DataRemix: a universal data transformation for optimal inference from gene expression datasets

Weiguang Mao[1,2], Ryan Hausler[2] and Maria Chikina[1,2*]

Abstract

RNAseq technology provides an unprecedented power in the assesment of the transcription abundance, which benifits various downstream biological studies, such as gene-correlation network inference and eQTL discovery. However, raw gene expression values have to be normalized for nuisance biological variation and technical covariates, and different normalization strategies can lead to dramatically different results in the downstream study. Here we present a simple three-parameter transformation, DataRemix, which can greatly improve the biological utility of gene expression datasets without any specific knowledge on the dataset. As we optimize the transformation with respect to the downstream biological objective, this parametric framework reweighs the contribution of each hidden factor and make the biological signals visible. We demonstrate that DataRemix can outperform complicatd normalization methods which make explicit use of dataset specific technical factors. Also we show that DataRemix can be efficiently optimized via Thompson Sampling approach, which makes it feasible for computationally expensive objectives such as eQTL analysis. Finally we reanalyze the Depression Gene Networks (DGN) dataset, and we highlight new *trans*-eQTL networks which were not reported in the initial study.

Genome-wide gene expression studies have become a staple of large scale systems biology and clinical projects. However, while gene expression is the most mature high-throughput technology, technical challenges remain. Raw gene expression values must be normalized for any technical and nuisance biological variation and the normalization strategy can have dramatic effects on the results of downstream analysis. This is especially true in cases where the sought-after gene expression effects are likely to be small in magnitude, such as expression quantitative trail loci (eQTLs). Increasingly sophisticated normalization methods have been proposed and many are computational intensive and/or can have multiple free parameters that must be optimized (Leek & Storey 2007; Stegle *et al.*. 2010; Listgarten *et al.*. 2010; Kang *et al.*. 2008; Mostafavi *et al.*. 2013). Moreover, it is not uncommon for one dataset to yield multiple normalized versions that maximize performance in a particular setting (such as the discovery of *cis*- and *trans*-eQTLs Battle *et al.*. 2014), highlighting the complexity of the normalization problem.

Singular value decomposition (SVD) is one of the most widely used gene expression analysis tools (Alter *et al.*. 2000, 2003) that can also be used for data normalization. Using the SVD we can simply remove the first few principle components that are presumed to represent technical factors such as batch-effects or other nuisance variation. In some cases this dramatically improves downstream performance, for example in the case of eQTL analysis (Mostafavi *et al.*. 2013). The drawback of this method is that the exact number of components to remove must be determined empirically and some meaningful biological signals may be lost in the process.

More sophisticated approaches attempt to partition data structure into useful and nuisance variation and remove only the latter (Leek & Storey 2007; Stegle *et al.*. 2010; Listgarten *et al.*. 2010; Kang *et al.*. 2008; Mostafavi *et al.*. 2013). These can improve on the naive SVD-based normalization but require additional input such as technical covariates, or the study design. The success of these methods ultimately depends on the availability and quality of such meta data and some methods still rely on parameter optimization to maximize performance. These widely used normalization approaches all have a common theme that the rely in part on the intrinsic data structure. One key property that contributes to the success of these approaches is that for many biological questions of interest nuisance variation (of technical or biological origin) is larger in magnitude than useful variation. Our proposed method, DataRemix, explicitly formalizes this view of the data normalization problem.

In this work we demonstrate that biological utility of gene expression datasets can be dramatically improved with a simple three-parameter transformation, DataRemix. Our method does not require any dataset specific knowledge but rather optimizes the transformation with respect to some independent *objective* of data quality, such as the quality of the gene-correlation network or the number of *trans*-eQTL discoveries. Because our method requires only the gene expression data and biological validity objective, it can

[*] Correspondence: mchikina@pitt.edu
[2] Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh,
Full list of author information is available at the end of the article

be applied to any publicly available dataset. We focus our study on gene expression data for which methods for quantifying biological validity are well established, but our approach can be readily applied to any high-throughput molecular data for which similar quality metrics can be defined. We show that this strategy can outperform methods that make explicit use of dataset specific factors, and can further improve datasets that have been extensively normalized via an optimized, parameter rich model. We also show how the optimal parameters of DataRemix can be found efficiently by Thompson Sampling with a dual learning setup, making the approach feasible for computationally expensive objectives such as eQTL analysis.

Result

The DataRemix framework

We formulate DataRemix as a simple parametrized version of SVD which can be directly optimized to improve the biological utility of gene expression data. SVD decomposition can be thought of as a solution to the low-rank matrix approximations problem defined as:

$$\min_{U_k,\Sigma_k,V_k} \left\|X - U_k\Sigma_k V_k^T\right\|_F^2 \qquad (1)$$

where U and V are unitary matrices. Given a gene-by-sample matrix X and its SVD decomposition $U\Sigma V^T$ the product of k-truncated matricies $U_k\Sigma_k V_k^T$ gives the rank-k approximation of X. We introduce addition parameters p and μ to define a new reconstruction:

$$\text{DataRemix}_{\{k,p,\mu\}}(X) = U_k\Sigma_k^p V_k^T + \mu(X - U_k\Sigma_k V_k^T) \qquad (2)$$

Here, k is the number of principle components of SVD and $p \in [-1, 1]$ is a real number which alters the scaling of each eigenvalue. For $p = 1$, this approach reduces to the original SVD-based reconstruction . For $p = 0$ the transformation gives the frequently used whitening operation (Friedman 1987). As depicted in Figure 1, generally, different choices of p reweigh the contribution of each variance component, possibly making some low-variance biological signals visible while down-weighting technical and other systematic noise. The parameter μ is a non-negative weight that adds the residual back to the reconstruction in order to make the transformation *lossless*.

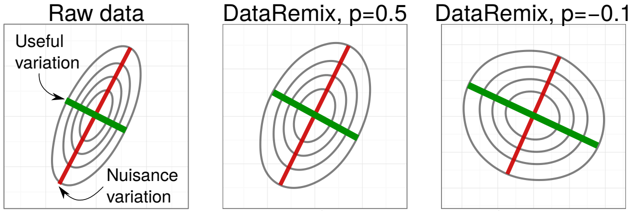


Figure 1: Visual representation of DataRemix transformation. We simulate a 2-dimensional dataset where the nuisance variation contributes more variance than useful variation. Different power parameters p reweigh the contributions of the two variance axes, making the useful variation more "visible".

Intuitively, we expect this approach to succeed because sophisticated normalization methods that use both data structure and some external variables, such as technical covariates, can be thought of as implicit regularizations on the naive SVD-based normalization (which simply removes the first k components), and this method simply makes this explicit.

Parameter Optimization

The parameters $\lambda = (k, p, \mu)$ need to be optimized with respect to a particular biological objective. Grid search and random search (Bergstra & Bengio 2012) are among the most popular strategies, but these methods have low efficiency. Most of the search steps are wasted and the optimally of parameters is highly constrained by the step size and available computing power. In order to utilize the search history and keep a good balance between exploration and exploitation, we can formulate parameter search as a dual learning task.

We define a general performance measure $y = L(\lambda, \mathcal{D})$, with λ representing the parameter tuple (k, p, μ), $\mathcal{D}$ as the data, L as the evaluating process and y as the biological objective. Ideally we can figure out the optimal point $\text{argmax}_\lambda L$ easily by gradient descent based method, but usually L is derivative-free and it is time intensive. Thus we introduce a surrogate model $f(\lambda)$ which can directly predict $L(\lambda, \mathcal{D})$ only given λ and there are two expects on f: $\text{argmax}_\lambda f$ should be easy to solve and f should have enough capacity.

With these two properties, we can sequentially update f with (λ_t, y_t) and propose to evaluate L at $\lambda_{t+1} = \text{argmax}_\lambda f$ in the next step. By gradually updating f with newly evaluated samples (λ, y), $\text{argmax}_\lambda f$ approaches the true underlying optimal $\text{argmax}_\lambda L$ as f can gradually fit to the underlying mapping function L. This provides a more efficient approach to explore the parameter space by exploiting the search history. In this work, we model f as a sample from a Gaussian Process with mean 0 and kernel $k(\lambda, \lambda')$, where $\lambda = (k, p, \mu)^T$. It is well known that the form of the kernel has considerable effect on performance. After experimentation we settled on the exponential kernel as the most suited for our application. The exponential kernel is defined as bellow (note the difference from the squared-exponential or RBF kernel).

$$k(\lambda, \lambda') = \exp\left(-\frac{\|\lambda - \lambda'\|_2}{2}\right) \tag{3}$$

We observe $y_t = f(\lambda_t) + \epsilon_t$, where $\epsilon_t \sim N(0, \sigma^2)$. For Bayesian optimization, one approach for picking the next point to sample is to utilize acquisition functions (Snoek *et al.*. 2012) which are defined such that high acquisitions correspond to potentially improved performance. An alternative approach is the Thompson Sampling approach (Basu & Ghosh 2017; Agrawal & Goyal 2013; Hernández-Lobato *et al.*. 2014). After we update the the posterior distribution $P(f|\lambda_{1:t}, y_{1:t})$, we draw one *sample* f from this posterior distribution as the optimization target to infer λ_{t+1}. Theoretically it is guaranteed that λ_t converges to the optimal point gradually. With this theoretical guarantee, we focus on Thompson Sampling approach to optimize parameters for DataRemix.

Estimation of Hyper-Parameters

Firsrt we rely on the maximum likelihood estimation (MLE) to infer the variance of noise σ^2 (Rasmussen 2004). Given the marginal likelihood defined by (4), it is easy to use any gradient descent method to determine the optimal σ^2

$$\begin{aligned}\log p(\vec{y}|\vec{\lambda}) = &-\frac{1}{2}\vec{y}^T(K + \sigma^2 I)^{-1}\vec{y} - \frac{1}{2}\log|K + \sigma^2 I| \\ &- \frac{t}{2}\log 2\pi\end{aligned} \tag{4}$$

where $\vec{y} = y_{1:t} = (y_1, \ldots, y_t)^T$,$\vec{\lambda} = \lambda_{1:t} = (\lambda_1, \ldots, \lambda_t)^T$ and $K_{ij} = k(\lambda_i, \lambda_j)$.

Sampling from the Posterior Distribution

Since Gaussian Process can be viewed as Bayesian linear regression with infinitely many basis functions $\phi_0(\lambda), \phi_1(\lambda), \ldots$ given a certain kernel (Rasmussen 2004), in order to construct an analytic formulation for the sample f, first we need to construct a certain set of basis functions $\Phi(\lambda) = (\phi_0(\lambda), \phi_1(\lambda), \ldots)$, which is also defined as feature map of the given kernel. Then we can write the kernel $k(\lambda, \lambda')$ as the inner product $\Phi(\lambda)^T\Phi(\lambda')$.

Mercer's theorem guarantees that we can express the kernels in terms of eigenvalues and eigenfunctions, but unfortunately there is no analytic solution given the exponential kernel we used. Instead we make use of the random Fourier features to construct an approximate feature map (Rahimi & Recht 2008). First we compute the Fourier transform p of the kernel (see Supplemental Note for derivation).

$$\begin{aligned}p(\vec{\omega}) &= \frac{1}{(2\pi)^3}\int \exp(-i\vec{\omega}^T\vec{\Delta})\exp\left(-\frac{\left\|\vec{\Delta}\right\|_2}{2}\right)d\vec{\Delta} \\ &= \frac{8}{\pi^2(4\|\vec{\omega}\|_2^2 + 1)^2}\end{aligned} \tag{5}$$

where $\vec{\omega} = (\omega_1, \omega_2, \omega_3)^T$ and $\vec{\Delta} = \lambda - \lambda'$. Then we draw m_t iid samples $\omega_1, \ldots, \omega_{m_t} \in \mathbb{R}^3$ by rejection sampling with $p(\omega)$ as the probability distribution. Also we draw m_t iid samples $b_1, \ldots, b_{m_t} \in \mathbb{R}$ from the uniform distribution on $[0, 2\pi]$. Then the feature map is defined by the following equation.

$$\Phi(\lambda) = \sqrt{\frac{2}{m_t}}[\cos(\omega_1^T\lambda + b_1), \ldots, \cos(\omega_{m_t}^T\lambda + b_{m_t})]^T \tag{6}$$

where the dimension m_t can be chosen to achieve the desired level of accuracy with respect to the difference between true kernel values $k(\lambda, \lambda')$ and the approximation $\Phi(\lambda)^T\Phi(\lambda')$.

Thompson Sampling

Any sample f from the Gaussian Process can be defined by $f(\lambda) = \Phi(\lambda)^T\theta$, where $\theta \sim N(0, I)$ and $\Phi(\lambda)^T$ is defined by (6). In order to draw a posterior sample f, we just need to draw a random sample θ from the posterior distribution $P(\theta|\vec{\lambda}, \vec{y})$.

$$\begin{aligned}P(\theta|\vec{\lambda}, \vec{y}) &\propto P(\vec{y}|\vec{\lambda}, \theta)P(\theta) \\ &\propto N(A^{-1}\Phi(\vec{\lambda})\vec{y}, \sigma^2 A^{-1})\end{aligned} \tag{7}$$

where $A = \Phi(\vec{\lambda})\Phi(\vec{\lambda})^T + \sigma^2 I$ and $\Phi(\vec{\lambda}) = (\Phi(\lambda_1) \cdots \Phi(\lambda_t))$ (see Supplemental Note for more details). The overall algorithm is summarized as the following pseudo code.

Algorithm 1 Thompson Sampling for Searching λ

```
Extra Parameters
$t_{max}$: the maximum number of iteration steps
$\xi$: a pre-defined probability which ensures the search doesn't stuck in the local optimum

1. Get a short sequence $\mathcal{D}_1 = (\lambda, y)$ as seeds by random search.
2. Draw $m_t$ iid samples $\omega_1, \ldots, \omega_{m_t} \in \mathbb{R}^3$ and $m_t$ iid samples $b_1, \ldots, b_{m_t} \in \mathbb{R}$ according to (5)
3. Iterate from $t = 1$ until $\lambda$ converges or it reaches $t_{max}$
    (1) At step $t$, estimate the hyper-parameter $\sigma^2$ given $\mathcal{D}_t$ according to (4)
    (2) Draw a sample $f$ given $\mathcal{D}_t$ according to (7) with feature map determined by (6)
    (3) $\lambda_{t+1} = \begin{cases} \text{argmax}_\lambda f(\lambda) & \text{w.p. } 1-\xi \\ \text{random search} & \text{w.p. } \xi \end{cases}$
    (4) Evaluate $y_{t+1}$ given $\lambda_{t+1}$
    (5) $\mathcal{D}_{t+1} = \mathcal{D}_t \bigcup (\lambda_{t+1}, y_{t+1})$
```

Quality of the correlation network derived from the GTex gene expression study.

The GTex datasets (Lonsdale *et al.*. 2013) is comprised of human samples from diverse tissues, many of which were obtained post-mortem and there are many technical factors which have considerable effects on the gene expression measurements. On the other hand this rich dataset provides an unprecedented multi-tissue map of gene regulatory networks and has been extensively analyzed in this context. It is natural to assume that a dataset that is better at recovering known pathways is likely to yield more credible novel predictions. Thus, we use DataRemix to optimize the known pathway recovery task as a function of the correlation network computed on a Remixed dataset.

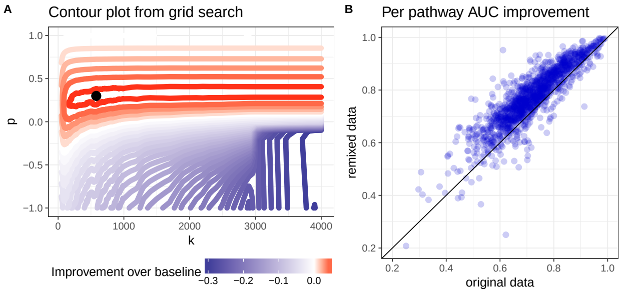


Figure 2: **A** The improvement in performance of DataRemix transform of the pathway prediction task visualized as a function of k and p parameters (μ is fixed at 0.01). Performance is measured as the mean AUC across all pathways in the "canonical" mSigDB dataset and the red contours indicate improvement over the performance on untransformed data. **B** Per-pathway performance improvement for the optimal DataRemix transformation.

Specifically we start with a quantile normalized TPM data that has not been corrected for technical factors or tissue of origin. We formally define the objective as the average AUC across "canonical" mSigDB pathways (which include KEGG, Reactome and PID) (Subramanian *et al.*. 2005) using guilt-by-association. Specifically, the genes are ranked by their average Pearson correlation to other genes in the pathway (excluding the gene when the gene itself is a pathway member). Figure 2A depicts the results of grid search for the parameters k, p (with μ fixed at 0.01) and the contour plot shows a clear region of increased performance. Using the optimal transformation found by grid search, we plot per-pathway AUC improvement in Figure 2B and find that the AUC is substantially increased for almost every pathway.

eQTL discovery in the DGN dataset.

We also consider the task of discovering *cis*- and *trans*-eQTLs on the Depression Gene Networks (DGN) dataset (Battle *et al.*. 2014). In the original analysis this dataset was normalized using the Hidden Covariates with Prior (HCP) (Mostafavi *et al.*. 2013) with four free parameters that were separately optimized for *cis*- and *trans*-eQTLs. The rationale behind seperate *cis* and *trans* optimized normalization can be understood in terms of which variance components represent useful vs. nuisance variation in the two contexts. Specifically, *cis*-eQTLs represent *direct* effects of genetic variation on the expression of a single gene. On the other hand, *trans*-eQTLs represent network level, *indirect* effects that are mediated by a regulator. Thus, *trans*-eQTLs are reflected in systematic variation in the data which becomes a nuisance factor when only direct effects are of interest. It thus follows that the data should be more aggressively normalized for *cis*-eQTL discovery. The original analysis of this dataset optimized the HCP parameters separately for the *cis* and *trans* tasks yielding two different datasets that we refer to as $D_{cis-optim}$ and $D_{trans-optim}$.

The HCP model takes various technical covariates as input, and of the covariates used in the original study 20 cannot be inferred from the gene-level counts. In order to investigate how much improvement can be achieved via DataRemix in the absence of access to these covariates we also consider a "naively" normalized dataset, quantile normalization of log-transformed counts, or D_{QN}.

cis-eQTLs.

In this task we focus on optimizing the discovery of *cis*-eQTLs. We define *cis*-eQTLs as a SNP-gene interaction where the SNP locates within 50kb of the gene's transcription start site. The interaction is quantified with Spearman rank correlation and deemed significant at 10% FDR (Benjamini-Hochberg correction for the total number of tests).

We perform our analysis in a cross-validation framework, whereby we can optimize DataRemix parameters (using grid search or Thompson Sampling) using SNPs on the odd chromosomes only and then evaluate the parameters on the held-out even chromosome set. We visualize the effect of varying the k and p parameters on the performance of the DataRemix transform in Figure 3. Red regions indicate improvement over the number of *cis*-eQTLs discovered with the $D_{cis-optim}$ dataset. We find that both versions of the dataset can be improved via the DataRemix transform to a similar degree. We also find that on this task the optimal p parameter is negative and the result is relatively insensitive to the choice of k. The last observation can be interpreted when we consider the interaction between p and μ (the multiplier for the residual part including $k+1$ through $\max(k)$ components). If we wish to bring forward small-variance components, as is the case with *cis*-eQTL discovery, we would like the diagonal values of $\mu\Sigma_{k+1:\text{rank of X}}$, representing the contribution of the later components, to be in the same range or larger than $\max(\Sigma^{p}_{1:k})$ which is the largest contribution of the high variance components. This can be achieved by picking different values of k.

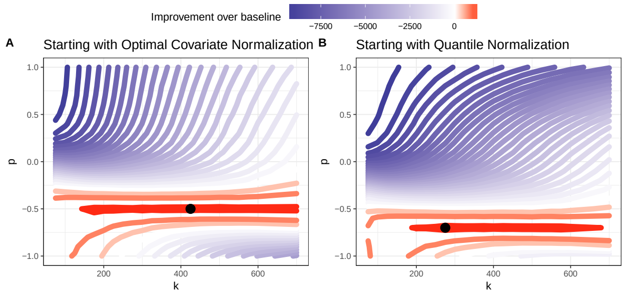


Figure 3: Contour plot representing the effects of the k and p parameters on the performance of DataRemix on *cis*-eQTL discovery on 50,000 randomly selected SNPs on odd chromosomes (training set). Red contours represent parameter combinations that increase the number *cis*-eQTLs beyond what can be achieved using the $D_{cis-optim}$ dataset. Panel A shows the results starting with $D_{cis-optim}$ while D_{QN} is used for panel B. Improvement can be achieved starting with either datasets. We note that the optimal p parameter is negative (though slightly different) for both datasets.

The final results for both the train and test set are depicted in Figure 4. We find that the optimal parameters are indeed generalizable as we achieve a similar level of improvement on the train and test datasets. Importantly, we find that while the quantile-normalized dataset D_{QN} performs considerably worse that $D_{cis-optim}$ the two datasets achieve comparable performance after applying DataRemix. Moreover, the final performance of the Remixed D_{QN} dataset is an improvement of the baseline $D_{cis-optim}$ demonstrating the near optimal normalization is possible without access to technical covariates. We do note, on this task, the final performance of the Remixed $D_{cis-optim}$ is slightly better than that of D_{QN} and thus it is still advisable to include such covariates in the normalization pipeline if they are available.

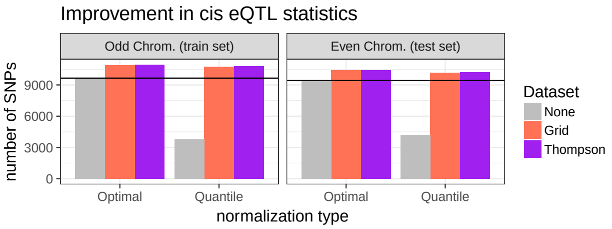


Figure 4: Final results from DataRemix parameter search using a cross-validation framework. Optimal parameters are determined using the odd chromosome SNPs only and then tested on the even chromosome SNPs. We find that the DataRemix transform does not overfit the objective as the degree of improvement is similar across the test and train SNP sets (note: the starting value of the baseline (DataRemix=”None”) datasets differ between the test and train SNP set). Moreover, we find that Thompson Sampling is able to match grid search results using only only 100 evaluations.

trans-eQTLs.

In our third task, we optimize the discovery of *trans*-eQTLs in the same DGN dataset. Ideally, *trans*-eQTLs represent network-level effects and thus give some insight about the regulatory structure of gene expression. However, in practice *trans*-eQTLs are simply defined as SNP-gene associations where the SNP and the gene are located on different chromosomes. While this is a useful heuristic definition, but it doesn't guarantee that the association is mediated at the network level. One possible source of bias is mis-mapped RNAseq reads which contaminate the quantification of the apparently *trans*-associated gene with reads from a homologous locus that has *cis* association. Even in the absence of technical artifacts, direct interchromsomal interactions have been observed (see Williams *et al.*. 2010 for a comprehensive review). In order to focus on potential indirect effects, we apply an additional filter to *trans*-eQTL discovery. Specifically we require SNPs involved in a *trans* effect to be associated with more than one gene at a FDR of 20% (Benjamini-Hochberg correction for the total number of test (approximately 8×10^9). We term these SNPs *trans*-SNPs$^+$. In comparison with same chromosome *cis*-eQTLs, inter-chromosome *trans*-eQTLs are rare and *trans*-SNPs$^+$ (as defined above) are more rare still. In fact, using the odd chromosome SNPs subsampled at 20%, we find only 88 such SNPs using $D_{trans-optim}$ dataset and this is the default value we wish to improve.

As is the case with *cis*-eQTLs, we investigate the k, p performance surface of the DataRemix transform at the grid-search optimal $\mu = 0.01$. Given that the relevant variance components that would maximize the *trans*-eQTL objective are different, it is not surprising that we find that the performance surface differs as well. In particular, we find that the optimum value of p is positive but close to 0 and thus the first k variance components are weighted equally with a weight close to 1. Consequently, at $\mu = 0.01$ and $p \approx 0$ the contribution of the first k components is considerably larger than that of the remaining ones and we find that the performance is more sensitive to the exact value of k.

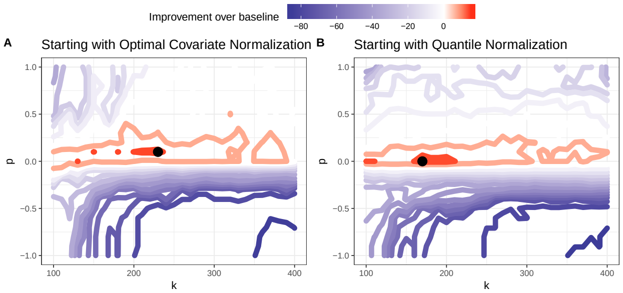


Figure 5: Contour plot representing the effects of the k and p parameters on the performance of DataRemix on *trans*-eQTL discovery on 50,000 randomly selected SNPs on odd chromosomes (training set). Red contours represent parameter combinations that increase the number of *trans*-eQTLs beyond what can be achieved using the $D_{trans-optim}$ dataset. Panel A shows the results starting with $D_{trans-optim}$ while D_{QN} is used for panel B. Improvement can be achieved starting with either datasets. We note that the performance is more sensitive to the choice of k.

Despite the difference in the performance landscape, we find that the DataRemix transform behaves similarly on this objective. Specifically, either starting dataset can be improved to similar final performance, though the optimal parameters are slightly different. As is the case with the *cis*-eQTL objective, the cross-validation procedure gives consistent results and no overfitting is observed for either grid search or Thompson Sampling (Figure 6).

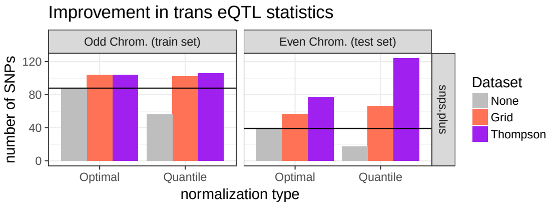


Figure 6: Final values for the eQTL statistics obtained from two versions of datasets. Here we make a comparison between quantile normalized D_{QN} and HCP normalized $D_{trans-optim}$ with parameters optimized for *trans*-eQTL discovery. We find DataRemix is able to improve upon either of starting datasets and the improvement on both the train and test dataset are comparable which indicates that overfitting is not a problem

Since *trans*-eQTLs are likely to reflect pathway level effects, we expect that a dataset that is optimally transformed for *trans*-eQTL discovery should also produce better correlation networks. We thus investigate if optimal DataRemix transform is transferable between tasks by checking if Remixed dataset optimized with respect to *trans*-eQTL discovery also improves the network quality criterion. Similar to our analysis of the GTex datasets, we use the correlation network to perform guilt-by-association pathway predictions and evaluate the results over 1,330 MSigDB canonical pathways. Figure 7 shows scatter plots of per-pathway AUPR (area under precision-recall curve) for several comparisons with respect to the baseline $D_{trans-optim}$ dataset. In the first panel we contrast the performance to D_{QN} and we observe that $D_{trans-optim}$ brings a considerable improvement over the quantile normalized dataset. In the second panel we contrast $D_{trans-optim}$ with the Remixed version of D_{QN} (optimized for *trans*-eQTL discovery with Thompson Sampling). We find that the pattern becomes opposite and the Remixed D_{QN} dataset performs consistently better that $D_{trans-optim}$. The final panel shows the results of Remixing $D_{trans-optim}$ itself which also improves the performance. Overall, we find that DataRemix improves multiple criteria of biological validity as optimizing for the *trans*-eQTL objective also results in improved correlation networks. Interestingly, we find that while the Remixed $D_{trans-optim}$ is no better than Remixed D_{QN} on *trans*-eQTL discovery, it performs slightly better on the pathway prediction task. Taking the two objectives into account, we conclude that starting with a properly covariate-normalized dataset is superior overall, which is also the our finding regarding the *cis*-eQTL objective.

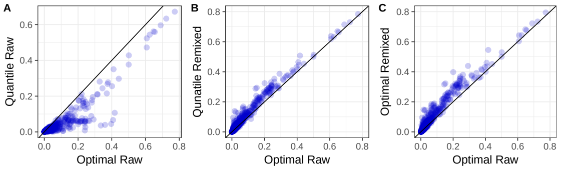


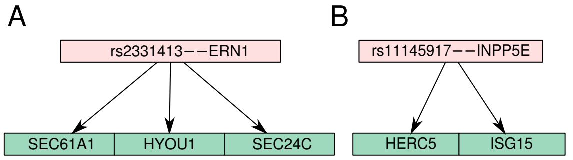
Figure 7: DataRemix-transformed datasets improve the pathway prediction objective which is not explicitly optimized. Each plot is a per-pathway AUPR (area under precision-recall curve) from various datasets (y-axis) contrasted with the results from the optimal covariate-normalized dataset $D_{trans-optim}$, which serves as the baseline (x-axis). Panel A shows the contrast between $D_{trans-optim}$ and D_{QN}. The performance of $D_{trans-optim}$ is considerably better. Panel B shows the results of the Remixed D_{QN} datasets (optimized for *trans*-eQTL discovery with Thompson Sampling). Even though D_{QN} starts out as considerably worse, the Remixed version is able to outperform $D_{trans-optim}$. Panel C shows the results of Remixed $D_{trans-optim}$. We choose to use AUPR instead of AUC because we find that Remixed version matches but doesn't further improve the AUC performance of $D_{trans-optim}$

A major finding of our study is that for the eQTL and pathway prediction tasks, the starting point of normalizing DGN datasets appears to matter relatively little. Even though the quantile-normalized dataset performs considerably worse in the beginning, after Remixing its performance matches that of the optimal covariate-normalized datasets. Of course, if covariates are available, it is preferable to use them and in the case of DGN, slightly further improvement can be achieved. However our results indicate that in some cases datasets *can* be effectively normalized even in the absence of meta-data about quality control or batch variables which is an important consideration for many legacy datasets where such information is not available.

Novel Biological Findings

At the optimal DateRemix parameters for D_{QN}, we find an additional 24 loci that have significant associations with more than one gene and are not in linkage disequilibrium with those significant hits in the $D_{trans-optim}$. We highlight two examples of new regulatory modules recovered via DataRemix that appear to be biologically credible based on the known functions of the genes involved. One of the newly significant interactions involves the SNP rs2331413 located in proximity of the ERN1 gene, which functions as a sensor of unfolded protein in the endoplasmic reticulum and triggers an intracellular signalling pathway termed the unfolded protein response. Three downstream genes associated with rs2331413 are likewise endoplasmic reticulum proteins. The ERN1 locus has been associated with several phenotypes in GWAS studies, most notably drug induced hepatotoxicity (Petros *et al.*. 2017).

We also find an SNP rs11145917 located near INPP5E gene which is associated with two genes in the alpha interferon response. Even though only two genes show genome-wide significance, several other canonical members of the alpha interferon response are just slightly short of the significance threshold suggesting that the locus affects the upstream signaling components. The INPP5E locus has been implicated in a variety of autoimmune diseases as well as blood immune-cell composition phenotype (de Lange *et al.*. 2017; Astle *et al.*. 2016), though to our knowledge no mechanism has been proposed. Our analysis suggests that INPP5E may affect baseline activity of the alpha interferon pathway, which is a testable prediction with potential clinical importance.



ERN1	endoplasmic reticulum to nucleus signaling 1
SEC61A1	Sec61 translocon alpha 1 subunit
HYOU1	hypoxia up-regulated 1
SEC24C	SEC24 homolog C, COPII coat complex component
INPP5E	inositol polyphosphate-5-phosphatase E
HERC5	E3 ubiquitin protein ligase , alpha interferon induced
ISG15	ISG15 ubiquitin-like modifier, alpha interferon induced

Figure 8: Clusters of *trans*-eQTLs detected by DataRemix that were not significant in the original dataset. Panel A. Both the *cis* and *trans* genes are involved in ER biology and specifically unfolded protein response. Panel B. Both of the *trans* genes are canonical targets of alpha interferon. The upstream *cis* gene, INPP5E, is a signaling molecule that mediates cell responses to various stimulation and its locus has been implicated in a variety of autoimmune diseases as well as blood immune-cell composition phenotypes.

Thompson Sampling Performance

We find that Thompson Sampling matches the best grid-search performance in under 100 steps giving a 40-fold reduction in the number of evaluations. We also note that it is possible for the Thompson sampling to surpass the grid-search results since the parameter combinations are not constrained by the choice of grid.

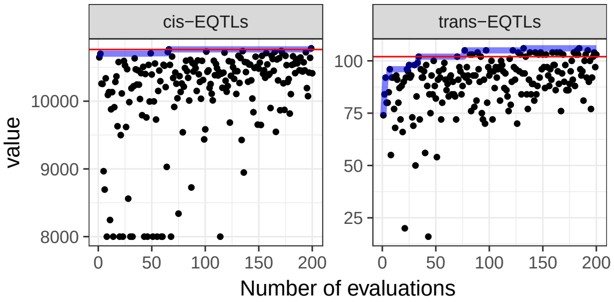


Figure 9: Objective evaluations as a function of iteration number for the *trans*-eQTL and *cis*-eQTL objectives using the quantile normalized D_{QN} dataset. Red lines indicate the maximum value that was obtained by grid-search and blue lines indicate the cumulative maximum of Thompson Sampling.

Discussion

We have proposed DataRemix, a new optimizable transformation for gene expression data. The transformation is able to improve the biological validity of gene expression representations and can be used for effective normalization in the absence of any knowledge of technical covariates. One limitation of the DataRemix approach is that it works best on data that is well approximated by a single Gaussian. However, it is relatively straightforward to adapt the approach to matrix decompositions different from SVD that are more suitable for non-Gaussian data, such as independent component analysis. We also note that it is possible to introduce additional parameters that specify more complex weighting schemes. However, as the number of parameters is increased, there is a potential for over-optimization of a specific objective above others. We emphasize that in our simple parametrization, we observe that multiple metrics of biological validity improve when only one is explicitly optimized. Specifically we find that optimizing for *trans*-eQTL discovery also improves the correlation network as measured by guilt-by-association pathway prediction. This property is less likely to be preserved as the number of parameters is increased.

Methods

GTex Dataset

We downloaded the complete gene-level TPM data (RNASeQCv1.1.8) from the GTex consortium (Lonsdale *et al.*. 2013). These data were quantile normalized.

DGN Dataset

Depression Gene Networks (DGN) dataset contains whole-blood RNA-seq and genotype data from 922 individuals. The genotype data was filtered for MAF>0.05. The genomic coordinate of each SNP was taken from the Ensembl Variation database (version 90, hg19/GRCh37). SNP identifiers that were not present in that release were excluded. After filtering there were 649,875 autosomal single nucleotide polymorphisms (SNPs). Data is available upon application through NIMH Center for Collaborative Genomic Studies on Mental Disorders. For gene expression we used the gene-level quantified dataset. The dataset comes already filtered for expressed genes and was further filtered for gene symbols that were not present in Ensembl 90 leaving 13,708 genes. The dataset comes in two covariate normalized versions with normalization parameters optimized for *cis*- and *trans*-eQTL discovery separately. To create the naive-normalized dataset, we applied a log transformation, $log(x+1)$, to the raw counts and quantile normalized the results.

eQTL mapping

eQTL association mapping was quantified with Spearman rank correlation. For *cis*-eQTLs, testing was limited to SNPs which locate within 50kb of any of the gene's transcription start sites (Ensembl, version 90). *cis*-eQTl is deemed significant at 10% FDR with Benjamini-Hochberg correction for the total number of tests. For *trans*-eQTLs, the significance cutoff is 20% FDR with Benjamini-Hochberg correction for the total number of tests. Since the Benjamini-Hochberg FDR is a function of the entire p-value distribution in order to ensure consistency comparisons, the rejection level was set once based on the p-value that corresponded to 10% or 20% FDR in the original *cis*-optimized $D_{cis-optim}$ and *trans*-optimized $D_{trans-optim}$ dataset respectively. To reduce the computational cost of grid evaluations, all the optimization computations were performed on a set of 100,000 subsampled SNPs.

Correlation network evaluation

We evaluated the quality of the correlation network derived from a particular dataset using guilt-by-association pathway prediction. Specifically, the genes were ranked by their average Pearson correlation to other genes in the pathway (excluding the gene when the gene itself is a pathway member). The resulting ranking was evaluated for performance using AUC or AUPR metric. For pathway ground-truth we used the "canonical" pathways dataset from MSigDB, comprising 1,330 pathways (Subramanian *et al.*. 2005).

Software Access

DataRemix is an R package which is freely available at GitHub (https://github.com/wgmao/DataRemix).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Text for this section . . .

Acknowledgements

Text for this section . . .

Author details

[1]Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology,. [2]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh,.

References

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, **3**: 1724–1735.

Stegle O, Parts L, Durbin R, Winn J. 2010. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, **6**: e1000770.

Listgarten J, Kadie C, Schadt EE, Heckerman D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A*, **107**: 16465–16470.

Kang HM, Ye C, Eskin E. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**: 1909–1925.

Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, Koller D. 2013. Normalizing rna-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One*, **8**: e68141.

Battle A, *et al.*. 2014. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome Res*, **24**: 14–24.

Alter O, *et al.*. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, **97**: 10101–10106.

Alter O, Brown PO, Botstein D. 2003. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences*, **100**: 3351–3356.

Mostafavi S, *et al.*. 2013. Normalizing rna-sequencing data by modeling hidden covariates with prior knowledge. *PLoS One*, **8**: e68141.

Friedman JH. 1987. Exploratory projection pursuit. *Journal of the American statistical association*, **82**: 249–266.

Bergstra J, Bengio Y. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, **13**: 281–305.

Snoek J, Larochelle H, Adams RP. 2012. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, S. 2951–2959.

Basu K, Ghosh S. 2017. Analysis of thompson sampling for gaussian process optimization in the bandit setting. *arXiv preprint arXiv:1705.06808*.

Agrawal S, Goyal N. 2013. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, S. 127–135.

Hernández-Lobato JM, Hoffman MW, Ghahramani Z. 2014. Predictive entropy search for efficient global optimization of black-box functions. In *Advances in neural information processing systems*, S. 918–926.

Rasmussen CE. 2004. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, S. 63–71. Springer.

Rahimi A, Recht B. 2008. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, S. 1177–1184.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, *et al.*. 2013. The genotype-tissue expression (gtex) project. *Nature genetics*, **45**: 580–585.

Subramanian A, Tamayo P, *et al.*. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**: 15545–15550.

Williams A, Spilianakis CG, Flavell RA. 2010. Interchromosomal association and gene regulation in trans. *Trends in genetics*, **26**: 188–197.

Petros Z, Lee MTM, Takahashi A, Zhang Y, Yimer G, Habtewold A, Schuppe-Koistinen I, Mushiroda T, Makonnen E, Kubo M, *et al.*. 2017. Genome-wide association and replication study of hepatotoxicity induced by antiretrovirals alone or with concomitant anti-tuberculosis drugs. *Omics: a journal of integrative biology*, **21**: 207–216.

de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, Jostins L, Rice DL, Gutierrez-Achury J, Ji SG, *et al.*. 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature genetics*, **49**: 256.

Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, Riveros-Mckay F, Kostadima MA, *et al.*. 2016. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*, **167**: 1415–1429.