Europe PMC Funders Group
Author Manuscript
Nat Protoc. Author manuscript; available in PMC 2012 September 01.

Published in final edited form as:
Nat Protoc. ; 7(3): 500–507. doi:10.1038/nprot.2011.457.

Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses

Oliver Stegle[1,2,6], **Leopold Parts**[3,6], **Matias Piipari**[4], **John Winn**[5], and **Richard Durbin**[3]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany

[2]Max Planck Institute for Developmental Biology, Tübingen, Germany

[3]Wellcome Trust Sanger Institute, Cambridge, UK

[4]Pear Computer LLP, London, UK

[5]Microsoft Research, Cambridge, UK

Abstract

We present PEER (probabilistic estimation of expression residuals), a software package implementing statistical models that improve the sensitivity and interpretability of genetic associations in population-scale expression data. This approach builds on factor analysis methods that infer broad variance components in the measurements. PEER takes as input transcript profiles and covariates from a set of individuals, and then outputs hidden factors that explain much of the expression variability. Optionally, these factors can be interpreted as pathway or transcription factor activations by providing prior information about which genes are involved in the pathway or targeted by the factor. The inferred factors are used in genetic association analyses. First, they are treated as additional covariates, and are included in the model to increase detection power for mapping expression traits. Second, they are analyzed as phenotypes themselves to understand the causes of global expression variability. PEER extends previous related surrogate variable models and can be implemented within hours on a desktop computer.

INTRODUCTION

Here we present a protocol to improve the power and interpretability of population-level gene expression analyses. The protocol is based on the software suite known as PEER, which consists of a collection of Bayesian approaches to infer hidden determinants and their effects from gene expression profiles by using factor analysis methods[1,2].

Our understanding of the genetic basis of gene expression has been developed by studying species such as yeast[3-5], worms[6], mice[7,8] and humans[9-12]. As large-scale expression data were generated from these and other species over the past decade, it became increasingly

© 2012 Nature America, Inc. All rights reserved.

Correspondence should be addressed to O.S. (oliver.stegle@tuebingen.mpg.de)..
[6]These authors contributed equally to this work.

Note: Supplementary information is available via the HTML version of this article.

AUTHOR CONTRIBUTIONS O.S., L.P., J.W. and R.D. designed the probabilistic models underlying the protocol. O.S., L.P. and M.P. developed the PEER software suite. O.S. and L.P. wrote the paper with input from all authors.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

apparent that there are nontrivial statistical hurdles to overcome in order to fully harness their power[13,14].

First, there are heterogeneous sources of variability in expression data, and dealing with them requires careful, iterative analysis. It is important to account for technical and other confounding sources of expression variation, including batch effects, environmental influences, sample history and other known or unknown factors. At the same time, we need to model the effect of the variables of interest, such as locus genotypes or case-control status. Second, it is beneficial to integrate existing knowledge on pathways and regulatory networks into the statistical analysis and mapping procedure to better understand the regulation of multiple transcripts from a single locus[15,16].

Learning unmeasured determinants of gene expression variation

Several studies have found that batch effects and other global confounders reduce the power to find expression quantitative trait loci (eQTLs)[17,18]. However, these factors cannot be directly included in modeling if they are not measured. We and others have developed novel statistical approaches to account for such hidden determinants of expression variation[1,19,20]. Our method infers a small number of variables for every individual in the data set. We assume that these variables have a broad influence, and thus each of them has an effect size for every gene. We then treat them analogously with measured global confounders such as batch labels or measured RNA quality, and include them in the model to both improve power to detect true eQTLs and reduce spurious false-positive associations[1].

Learning cellular features from gene expression

eQTL studies have shown the existence of regulatory hotspots that are associated with multiple genes in *trans*[3-5,21,22]. Although a single enhancer may regulate multiple genes via a direct mechanism, an important alternative explanation is the existence of some other biological variable that affects the expression levels and is itself under genetic regulation. PEER can help to find the biological origin of such factors and hence the observed eQTL hotspots[2]. For example, we may model the expression variability of genes as a function of transcription factor activations. Our method again infers a small number of variables, now corresponding to the activation of each transcription factor, for every individual. However, we no longer assume that these variables affect all gene expression levels. Instead, they are likely to influence the gene expression level only if the corresponding transcription factor targets the gene. For all other genes, the effect of the variable is likely to be very close to zero. The information pertaining to which factor can affect which gene is provided to PEER, for example, in the form of transcription factor binding or pathway membership data. The learned factors can themselves be used as traits and mapped to genetic loci, thereby explaining *trans* hotspots as a composite effect of interpretable biological features[2].

Applications

PEER has been successfully used in several gene expression variation studies to increase the number of eQTL discoveries. Initially, we demonstrated additional eQTL findings in yeast and mouse data sets[1]. We then found three times more eQTLs compared with a standard linear model in the HapMap II gene expression data using genotyping arrays[1], as well as in full genome sequences from the 1000 Genomes project[23]. In the MuTHER project, the eQTL findings increased twofold in skin and fat tissues, as well as in lymphoblastoid cell lines[24].

Applications of PEER have also helped in understanding the genetics of inferred cellular traits when including prior information on targets of cellular features[2]. We reanalyzed the data from Smith and Kruglyak[5], inferring the activations of 167 transcription factors in a set

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

of yeast segregants. Here 15% of hotspot-associated *trans* eQTLs could be explained by genetic control of the transcription factor activations, which had downstream effects on gene expression levels.

PEER can be applied to the analysis of reference populations as well as case-control studies and differential expression analyses. For such uses, an additional covariate for the case status must be introduced, but the availability of genotype data is not required. PEER can also account for further measured covariates; for example when data are combined from multiple experiments or laboratories. More generally, PEER can be used for analyzing any high-dimensional phenotype with population-scale data.

Algorithm

The application of PEER to gene expression studies consists of two steps (Fig. 1). First, PEER is used to infer hidden expression determinants from the expression profiles. Second, the learned factors are used in alternative genetic analyses (Fig. 1a–c).

The learning step done in PEER infers hidden expression determinants from the normalized and preprocessed expression profiles, taking any known covariates into account. The learned variables can be constrained to affect known sets of genes via a prior connectivity matrix. By default, with no prior connectivity given, they are assumed to be global and to affect large fractions of all genes. The learning algorithm in PEER estimates a suitable number of factors implicitly and only explains broad variance components, thereby helping to avoid overfitting[1].

PEER produces learned factor activations, their effects on each gene and a residual data set of the expression values after subtracting the factor contribution (Fig. 1). eQTLs can then be mapped on the residuals directly (Fig. 1a), or on original data, treating the learned factors as covariates in the association tests (Fig. 1b). The factors can also be used as phenotypes in genetic mapping (Fig. 1c) or they can be tested for association with other phenotypes.

PEER itself does not offer low-level data processing. Gene expression normalization and the necessary preprocessing of genotype information need to be done using external tools[25,26]. When RNA-seq estimates are used for transcript abundance, we recommend using DESeq to estimate library sizes and variance-stabilize expression data sets[27]. If it is available from low-level processing, PEER can also correctly use information on measurement uncertainty for specific probes and samples.

The learning of factors implemented in PEER is based on efficient approximate inference techniques that ensure computational tractability for practical applications while retaining the necessary accuracy of the results obtained[1,2,28]. Once the hidden expression determinants are learned, association testing is carried out in the second analysis step by using a range of existing methods. Our instructions assume the use of a standard linear model that yields a test statistic for linkage or association between individual variants and genes, such as that implemented in R/qtl (ref. 29) or PLINK[30]. To assess genome-wide significance, these statistics have to be converted to association probabilities and corrected for multiple testing for both genetic variants and transcript levels (e.g., using Bonferroni or false discovery rate (FDR)[31]).

Comparison with other methods

The functionality implemented in PEER is in part also available in alternative packages that account for confounding influences in eQTL studies. These either recover the set of hidden factors explicitly[19] or use the covariance structure induced by them[20,32]. The algorithm implemented in PEER is most closely related to surrogate variable analysis[19] and has

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

previously been compared in more detail[1]. Notably, PEER allows for the following: the automatic setting of an appropriate number of hidden determinants to learn, the incorporation of probe-level uncertainties (e.g., if probe measurements are not variance-stabilized from count data) and the combination of the inference of hidden confounding factors while accounting for the effect of known covariates.

Similarly, alternative approaches to learn hidden determinants to be used as trait variables have been suggested. For example, other bilinear models have previously been used by Biswas *et al.*[33], and methods to combine eQTL mapping with integrated network models have been considered by Zhu *et al.*[34] and Aten *et al.*[35]. Notably, the supervised factor inference in PEER is scalable and can be used on genome-wide data sets while retaining sufficient accuracy, thus allowing for meaningful conclusions to be drawn from the inferred quantities themselves[2].

Limitations

PEER is applicable to a wide range of analysis settings. At present, there is no support for mixed modeling, wherein some variables (e.g., zygosity, gender, batch) have a random effect. In addition, information on population structure, if not encoded by the covariates (e.g., by introducing principal components of the genotype data), is not included in the model, and it may be recapitulated in the inferred factors. Finally, as a rule of thumb, the number of samples needs to be larger than the expected number of factors to be learned. In combination with prior knowledge, it is, however, feasible to statistically identify a greater number of factors than individuals in the data set[2].

Experimental design

Required data matrices—The application of PEER for eQTL mapping requires gene expression profiles and genotype information for a set of N individuals. Example data files are provided in Supplementary Data 1. For applications not related to QTL mapping, the genotype matrix is not required.

- *Expression matrix*. Matrix of shape $N \times G$, where G denotes the number of measured gene expression levels. Expression estimates can be positive or negative values and on a logarithmic scale, as provided by most common normalization methods. Ideally, the expression estimates should be variance-stabilized.
- *Genotype matrix*. Matrix of shape $N \times S$, where S denotes the number of genotypes. For fitting association models, genotypes should be encoded as the minor allele count (0/1 for haploid, 0/1/2 for diploid organisms). Existing packages (R/qtl, PLINK) will have their own format requirements for genotype data.

Optional data matrices

- *Covariates matrix*. Matrix of size $N \times C$, where C is the number of covariates. Examples of covariates include other cofactors such as gender information, population membership or batch variables to be accounted for in the analysis. Categorical variables (e.g., batch number) have to be encoded as indicators, with a different binary variable for each batch, having a value of 1 if the individual was in the batch and a value of 0 otherwise. See the tutorials provided with the PEER package for further examples.
- *Uncertainty matrix.* If they are provided by the low-level processing of gene expression, it is possible to include uncertainty estimates specific to each gene. This matrix has the same dimension as the expression estimate, $N \times G$, providing

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

the variance of each measurement. If no uncertainty matrix is provided, PEER estimates the data variance automatically.

- *Prior connectivity matrix.* To infer factors that affect specific genes, PEER requires knowledge about which genes every factor can influence. This information takes the form of a matrix of size $K \times G$, where the (k,g) entry corresponds to the a priori probability of factor k having a nonzero effect on gene g. For example, if a transcription factor k is known to bind to the promoter of gene g, the (k,g) entry should be close to 1, denoting a high probability of true regulation. In the simplest case, this matrix is binary with a value of 1 if variable k is known to affect gene g and a value of 0 otherwise. Prior link probabilities between 0 and 1, reflecting uncertain information, are also possible[2]. Note that the genes in the expression matrix and in the prior connectivity matrix have to be ordered in the same way.

File formats and example data files

- Data files should be formatted as comma-separated values (CSVs) or tab-delimited values. CSV files can be exported from common software for genomic data.
- Missing data are not supported for covariates and expression levels. Individuals with missing values should be dropped before the protocol, or their missing values need to be imputed. We suggest using the R 'impute' package for this task (see Step 5).
- We provide data files with an example eQTL experiment based on yeast data[4] in a format compatible with R/qtl (Supplementary Data 1). These experimental data include a genotype matrix (genotype.csv), expression levels (expression.csv), covariates (covariates.csv) and a prior connectivity matrix for learning biological variables (prior.csv). These example data can be used to understand the PEER workflow for other experiments and to reproduce the analysis steps in this protocol.

MATERIALS

EQUIPMENT

- Computer operating system: Linux or Mac OSX
- R (http://www.r-project.org): an open-source software environment for statistical computing (version 2.9.0 or higher)
- Example data (example data sets and scripts to reproduce the results shown here are available as Supplementary Data 1)

Required R packages

- PEER R package (can be downloaded from https://github.com/PMBio/peer/wiki. The examples in the protocol are based on PEER 1.3.)
- R/qtl (optional for QTL mapping in crosses; it can be installed from R by entering `install.packages("qtl"` at the command prompt)
- Impute R package (optional for imputation of missing values in gene expression; it can be installed from R by entering `install.packages("impute")` at the command prompt)

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Optional implementation

- Optionally, all analysis steps outlined in this protocol can also be implemented using the Python interface of PEER, offering near-identical syntax, or the command-line tool (see https://github.com/PMBio/peer/wiki).

PROCEDURE

Preparation and data loading ● TIMING 15 min

1| Load PEER and R/qtl:

```
> library(peer)

> library(qtl)
```

2| Load the prepared data matrices into an R/qtl cross object in the running R session. We assume that the data files follow the naming convention of the example data provided. For details on required and optional data matrices, see Experimental design.

```
> cross < - read.cross(format="csvs",genfile="genotype.csv",
phefile="expression.csv", genotypes=c(0,1))
```

Learning of hidden determinants from gene expression using PEER ● TIMING 30 min–2 h

3| Build the model (Fig. 1).

```
> model=PEER()
```

4| Set the maximum number of unobserved factors to model.

```
> PEER_setNk(model, n_unobserved_factors)
```

Note that unlike PCA-type models, the number of unobserved factors is not crucial when no prior is specified because PEER uses automatic relevance determination[36] to choose a suitable effective number of factors. Hence, n_unobserved_factors needs only to be set to a sufficiently large value (for technical details see Stegle *et al.*[1]). If no prior information on the magnitude of confounding effects is available, we recommend using 25% of the number of individuals contained in the study but no more than 100 factors.

5| (Optional) Impute missing values in gene expression. If the gene expression data set contains missing values, we suggest using the impute package to fill in the missing measurements.

```
> library(impute)

> cross$pheno < - impute.knn(cross$pheno)
```

6| Set expression data.

```
> PEER_setPhenoMean(model, as.matrix(cross$pheno))
```

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

7| (Optional) Set expression data uncertainty.

```
> PEER_setPhenoVar(model, as.matrix(expression_variance))
```

8| (Optional) Set observed covariates.

```
> PEER_setCovariates(model, as.matrix(covariates))
```

9| (Optional) Set prior connectivity.

```
> PEER_setSparsityPrior(model, as.matrix(prior))
```

If prior connectivity is specified, setting the number of unobserved factors is not needed. The number of factors in the prior information matrix over-rides any previous specification.

10| Train the model, observing convergence:

```
> PEER_update(model)
```

If the model is not converged after the default 1,000 iterations, and the variance of the residuals keeps decreasing, choose a higher value of iterations, e.g., 10,000.

```
> PEER_setNMax_iterations(model, 10000)
```

A total of 100 iterations should be sufficient to reach convergence on most data sets.

? TROUBLESHOOTING

Diagnostics and interpretation of learned hidden determinants ● TIMING 30 min

11| Run correlation analyses between the inferred variables and batch confounding effects. For example, to check first factor and first covariate, use:

```
> cor (PEER_getX(model)[,1], PEER_getCovariates(model)[,1]))
```

If several inferred factors correlated with batch effects/confounders, this can be indicative of a more complex, nonlinear effect of these known covariates on the mRNA levels. Scatter plots can help understand the nature of these dependencies.

12| Plot the posterior variance of the factor weights and convergence diagnostics. If there is a natural choice for the number of factors (usually observed as an ‘elbow’), and you are using the inferred factors as covariates in the linear model (see below), consider only including the more relevant factors, or rerun the model with this number of factors (Fig. 2).

```
> PEER_plotModel(model)
```

Application of learned hidden determinants in eQTL analyses ● TIMING 5 h

13| *Correcting for learned determinants in eQTL scans*. This can be done using option A (performing an eQTL scan on residual data set after accounting for confounders (Fig. 1a)) or option B (including observed and inferred confounders in the model (Fig. 1b)).

A. Perform an eQTL scan on a residual data set after accounting for confounders

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

i. This allows the use of nonparametric QTL methods such as rank correlation. To perform per-marker nonparametric tests on PEER residuals for genes 7–12, implicitly estimating genotype probabilities where data are missing, and to calculate *q*-values, use:

```
> cross$pheno=PEER_getResiduals(model)

> colnames(cross$pheno)=1:dim(cross$pheno)[2]

> lod_scores=scanone(cross, model="np", pheno.col=7:12)
```

ii. Correct for multiple testing by using FDR for the first tested trait, on the basis of χ^2 *P* values from the nested model.

```
> qvals=p.adjust(dchisq(2.*log(10)*lod_scores[,3]),
method="fdr")
```

Alternatively, for all traits without correcting for tests with multiple transcripts, use:

```
> qvals=apply(lod_scores[,3:8],2,function(x)
{p.adjust(dchisq(2*log(10)*x,1), method="fdr")
```

B. Include observed and inferred confounders in the model

i. This approach can only be used with parametric models. To perform per-marker parametric test for genes 7–12, using inferred PEER factors as covariates, use:

```
> lod_scores=scanone(cross, method="hk", model="normal",
pheno.col=7:12, addcovar=PEER_getX(model))
```

The number of discovered associations between expression levels and genotypes of nearby loci should increase as the variability attributable to other global factors is explained away. When testing for association with variants in a 10,000-bp window around the probe, we found many additional eQTLs for a range of LOD score cutoffs (Fig. 2a).

? TROUBLESHOOTING

14| Genetic mapping by using the hidden determinants (Fig. 1c). This can be achieved by using option A (mapping the genetic basis of inferred variables by setting them as the cross phenotype, followed by standard mapping) or option B (further genetic analyses based on the learned factors).

A. Mapping the genetic basis of inferred variables by setting them as the cross phenotype, followed by standard mapping

i. To map the genetic basis of inferred variables, set them as the cross phenotype, followed by standard mapping:

```
> cross$pheno < - PEER_getX(model)

> colnames(cross$pheno) < - 1:n_factors
```

```
> lod_scores < - scanone(cross, pheno.col=1:n_factors,
method="hk")
```

Factors associated with a genotype are indicative of a *trans* eQTL hotspot, with many gene expression levels associated with a single variant. If a biological prior is not being used, include this locus genotype in the model as a covariate and rerun; this increases the interpretability of the results. Depending on strong effects or suitable prior information, learned factors explain *trans* eQTL hotspots with few individual factors (Fig. 3).

B. Further genetic analyses based on the learned factors

i. The learned factors can be used in other contexts, e.g., to identify genetic interactions between cellular features, the genetic state and expression levels[2]. For an interaction scan of gene 10 and factor 5 using R/qtl, use:

```
> int_lods=scanone(cross, pheno.col=10,
intcovar=PEER_getX(model)[,5], addcovar=PEER_getX(model)[,5])
```

In this case, additional care has to be taken for multiple testing corrections, as the space of possible interactions is large.

? TROUBLESHOOTING

? TROUBLESHOOTING

Troubleshooting advice can be found in Table 1. For additional troubleshooting and diagnostic guidelines, please also consult the PEER WIKI online (https://github.com/PMBio/peer/wiki).

TIMING

Steps 1 and 2, loading data and setting up the R environment: 15 min

Steps 3–10, setting up the PEER model and running inference: 30 min–2 h (for 6,000 genes and 170 factors). Application of PEER scales linearly in time and memory consumption with the number of genes and individuals, and quadratically with the number of learned factors. We recommend creating a moderate-sized data set to estimate the running time.

Steps 11 and 12, diagnostics of inference results: 30 min

Steps 13 and 14, application to eQTL analysis: 5 h (for 6,000 genes and 3,000 SNPs)

Computation time for eQTL scans scales linearly with the number of tested loci and transcripts.

ANTICIPATED RESULTS

PEER produces estimates of hidden determinants of gene expression that aid analysis of the data. Depending on whether prior information is included while learning (Step 9), these estimates resemble either interpretable cellular features or broad variance components (Fig. 2). In eQTL mapping, the inferred factors can be used to increase power in genetic mapping (Fig. 2) and to identify the genetic determinants of learned cellular features (Fig. 3). In addition to applications in genetic mapping, hidden determinants of gene expression can be used in other analyses of gene expression (see, for example, Leek *et al.*[19]).

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank R. Brem and L. Kruglyak for providing genotype and expression phenotype data to be included alongside this protocol. This work received financial support from the Wellcome Trust (grant no. WT077192/Z/05/Z) and the Technical Computing Initiative (Microsoft Research). O.S. received funding from the Volkswagen Foundation.

References

1. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput. Biol. 2010; 6:e1000770. [PubMed: 20463871]
2. Parts L, Stegle O, Winn J, Durbin R. Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet. 2011; 7:e1001276. [PubMed: 21283789]
3. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002; 296:752–755. [PubMed: 11923494]
4. Brem RB, Storey JD, Whittle J, Kruglyak L. Genetic interactions between polymorphisms that affect gene expression in yeast. Nature. 2005; 436:701–703. [PubMed: 16079846]
5. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. PLoS Biol. 2008; 6:e83. [PubMed: 18416601]
6. Rockman MV, Kruglyak L. Genetics of global gene expression. Nat. Rev. Genet. 2006; 7:862–872. [PubMed: 17047685]
7. Valdar W, et al. Genome-wide genetic association of complex traits in heterogeneous stock mice. Nat. Genet. 2006; 38:879–887. [PubMed: 16832355]
8. Doss S, Schadt EE, Drake TA, Lusis AJ. *Cis*-acting expression quantitative trait loci in mice. Genome Res. 2005; 15:681–691. [PubMed: 15837804]
9. Stranger BE, et al. Population genomics of human gene expression. Nat. Genet. 2007; 39:1217–1224. [PubMed: 17873874]
10. Cheung VG, Spielman RS. Genetics of human gene expression: mapping DNA variants that influence gene expression. Nat. Rev. Genet. 2009; 10:595–604. [PubMed: 19636342]
11. Montgomery SB, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
12. Pickrell JK, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
13. Breitling R, et al. Genetical genomics: spotlight on QTL hotspots. PLoS Genet. 2008; 4:e1000232. [PubMed: 18949031]
14. Franke L, Jansen RC. eQTL analysis in humans. Methods Mol. Biol. 2009; 573:311–328. [PubMed: 19763935]
15. Lee SI, et al. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc. Natl. Acad. Sci. USA. 2006; 103:14062–14067. [PubMed: 16968785]
16. 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]
17. Balding, DJ. Handbook of Statistical Genetics. Wiley-Interscience; 2007.
18. Plagnol V, et al. Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. PLoS One. 2008; 3:e2966. [PubMed: 18698422]
19. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007; 3:e161.
20. Kang HM, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. Genetics. 2008; 180:1909–1925. [PubMed: 18791227]

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

21. Schadt EE, et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 2005; 37:710–717. [PubMed: 15965475]
22. Small KS, et al. Identification of an imprinted master *trans* regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat. Genet. 2011; 43:561–564. [PubMed: 21572415]
23. 1000 Genomes Project Consortium, 1000 Genomes Project. et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
24. Nica AC, et al. The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. PLoS Genet. 2011; 7:e1002003. [PubMed: 21304890]
25. Huber W, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics. 2002; 18(suppl. 1):S96–S104. [PubMed: 12169536]
26. Pearson RD, et al. PUMA: a Bioconductor package for propagating uncertainty in microarray analysis. BMC Bionf. 2009; 10:211.
27. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11:R106. [PubMed: 20979621]
28. Rattray M, Stegle O, Sharp K, Winn J. Inference algorithms and learning theory for Bayesian sparse factor analysis. J. Phys. Conf. Ser. 2009; 197:012002.
29. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003; 19:889–890. [PubMed: 12724300]
30. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 2007; 81:559–575. [PubMed: 17701901]
31. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA. 2003; 100:9440–9445. [PubMed: 12883005]
32. Listgarten J, Kadie C, Schadt EE, Heckerman D. Correction for hidden confounders in the genetic analysis of gene expression. Proc. Natl. Acad. Sci. USA. 2010; 107:16465–16470. [PubMed: 20810919]
33. Biswas S, Storey JD, Akey JM. Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC Bioinf. 2008; 9:244.
34. Zhu J, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat. Genet. 2008; 40:854–861. [PubMed: 18552845]
35. Aten JE, Fuller TF, Lusis AJ, Horvath S. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. BMC Sys. Biol. 2008; 2:34.
36. MacKay DJC. Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks. Network. 1995; 6:469–505.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

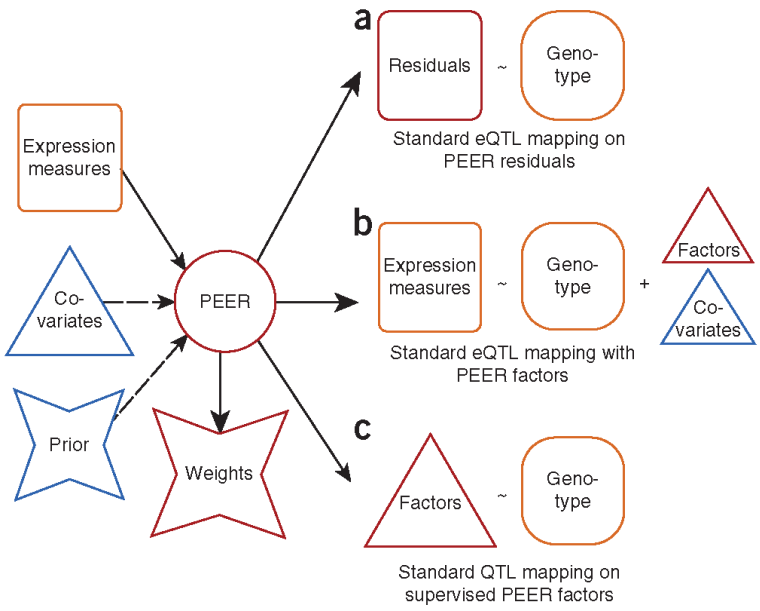

Figure 1.

Protocol alternatives for applying PEER to analyses of expression QTL studies. PEER infers hidden factors (red triangles), their weights (red star) and a residual gene expression matrix (red square) from a set of gene expression levels (orange squares). If available, experimental confounders (blue triangles) or prior information on groups of genes affected by a factor (blue star) can be included. (**a**) Results of PEER are processed in downstream QTL analysis on the residual data set. (**b**,**c**) Alternatively, the inferred factors can be used (**b**) as additional covariates or (**c**) as phenotypes themselves. Orange shapes denote experimental measurements; blue shapes denote prior information including covariates; and the red shapes denote PEER results. Similar shapes of the figures denote similar matrix dimensions. Dashed arrows indicate dependencies that optionally can be taken into account.

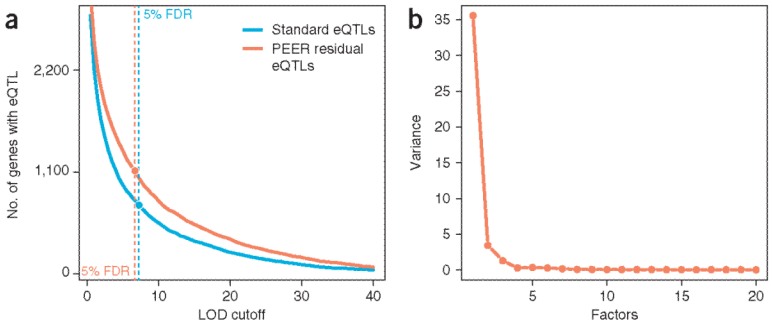

Figure 2.

Illustrative analysis results of the application of PEER. Data are from Smith and Kruglyak[5]. (**a**) The number of significant associations between locus genotype and probe expression levels is expected to increase upon the application of PEER (orange line) compared with the standard model (light blue line) for a range of LOD score cutoffs (FDR threshold of 5% shown as dashed line). (**b**) Diagnostic plot of the factor relevance (ARD parameters). PEER deactivates all but the first three factors in this data set.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

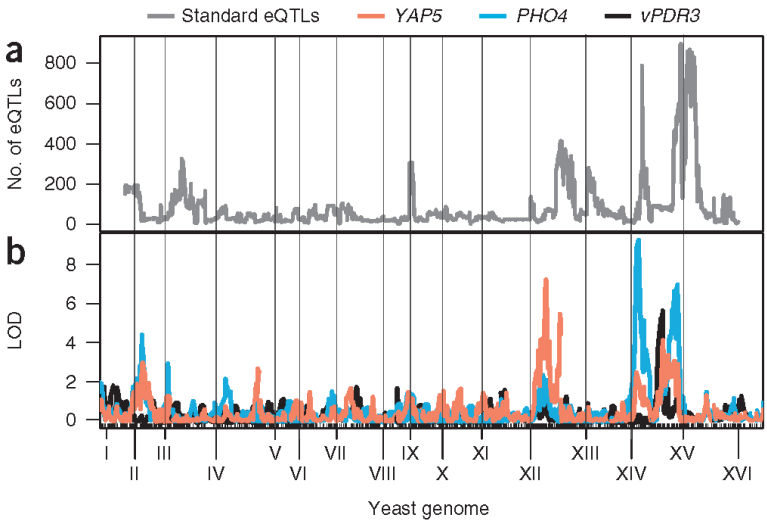

Figure 3.

Illustrative analysis results of the application of PEER in supervised mode. Data are from Smith and Kruglyak[5]. (**a**) Density of the genetic associations between genetic markers and genes (per-gene FDR < 5%). (**b**) When PEER is used to infer transcription factor activations, the resulting variables are themselves influenced by genotype, which are demonstrated here by linkage plots of *YAP5* (orange), *PHO4* (light blue) and *PDR3* (black) factors. Inferred factors capture some of the eQTL hotspots from standard eQTLs.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

TABLE 1

Troubleshooting table.

Step	Problem	Possible reason	Solution
10	PEER does not converge	Choice of the number of iterations *N* is too high, or convergence criteria are too strict	Reduce the number of iterations and use PEER_plotModel to observe convergence. Adjust the criteria if appropriate
	Residual variance does not decrease during training	Expression levels used are not variance stabilized	Use a different low-level normalization strategy that offers variance stabilization of expression estimates
	Factor inference in supervised mode yields unstable factor inferences	If factors have too similar targets, they may be statistically unidentifiable	Filter or merge factors, such that the prior information matrix only contains factors with a distinct target set
13	No increase in power when testing for genetic associations on residuals	Not enough or too many active factors. If large *trans* hotspots are dominating, associations may get erroneously explained away as confounding factors	Consider alternative prior values for 'alpha' (see R code examples in r_demo.R; function PEER_setPriorAlpha). Include master regulators as covariates
14	Learned factors have no genetic associations	Factors preferably explain large variance components that are not due to genetic variation	Use different prior information. Filtered pathway or transcription factor binding is recommended. Alternatively, try unsupervised mode