the 99th percentile of scores contained a small number of loci with scores of 0. We examined whether this effect put these sparse methods at a disadvantage for the enrichment tests. We found no systematic relationship between enrichment P value and the number of 0 scores in the 99th percentile.

Mutant expression change enrichment

Finally, we combined data from two systematic loss of function studies [31,32] to see which method produced eQTL that agreed most with the mutant data.

In this test, we collected the maximum absolute expression change observed for each target gene when genes co-localized with eQTL in the 99th percentile are mutated. These values were aggregated over all target genes, forming a distribution for each eQTL mapping method. We compared these distributions to a null distribution (see methods for details) via the Kolmogorov-Smirnov test. We assert that the method that yields eQTL that are enriched for large changes in expression in the mutant study is the most desirable method.

All methods produced score distributions that deviated significantly from the null distribution, suggesting that there is indeed consistency between the yeast eQTL data and independent mutant data. Although all methods showed significant deviation from the null, the magnitude of enrichment varied widely (Figure 6). RFSF showed the most significant enrichment, followed closely by HK.



Discussion

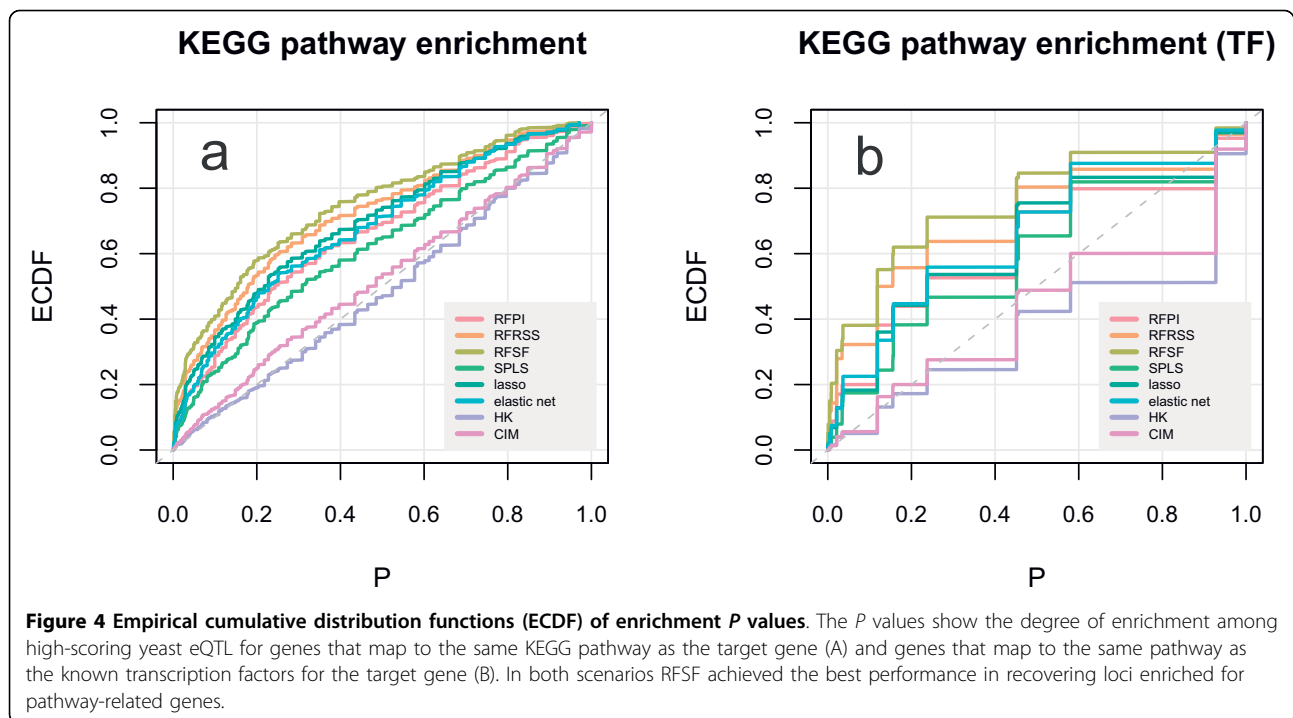
High-throughput data make functional benchmarking of eQTL mapping methods possible

Augmenting eQTL with independent information has been done previously to strengthen hypotheses suggested by the eQTL data [33-36]. Although these applications demonstrate a certain degree of correspondence between eQTL data and external data sources, and imply that such correspondence is desirable in an eQTL mapping method, no benchmarks based on the systematic recovery of biological information have been proposed and applied to a wide variety of mapping methods and data sets. Validating the performance of mapping methods is important not only for those whose analysis ends with an eQTL map, but also for more sophisticated algorithms such as Lirnet [37] and

Geronemo [38] which build on top of basic mapping concepts. Our analysis, combined with previously cited works that integrate eQTL with other data, show that there is indeed agreement among eQTL and data from different sources. Maximizing this agreement should be a core objective of future mapping techniques. We hope that this approach to benchmarking, in addition to traditional simulated benchmarks, will help practitioners find the appropriate method now, and lead to the development of better mapping methods in the future.

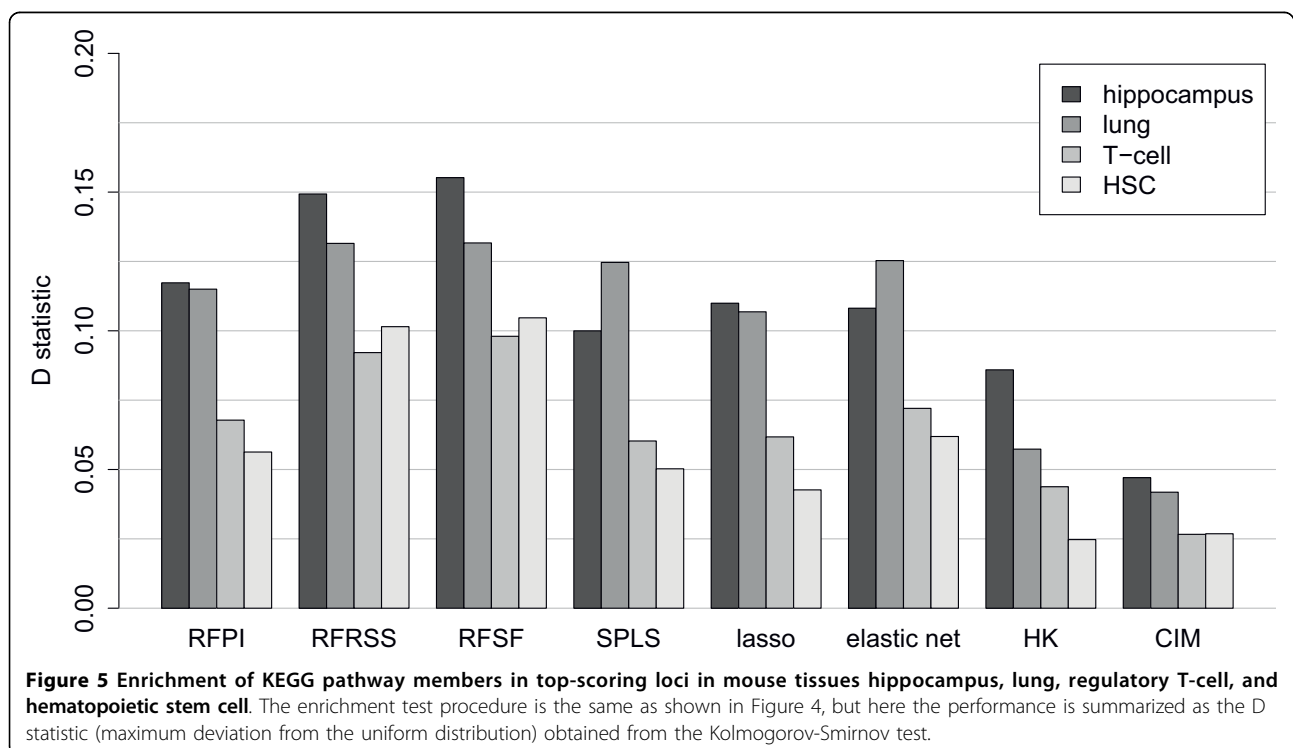
Multi-locus eQTL mapping methods outperform legacy methods

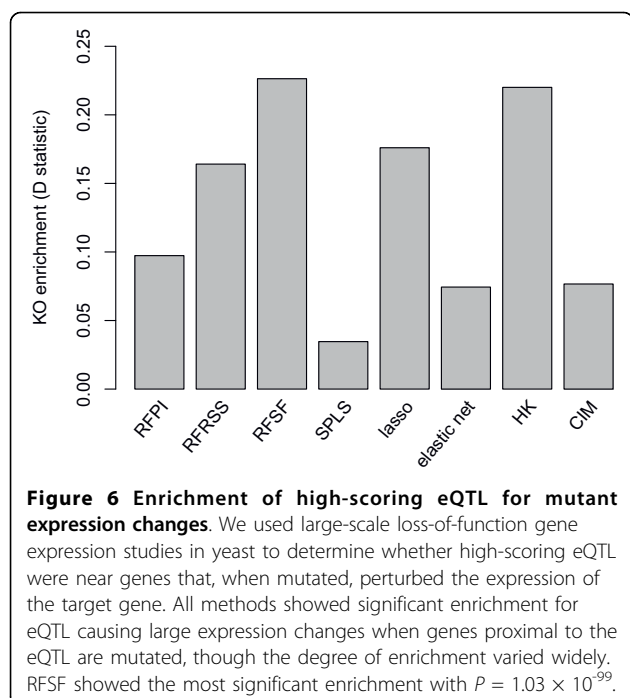
With few exceptions, the legacy methods – HK and CIM – stood out as the poor performers, particularly in the simulations, *cis*-eQTL proportions, and enrichment



for KEGG pathway relationships. In preliminary analyses, we found related univariate mapping methods such as EM interval mapping [21] and ANOVA to have performance almost indistinguishable from HK (data not shown). This observation is important because even

at the time of this writing there are still eQTL papers being published that use legacy mapping methods for their analysis [39-42], ostensibly because the more modern methods are not as accessible. In light of our results, we expect that these studies have not exploited the full





potential of the collected data. This represents a challenge for the computational community of working to promote not just the development, but also the adoption of these more advanced methods.

There is a fundamental difference in how the legacy linear methods (HK, CIM) and the multi-locus linear methods (SPLS, lasso, elastic net) score loci. The univariate mapping methods rely on a LOD score (or a P value in the case of one-way ANOVA) that expresses the significance of the estimated correlation between a single marker and the trait, resulting in thousands of individual modeling attempts per expression trait. The multi-locus methods, in contrast, assign coefficients to multiple loci in a single final model. These coefficients are then used as locus scores. The disparity in performance between the two classes of methods is likely a result of scoring by contribution to the model (multi-locus approach), rather than scoring by significance (univariate approach).

RF offers a third paradigm for scoring that is conceptually similar to the coefficient approach of the multi-locus linear methods, though distinct in implementation. Each of the three importance measures derived from RF measures a locus' average contribution in an ensemble of models. This differs from the coefficient approach in that it is a summary of multiple models, each including multiple loci, rather than a summary of a single model including multiple loci. Additionally, the multi-locus linear methods do not implicitly allow for the inclusion of epistatic interactions in the locus scoring process, while RF does.

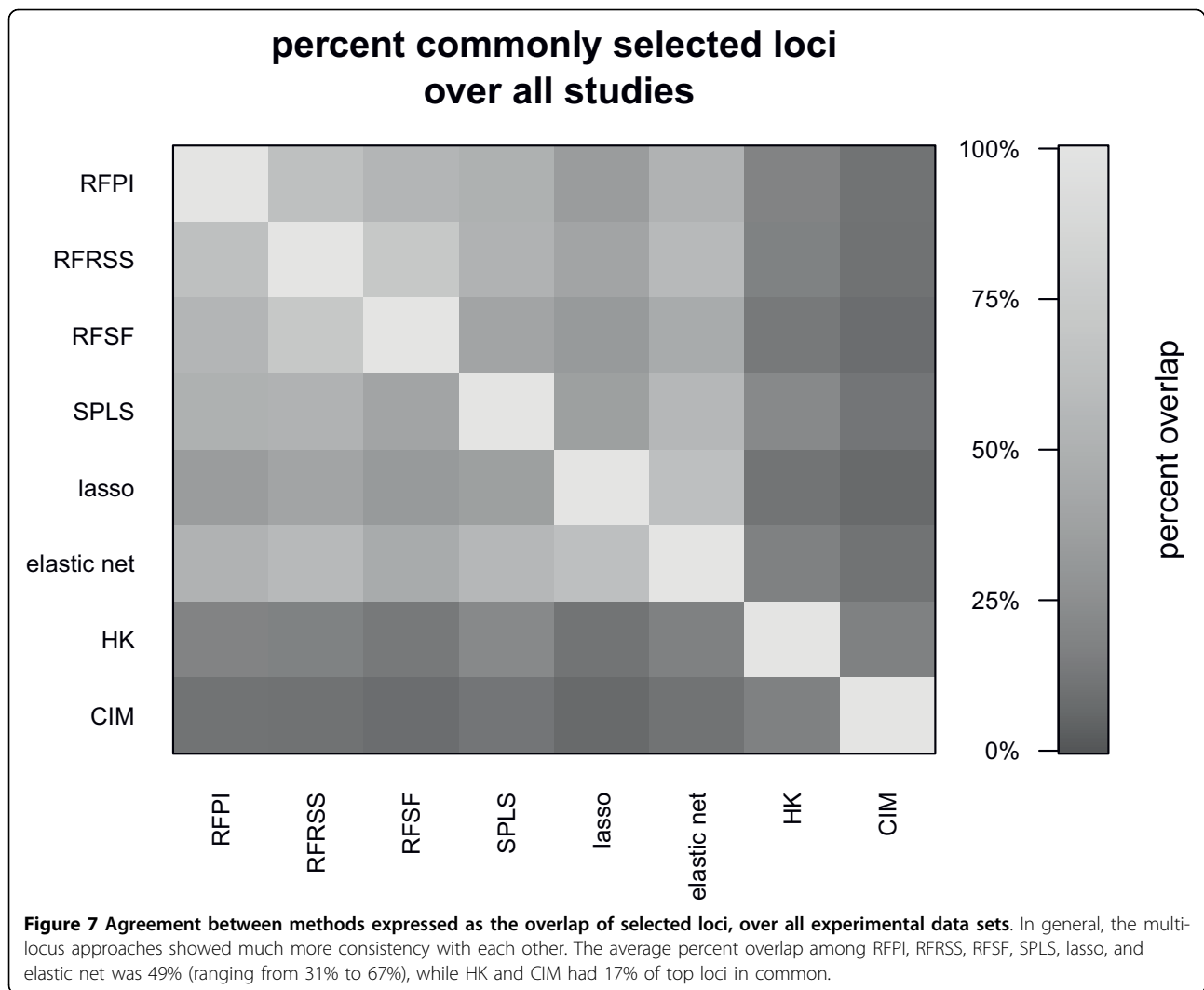
It should be noted that the benchmarking process described in this work did not focus on the methods' abilities for statistical inference, that is, determining whether a locus *significantly* explains an expression trait. Instead, our benchmarks focused on which methods prioritized the loci with the greatest degree of effectiveness over a large panel of data. If statistical inference is desired, appropriate permutation of the data can be performed to obtain a null distribution of scores for the chosen method, which can then be used to assess significance of the scores.

We evaluated all experimental data sets and compared the loci that each method scored in the 99th percentile. In general, the multi-locus approaches showed agreement amongst themselves, with an average 49% overlap. Figure 7 highlights the lack of consistency between the legacy methods and the multi-locus methods, and amongst themselves.

Random Forests selection frequency maps the most biologically consistent eQTL

Random Forests (RF) [23] is a classification and regression algorithm based on fitting an ensemble of trees. When mapping eQTL, RF fits decision trees by using markers as predictor variables, i.e., each node in a tree corresponds to a split of the population based on the genotype at the selected marker. By combining an ensemble of many diverse decision trees, RF guards against overfitting and also provides several measures of predictor variable importance. In this work, these measures of variable importance are used to map eQTL.

Although multi-locus methods in general outperformed the legacy methods HK and CIM, RFSF showed the most consistent performance overall. In the simulations and *cis*-eQTL proportion test it was among the best, and in the KEGG and mutant enrichment tests it outperformed the competitors. This finding is somewhat surprising because RFSF is virtually ignored as a variable importance measure in most applications of RF, including QTL and GWAS [14,15,17,26,27]. Avoiding RFSF may have several explanations. For instance, it has been shown previously that RFSF can be biased. This bias manifests itself in the case of continuous or categorical predictors that vary widely in their scales or number of categories [43]. This is typically not an issue in the case of genotype data, where all predictors are categorical with the same number of categories. However, RFSF can also be biased when there is a significant degree of correlation between predictors, which is the case with genotype data. Under these conditions, RFSF preferentially selects variables (markers) with low correlation to other variables; markers in linkage disequilibrium are under-selected. In order to estimate and account for this bias, we add or subtract the deviation from the mean

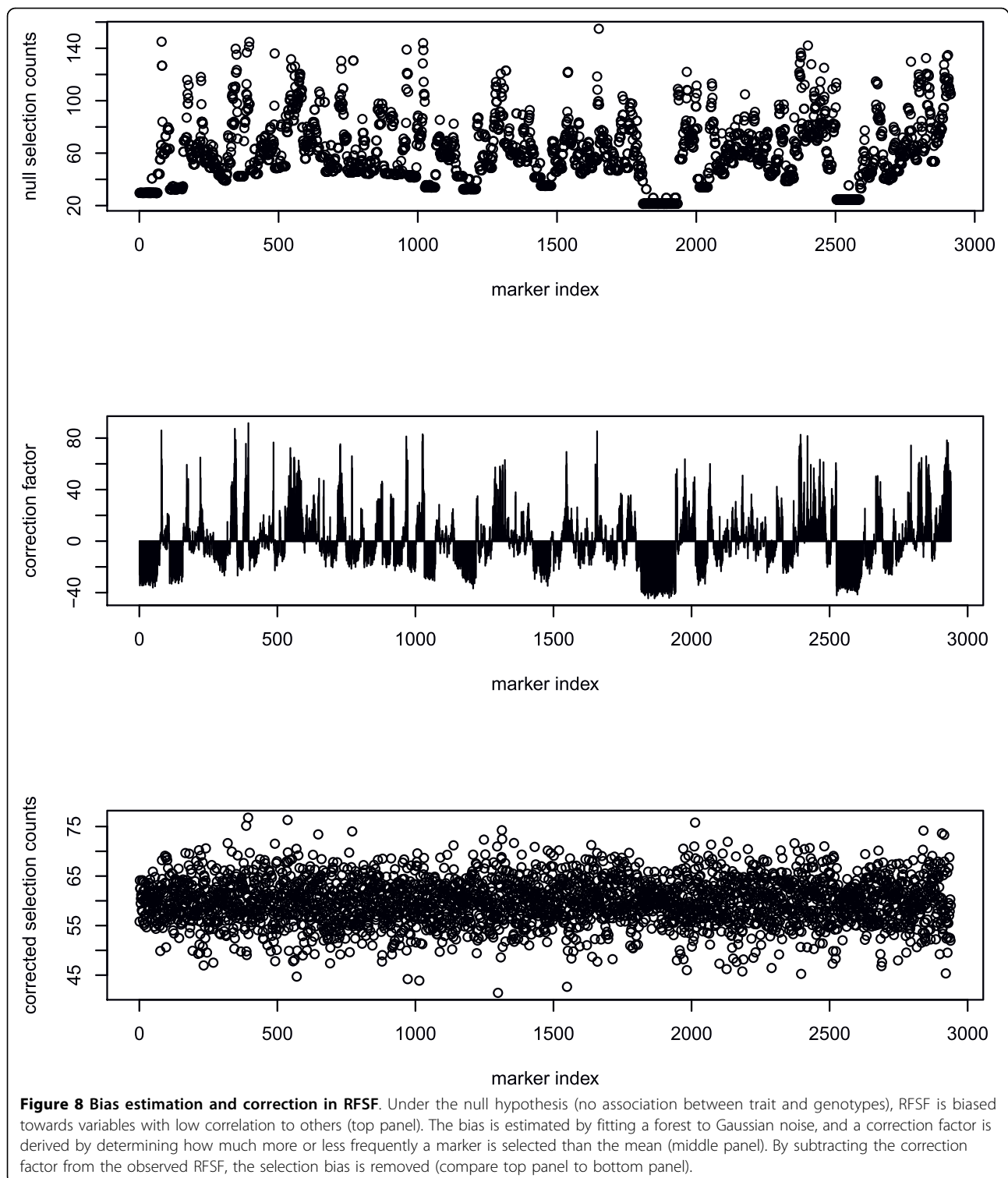


selection frequency observed under the null hypothesis (no association between trait and genotype data). See methods and Figure 8 for details.

We decided to investigate further the potential reasons why RFSF performed better than the more typically used RFPI or RFRSS. We hypothesized that perhaps RFSF picked up on smaller effects near the leaves of the trees, i.e. it is able to detect loci with very subtle effects on the trait. To demonstrate this, we use the largest data set (yeast) and grew several RFs with different characteristic tree depths. We then tested these forests with the *cis*-eQTL proportion test and the KEGG enrichment test (see methods for details). We found that increasing the depth of the trees had a modest effect on the performance of RFPI and RFRSS, with an increase in percentage of *cis*-eQTL from 22.4% to 23.5% and 24.1% to 25.6%, respectively, and an increase in D statistic (for the KEGG enrichment test) from 0.186 to 0.225 and 0.241 to 0.318, respectively. Conversely, RFSF benefited more

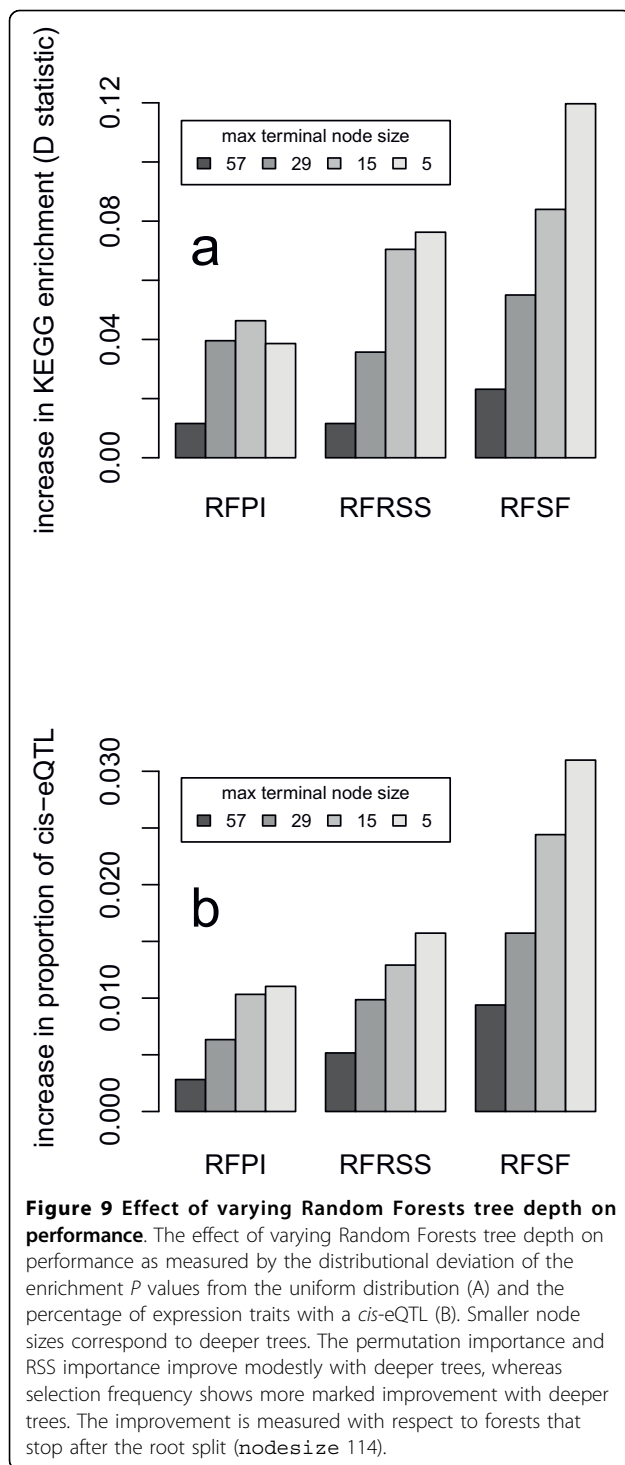
from the deeper forests, with an increase in percentage of *cis*-eQTL from 24.3% to 27.4% and an increase in D statistic (for the KEGG enrichment test) from 0.241 to 0.361 (Figure 9). In addition, we found that agreement with the linear methods (SPLS, lasso, elastic net, HK, and CIM) was at its highest when the tree growth was stopped early; similarity decreased with increasing tree depth. This effect was more pronounced for RFSF than for the other RF importance measures, which further suggests that the effects found near the leaves of the trees are connected to RFSF's superior performance (Figure 10).

To further explore this idea, we performed simulations where the expression trait was a function of eight loci, two with strong effects, and six with small effects. As expected, the loci with the stronger effects were used in splits closer to the root node. The causal loci with weaker effects were used to split closer to the leaves. In these simulations, RFSF scored the weak causal loci in the 99th percentile 18.3% of the time, while RFPI scored



the same loci in the 99th percentile only 10% of the time. These simulations also showed that RFPI is tightly coupled to a variable's proximity to the root node, while RFSF can give high scores even if the variable is not used close to the root node.

From these investigations we conclude that RFPI and RFRSS both essentially determine variable importance near the roots of the trees, and that biologically important splits further down the tree are not adequately reflected in the overall importance scores. RFSF on the



other hand, recovers more biologically meaningful predictor variables (loci) when trees are grown deep, suggesting that even splits far down the tree can be reflected in this importance measure. Epistatic effects are an example of where this phenomenon is important – often genetic interactions are weak and only present in a

subset of the population. Such conditional effects are likely to manifest themselves deeper in the trees. RFSF is an attractive measure in these situations.

Because of its demonstrated performance advantages in finding biologically relevant loci, its ability to implicitly consider epistatic interactions, as well as its straightforward and readily available implementation, we recommend using Random Forests for eQTL mapping. We have prepared a short tutorial and example R code demonstrating mapping eQTL with the bias-corrected selection frequency at <http://cellnet.biotech.tu-dresden.de/RFSF>.

Marker density and analysis strategy

In this work we examined studies with genotype data in the range of thousands of markers. With the advent of next-generation sequencing and other ultra high-throughput methods, we expect to see more and more studies with hundreds of thousands, millions, or even tens of millions of SNPs. We wish to put the presented work in context by drawing a distinction between filtering methods, mapping methods, and explicit models (Figure 11).

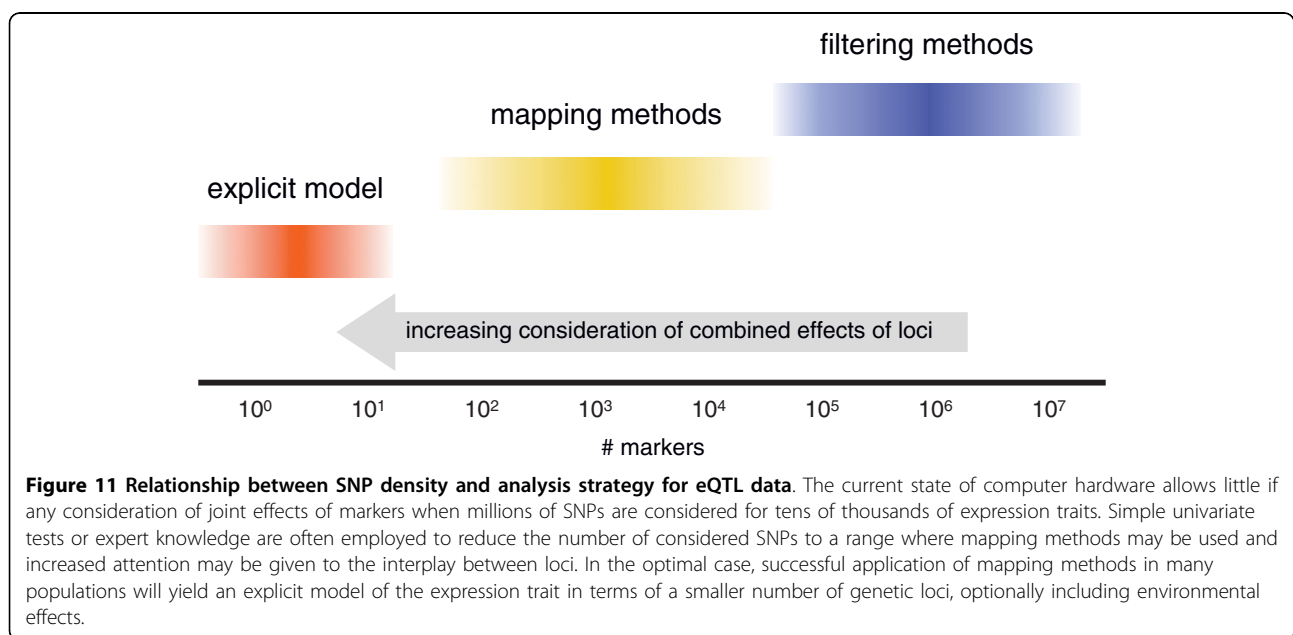
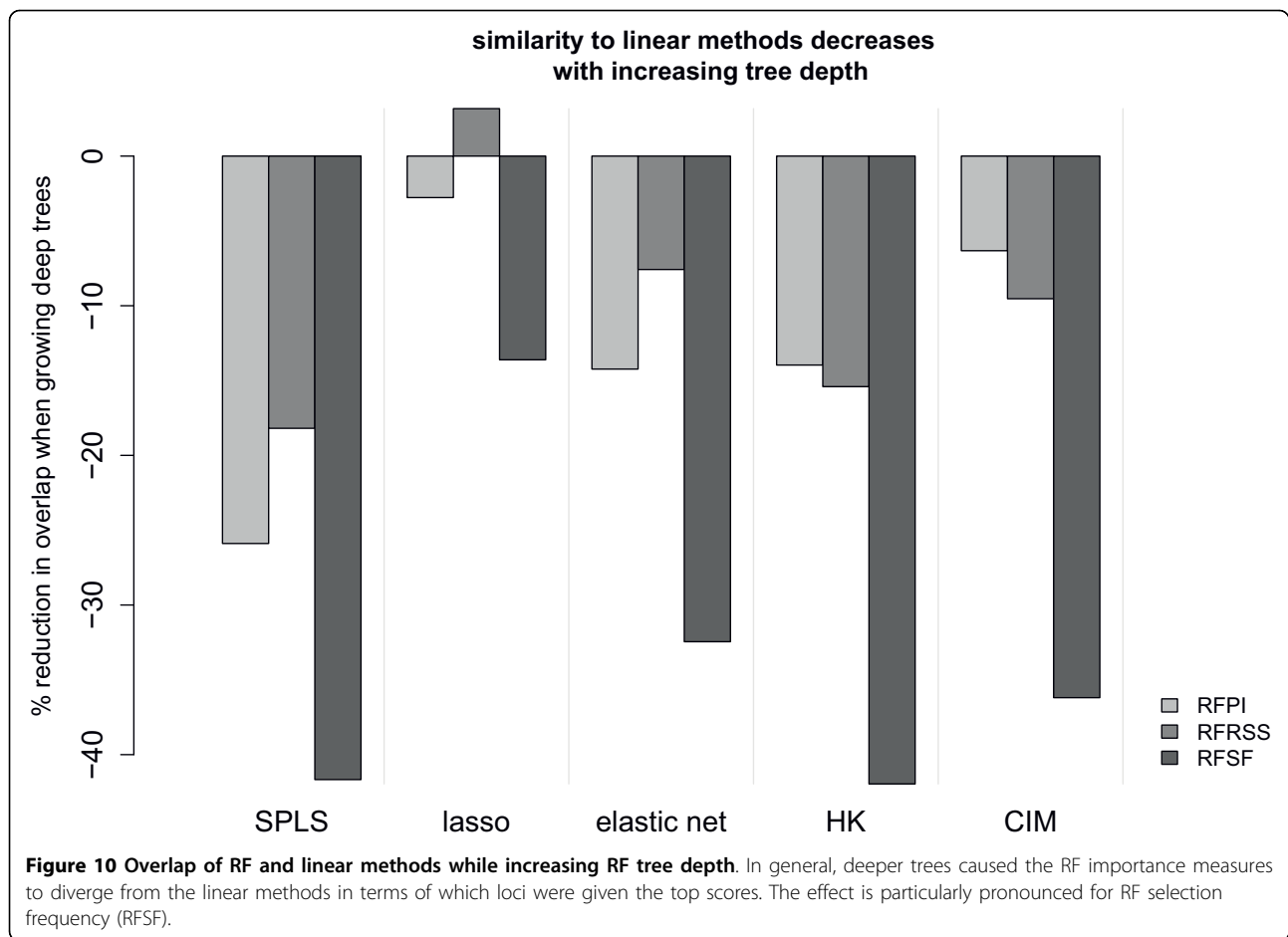
The state of computer hardware at the time of this writing makes the multi-locus methods presented here impractical for exhaustive evaluations of data sets with millions of SNPs and tens of thousands of expression traits. The current solution to this problem is to filter the SNPs to a more tractable number using univariate tests or expert knowledge [44–46]. Considering the joint effects of markers at this point is generally a fruitless effort, given the astronomical number of potential combinations and the problem of dealing with false positives.

As the number of markers considered falls into the tens of thousands, the problem transitions from filtering to mapping. Mapping is a combination of modeling and feature selection, and the methods we explored in this work address the mapping problem. Here the interplay between loci becomes important for accurately identifying the causal regions that should be included in an explicit model of the trait.

Once causal loci have been identified reliably and the relationships between them have been characterized (additive vs. dominant, epistatic vs. additive, etc.), one can construct a linear model, usually consisting of a handful of terms, that accurately describes the trait as a function of the genetic state of the organism. Such an explicit model, though desirable, is rarely attained.

Implications for related mapping problems

Most of the conclusions from our work have implications beyond eQTL mapping. Ideally, the concept of a knowledge-driven benchmark could be used for any



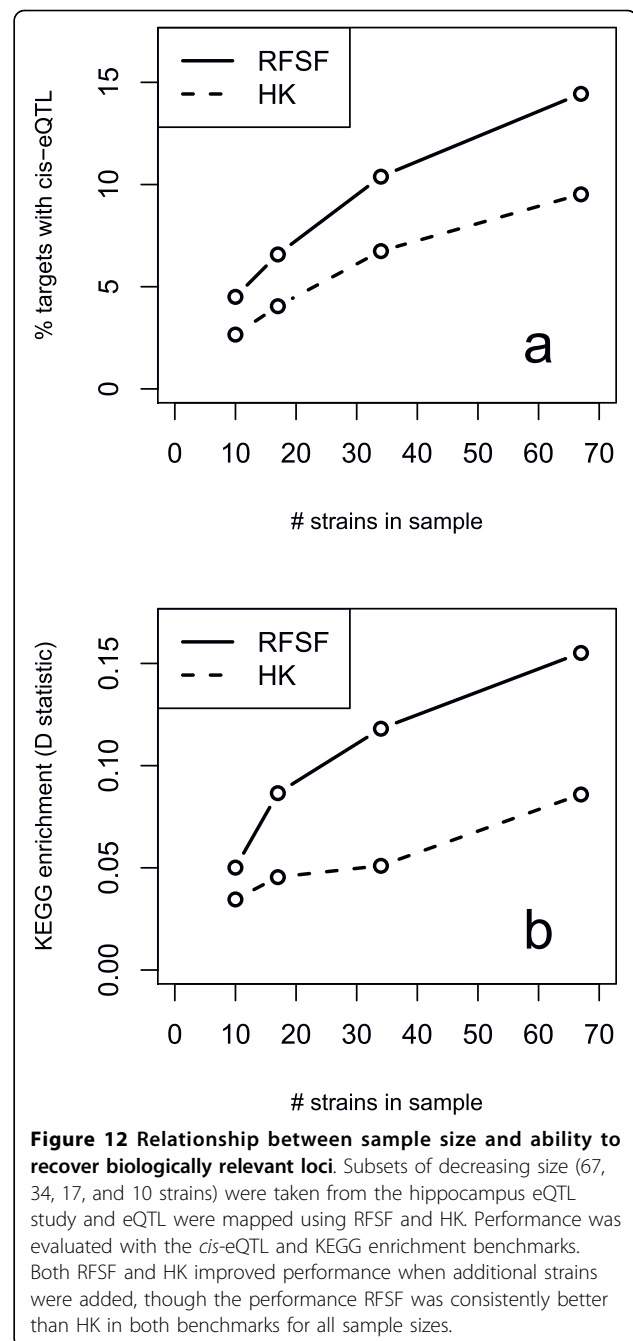
physiological trait, but our approach depends on a fairly detailed knowledge of the molecular mechanisms underlying the mapped trait. Neither our notion of measuring the enrichment of regulator-target gene groups in common pathways, nor our counting of *cis*-eQTL is immediately extendible to physiological traits. Still, taken together, the evidence from this study indicates that QTL mapping - whatever the trait - should be performed using a multi-locus method. Using univariate methods such as HK will lead to severe underexploitation of the data.

Some of the more specific conclusions from our work will need further validation in other organisms and populations. For example, the study populations used here all had roughly a 50/50 distribution of two possible alleles at each marker. Human populations are characterized by very uneven distributions of SNPs, where minor alleles can be extremely rare in a given population. Such a change in the characteristics of the data could influence the ranking of the individual methods. However, such fluctuations in the individual rankings are still unlikely to affect the general conclusion that multi-locus methods produce more informative results than univariate methods, even in GWAS and linkage studies in outbred populations [47-51].

Finally, in this work we observed the expected relationship between study size and power to detect biologically interesting loci. We explored this phenomenon explicitly by taking subsets of decreasing sample size from the hippocampus study, and then comparing two representative methods - here RFSF and HK - using the *cis*-eQTL and KEGG enrichment benchmarks. The results are depicted in Figure 12 and clearly show that while both methods show improvements with additional samples, it is RFSF, the multi-locus method, that shows consistently better performance, regardless of the sample size. This suggests that even in studies with small sample sizes, multi-locus approaches are preferable to single-locus methods.

Conclusions

We have compared modern machine learning and regression eQTL mapping methods with more classical mapping approaches from statistical genetics, and evaluated the methods based on their ability to lead users to loci that are more readily supported by external information. We found that the modern methods, which freely allow the consideration of all loci simultaneously, generally outperform their classical counterparts in this regard. In particular, we found that Random Forests consistently mapped the most promising eQTL. Random Forests bias-corrected selection frequency, a novel importance measure, performed better in these tasks than the established permutation importance and RSS importance.



Methods

eQTL mapping

We used expression data from four eQTL studies in four different tissues in recombinant inbred BXD mouse strains: regulatory T-cell, lung, hematopoietic stem cells [28], and hippocampus [29]. We used only probe sets that mapped unambiguously to Ensembl gene IDs with KEGG annotations [22]. This resulted in a set of 6,121 probe sets for studies using the Affymetrix Mouse 430 2.0 array (lung, regulatory T-cell, and hippocampus) and

3,051 probe sets for the hematopoietic stem cell study, which used the Affymetrix U74Av2 array. Genotype data for the BXD recombinant inbred strains of mice used in these studies consisted of 3,794 markers and was downloaded from the GeneNetwork database [52]. In addition to the mouse data, we used the yeast eQTL study previously published in [30]. After filtering out probes with missing or otherwise ambiguous data, we were left with 4,501 gene expression measurements and 2,914 markers.

Random Forests

We used the reference implementation of Random Forests [53] in R for all mapping discussed in this work. We grew forests with 5,000 trees, the `mtry` parameter was set to the default (one third of the total number of markers) and the node size was also the default of 5, unless otherwise noted. We then extracted unscaled permutation importance measures (RFPI), residual sums of squares importance measures (RFRSS), and selection frequencies (RFSF) from the forests for use as the scores for each marker.

We estimated and accounted for bias in RFSF as follows. Using the actual genotype data as predictors, we fit 500 10-tree forests to independent draws from Gaussian noise. This resulted in 5,000 trees, equal in size to the forests used in this work. We collected the selection frequencies for each marker and subtracted the mean selection frequency to yield a vector of correction factors – one value for each marker. Subtracting this vector of correction factors from the observed selection frequencies (from the observed data) gives bias-corrected selection frequencies (Figure 8). In the context of results in this work, all references to RFSF imply the bias-corrected RFSF, as described here.

Sparse partial least squares

Chun and Keles [12] recently introduced a method of eQTL mapping using sparse partial least squares, which included an R package and a thorough tutorial available online. We used the `spls` R package to map eQTL, performing cross-validation on every target to determine the optimal parameters for each fit. `eta`, the thresholding value, was allowed to vary between 0.3 and 0.7, to prevent both overfitting and a model that was too sparse to score multiple loci. The number of hidden components was allowed to vary from 1 to 5. A final fit was performed with the optimal parameters, and the absolute value of the coefficients was used as the score for each marker.

The lasso

The lasso [24], a regression shrinkage method, has previously been applied to QTL mapping [54], but to our knowledge has never been tested against competing mapping methods in the context of an eQTL study. For this work, we used the lasso as implemented in the

`elasticnet` package for R. The lasso is a special case of the elastic net with lambda equal to (or very near) 0. For each target gene examined, we took the absolute value of the lasso coefficients for a fit performed with the `s` parameter determined by 10-fold cross-validation, with an imposed minimum of 0.5. These coefficients were used as the score for each marker.

The elastic net

The use of the elastic net [25] was the same as above for the lasso, except that lambda was set to 1. We found this value of lambda to be optimal after testing a sample of target genes over a range of lambda values (0.5,1,10,100).

Haley-Knott regression

We used the implementation of Haley-Knott regression [20] available in the `qtl` package for R. LOD scores were calculated at the marker locations.

Composite interval mapping

To perform composite interval mapping [19] we used the implementation in the `qtl` package for R, with the `method` argument set to “EM”, and all other arguments set to their default. LOD scores were calculated at the marker locations.

Simulations

To simulate eQTL with known underlying models, we used the full BXD genotype matrix, available from the GeneNetwork [52]. This matrix consists of 89 strains and 3,794 markers. Using this genotype data, we randomly selected one, two, or three markers (depending on the model to be simulated), and then simulated a trait by using a linear combination of the markers directly, or of logical operations on the markers (in the case of epistasis). All traits started with a baseline value of 9, before adding in the genetic effects. Genetic effects were added as follows: in the single locus model, a single marker was selected at random, and its vector of genotypes (where 1 = BB and 0 = DD) was multiplied by a coefficient, in this case 1. For the two-locus epistatic model, two marker vectors were selected at random, with each being multiplied by 0.25 and then summed. The epistatic component was added by applying the AND logical operation to the genotype vectors (where a 1 is a TRUE and a 0 is a FALSE) and then multiplying the result by a coefficient, in this case 1, and then adding to the additive component. Three locus additive and epistatic traits were constructed in a similar fashion. Gaussian noise with mean 0 was then added to the traits, over 8 levels of increasing standard deviation, which corresponded to 2.5, 5, 7.5, 10, 12.5, 15, 17.5, and 20% of the trait mean.

Each model type (i.e. single locus, two locus epistatic, etc.) was simulated independently 50 times, and each mapping method was applied to the same data. For each

simulation and for each mapping method, the maximum (i.e. worst) rank among the set of causal markers was recorded in each noise level. The median of these values (over the 50 simulations) was used to reflect the performance of a given mapping method over the increasing levels of noise. Lower values represent the ability of a method to assign high scores to *all* causal loci.

cis-eQTL counts

Performance based on the proportion of recovered *cis*-eQTL was assessed by counting the number of expression traits where a marker within 500 kb (for mouse) or 50 kb (for yeast) of the midpoint of the target gene's genomic location had a score in the 99th percentile of the scores for the respective target gene. These cutoffs, though arbitrary, reflect the difference in complexity between the yeast and mouse genomes - the conclusions drawn from the benchmark are not heavily influenced by this choice. This number was then divided by the number of total expression traits examined for the respective data set.

KEGG enrichment

Each expression trait we tested mapped to at least one KEGG pathway, and each gene found in the KEGG pathway was mapped to the nearest marker. If no marker fell within 5 Mb of a gene, the gene was omitted. For each expression trait, the markers having scores in the 99th percentile were selected for the enrichment test. The hypergeometric test was used to test this set for the enrichment of markers mapping to genes participating in the same KEGG pathway as the target gene. If multiple pathways existed for any expression trait, all were tested and the minimum *P* value was used as the representative *P* value.

In the case of the yeast eQTL data, we additionally assessed enrichment of pathways in which transcription factors binding to the target gene participate. As a basis for mapping transcription factors to their targets, we used [55]. We did not attempt this test with the mouse data because of the lack of dense and reliable TF-target data for mouse.

Since in this test even randomly selected markers yield *P* values that deviate somewhat from the uniform distribution, we calculated an empirical null distribution of *P* values. To construct this distribution, we assigned scores to the markers, drawn randomly from a Gaussian distribution with mean 0 and standard deviation of 1. We then took the markers in the 99th percentile and performed the proposed enrichment test. This was performed for an equivalent number of expression traits contained in the actual data sets. The actual enrichment *P* values were corrected against this empirical null distribution of enrichment *P* values.

We plotted the empirical cumulative distribution function (ECDF) of the corrected enrichment *P* values for each method. As a summary measure for each method's deviation from the uniform distribution, we used the D-statistic as given by the Kolmogorov-Smirnov test. The test was one-sided with the alternative hypothesis that the observed cumulative distribution function accumulated faster than the reference (i.e. uniform) distribution.

Mutant expression change enrichment

Systematic loss of function data in yeast [31,32] was used to assess which eQTL mapping method tended to agree most with the regulatory relationships suggested by experimentally deactivating upstream regulators. We mapped each repressed gene to its nearest marker. Then, for each expression trait from the yeast eQTL study, we looked at markers in the 99th percentile of scores for that target. For markers mapping to experimentally repressed regulator genes, we collected the maximum absolute log₂ expression ratio (repressed expression divided by wild-type expression) for the appropriate target gene, aggregating them over the whole set of mapped expression traits. We then compared the distribution of the selected maximum absolute log₂ ratios generated by each eQTL mapping method by the Kolmogorov-Smirnov (KS) test, collecting the associated *P* value and D statistic. As a reference distribution in the KS test, a null distribution was constructed by a similar aggregation of maximum absolute fold changes, only with the association between scores and markers randomized for each target gene. The test was one-sided with the alternative hypothesis that the observed cumulative distribution function accumulated slower than the reference distribution. Distributions with a tendency toward higher scores and deviating significantly from the reference distribution suggest an agreement between the eQTL and loss-of-function studies.

Variation of tree depth

To assess the impact of tree depth on each RF importance measure, we used the yeast eQTL data and recomputed eQTL maps for all expression traits, varying the `nodesize` argument to 5, 15, 29, 57, and 114.

The `nodesize` argument dictates whether or not a node may be split - if the number of observations in the node under consideration is greater than `nodesize`, the node may be split. Otherwise the node is not split and is marked as a terminal node. The default value of `nodesize` is 5 - this is the value used in the main body of the study. By selecting a `nodesize` of 114 (the number of samples in the yeast study), we ensure that splitting stops after the first split. The other values are intermediate steps, each about half the size of the last. We then assessed the improvement in the

enrichment of KEGG pathway members and proportion of cis-eQTL identified when growing the trees deeper, using the forest with nodesize 114 as the baseline.

Acknowledgements

We gratefully thank Rupert Overall and Gerd Kempermann (both CRT, Dresden), as well as Rob Williams, University of Tennessee, Memphis, USA, for providing us eQTL data and for help with the data pre-processing. We thank the Center for High-Performance Computing, TU Dresden for providing computational resources. We also thank the anonymous reviewers for their constructive comments and suggestions. This work was funded by the Klaus Tschira Foundation, the European Commission FP7 (Grant ID 223539), and the Helmholtz Alliance on Systems Biology.

Author details

¹Cellular Networks and Systems Biology, Biotechnology Center - TU Dresden, Dresden, Germany. ²Helmholtz Center for Infection Research, Braunschweig, Germany.

Authors' contributions

JM and AB conceived the benchmarks and analysis. JM performed the analysis and drafted the manuscript. RA assisted in the analysis. KS conceived and carried out the lung and T-cell studies. All authors reviewed the final manuscript.

Received: 13 April 2010 Accepted: 17 September 2010

Published: 17 September 2010

References

1. Rockman MV, Kruglyak L: **Genetics of global gene expression.** *Nat Rev Genet* 2006, **7**(11):862-72.
2. Brem RB, Storey JD, Whittle J, Kruglyak L: **Genetic interactions between polymorphisms that affect gene expression in yeast.** *Nature* 2005, **436**(7051):701-3.
3. Kempermann G, Chesler EJ, Lu L, Williams RW, Gage FH: **Natural variation and genetic covariance in adult hippocampal neurogenesis.** *Proc Natl Acad Sci USA* 2006, **103**(3):780-5.
4. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, Fischer J, Maatz H, Kren V, Pravenec M, Hubner N, Aitman TJ: **Heritability and tissue specificity of expression quantitative trait loci.** *PLoS Genet* 2006, **2**(10):e172.
5. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK: **High-resolution mapping of expression-QTLs yields insight into human gene regulation.** *PLoS Genet* 2008, **4**(10):e1000214.
6. Druka A, Potokina E, Luo Z, Bonar N, Druka I, Zhang L, Marshall DF, Steffenson BJ, Close TJ, Wise RP, Kleinjohs A, Williams RW, Kearsey MJ, Waugh R: **Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* in barley.** *Theor Appl Genet* 2008, **117**(2):261-72.
7. Sieberts SK, Schadt EE: **Moving toward a system genetics view of disease.** *Mamm Genome* 2007, **18**(67):389-401.
8. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE: **Variations in DNA elucidate molecular networks that cause disease.** *Nature* 2008, **452**(7186):429-35.
9. Schadt EE, Lum PY: **Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes.** *J Lipid Res* 2006, **47**(12):2601-13.
10. Michaelson JJ, Loguerio S, Beyer A: **Detection and interpretation of expression quantitative trait loci (eQTL).** *Methods* 2009, **48**(3):265-76.
11. Broman KW, Speed TP: **A model selection approach for the identification of quantitative trait loci in experimental crosses.** *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2002, **64**(4):641.
12. Chun H, Keles S: **Expression quantitative trait loci mapping with multivariate sparse partial least squares regression.** *Genetics* 2009, **182**(1):79-90.
13. Huang Y, Wuchty S, Ferdig MT, Przytycka TM: **Graph theoretical approach to study eQTL: a case study of *Plasmodium falciparum*.** *Bioinformatics* 2009, **25**(12):115-20.
14. Lee SSF, Sun L, Kustra R, Bull SB: **EM-random forest and new measures of variable importance for multi-locus quantitative trait linkage analysis.** *Bioinformatics* 2008, **24**(14):1603-10.
15. Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Van Eerdewegh P: **Identifying SNPs predictive of phenotype using random forests.** *Genet Epidemiol* 2005, **28**(2):171-82.
16. Benjamini Y, Yekutieli D: **Quantitative trait Loci analysis using the false discovery rate.** *Genetics* 2005, **171**(2):783-90.
17. Bureau A, Dupuis J, Hayward B, Falls K, Van Eerdewegh P: **Mapping complex traits using Random Forests.** *BMC Genet* 2003, **4**(Suppl 1):S64.
18. Jiang C, Zeng ZB: **Multiple trait analysis of genetic mapping for quantitative trait loci.** *Genetics* 1995, **140**(3):1111-27.
19. Zeng ZB: **Precision mapping of quantitative trait loci.** *Genetics* 1994, **136**(4):1457-68.
20. Haley CS, Knott SA: **A simple regression method for mapping quantitative trait loci in line crosses using flanking markers.** *Heredity* 1992, **69**(4):315-24.
21. Lander ES, Botstein D: **Mapping mendelian factors underlying quantitative traits using RFLP linkage maps.** *Genetics* 1989, **121**(Genetics):185-199.
22. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**(1):27-30.
23. Breiman L: **Random Forests.** *Machine Learning* 2001, **45**(1):5.
24. Tibshirani R: **Regression Shrinkage and Selection Via the Lasso.** *Journal of the Royal Statistical Society, Series B* 1996, **58**:267-288.
25. Zou H, Hastie T: **Regularization and variable selection via the Elastic Net.** *Journal of the Royal Statistical Society, Series B* 2005, **67**:301-320.
26. Lunetta K, Hayward L, Segal J, Van Eerdewegh P: **Screening large-scale association study data: exploiting interactions using random forests.** *Bmc Genetics* 2004, **5**(1):32.
27. Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD: **A comparison of analytical methods for genetic association studies.** *Genet Epidemiol* 2008, **32**(8):767-78.
28. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G: **Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'.** *Nat Genet* 2005, **37**(3):225-32.
29. Overall RW, Kempermann G, Peirce J, Lu L, Goldowitz D, Gage FH, Goodwin S, Smit AB, Airey DC, Rosen GD, Schalkwyk LC, Sutter TR, Nowakowski RS, Whatley S, Williams RW: **Genetics of the hippocampal transcriptome in mouse: a systematic survey and online neurogenomics resource.** *Frontiers in Neurogenomics* 2009, **1**(3).
30. Brem RB, Kruglyak L: **The landscape of genetic complexity across 5,700 gene expression traits in yeast.** *Proc Natl Acad Sci USA* 2005, **102**(5):1572-7.
31. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**(1):109-26.
32. Mnaimneh S, Davierwala AP, Haynes J, Moffat J, Peng WT, Zhang W, Yang X, Pootoolal J, Chua G, Lopez A, Trochesset M, Morse D, Krogan NJ, Hiley SL, Li Z, Morris Q, Griggull J, Mitsakakis N, Roberts CJ, Greenblatt JF, Boone C, Kaiser CA, Andrews BJ, Hughes TR: **Exploration of essential gene functions via titratable promoter alleles.** *Cell* 2004, **118**(1):31-44.
33. Wessel J, Zapala MA, Schork NJ: **Accommodating pathway information in expression quantitative trait locus analysis.** *Genomics* 2007, **90**(1):132-42.
34. Wu C, Delano DL, Mitro N, Su SV, Janes J, McClurg P, Batalov S, Welch GL, Zhang J, Orth AP, Walker JR, Glynne RJ, Cooke MP, Takahashi JS, Shimomura K, Kohsaka A, Bass J, Saez E, Wiltshire T, Su AI: **Gene set enrichment in eQTL data identifies novel annotations and pathway regulators.** *PLoS Genet* 2008, **4**(5):e1000070.
35. Ghazalpour A, Doss S, Zhang B, Wang S, Plaisier C, Castellanos R, Brozell A, Schadt EE, Drake TA, Lusis AJ, Horvath S: **Integrating genetic and network analysis to characterize genes related to mouse weight.** *PLoS Genet* 2006, **2**(8):e130.

36. Suthram S, Beyer A, Karp RM, Eldar Y, Ideker T: **eQED: an efficient method for interpreting eQTL associations using protein networks.** *Mol Syst Biol* 2008, **4**:162.
37. Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, Koller D: **Learning a prior on regulatory potential from eQTL data.** *PLoS Genet* 2009, **5**(1):e1000358.
38. 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 USA* 2006, **103**(38):14062-7.
39. La Merrill M, Gordon RR, Hunter KW, Threadgill DW, Pomp D: **Dietary fat alters pulmonary metastasis of mammary cancers through cancer autonomous and non-autonomous changes in gene expression.** *Clin Exp Metastasis* 2010, **27**(2):107-16.
40. Chen X, Hackett CA, Niks RE, Hedley PE, Booth C, Druka A, Marcel TC, Vels A, Bayer M, Milne I, Morris J, Ramsay L, Marshall D, Cardle L, Waugh R: **An eQTL analysis of partial resistance to Puccinia hordei in barley.** *PLoS ONE* 2010, **5**(1):e8598.
41. Wang J, Yu H, Xie W, Xing Y, Yu S, Xu C, Li X, Xiao J, Zhang Q: **A global analysis of QTLs for expression variations in rice shoots at the early seedling stage.** *The Plant Journal* .
42. Viñuela A, Snoek LB, Riksen JAG, Kammenga JE: **Genome-wide gene expression regulation as a function of genotype and age in C. elegans.** *Genome Res* 2010, **20**(7):929-37.
43. Strobl C, Boulesteix AL, Zeileis A, Hothorn T: **Bias in random forest variable importance measures: illustrations, sources and a solution.** *BMC Bioinformatics* 2007, **8**(1):25.
44. Rudd MF, Williams RD, Webb EL, Schmidt S, Sellick GS, Houlston RS: **The predicted impact of coding single nucleotide polymorphisms database.** *Cancer Epidemiol Biomarkers Prev* 2005, **14**(11 Pt 1):2598-604.
45. Jegga AG, Gowrisankar S, Chen J, Aronow BJ: **PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease.** *Nucleic Acids Res* 2007, **35** Database: D700-6.
46. Chan EKF, Hawken R, Reverter A: **The combined effect of SNP-marker and phenotype attributes in genome-wide association studies.** *Anim Genet* 2009, **40**(2):149-56.
47. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10**(6):392-404.
48. Phillips PC: **Epistasis-the essential role of gene interactions in the structure and evolution of genetic systems.** *Nat Rev Genet* 2008, **9**(11):855-67.
49. Carlborg O, Haley CS: **Epistasis: too often neglected in complex trait studies?** *Nat Rev Genet* 2004, **5**(8):618-25.
50. Moore JH: **The ubiquitous nature of epistasis in determining susceptibility to common human diseases.** *Hum Hered* 2003, **56**(1-3):73-82.
51. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta D, Sieberts SK, Monks S, Reitman M, Zhang C, Lum PY, Leonardson A, Thieringer R, Metzger JM, Yang L, Castle J, Zhu H, Kash SF, Drake TA, Sachs A, Lusis AJ: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nat Genet* 2005, **37**(7):710-7.
52. Wang J, Williams RW, Manly KF: **WebQTL: web-based complex trait analysis.** *Neuroinformatics* 2003, **1**(4):299-308.
53. Liaw A, Wiener M: **Classification and Regression by randomForest.** *R News* 2002, **2**(3):18-22.
54. Foster SD: **Incorporating LASSO Effects into a Mixed Model for Quantitative Trait Loci Detection.** *Journal of Agricultural, Biological, and Environmental Statistics* 2007, **12**(2):300.
55. Beyer A, Workman C, Hollunder J, Radke D, Möller U, Wilhelm T, Ideker T: **Integrated assessment and prediction of transcription factor binding.** *PLoS Comput Biol* 2006, **2**(6):e70.

doi:10.1186/1471-2164-11-502

Cite this article as: Michaelson et al.: Data-driven assessment of eQTL mapping methods. *BMC Genomics* 2010 **11**:502.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

