BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 2 2015, pages 209–215
doi:10.1093/bioinformatics/btu518

Gene expression

Advance Access publication September 29, 2014

Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation

Marcus A. Badgeley[1], Stuart C. Sealfon[1] and Maria D. Chikina[2,*]

[1]Department of Neurology, Mount Sinai School of Medicine, New York, NY 10029 and [2]Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, PA 15260, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Modern molecular technologies allow the collection of large amounts of high-throughput data on the functional attributes of genes. Often multiple technologies and study designs are used to address the same biological question such as which genes are over-expressed in a specific disease state. Consequently, there is considerable interest in methods that can integrate across datasets to present a unified set of predictions.

Results: An important aspect of data integration is being able to account for the fact that datasets may differ in how accurately they capture the biological signal of interest. While many methods to address this problem exist, they always rely either on dataset internal statistics, which reflect data structure and not necessarily biological relevance, or external gold standards, which may not always be available. We present a new rank aggregation method for data integration that requires neither external standards nor internal statistics but relies on Bayesian reasoning to assess dataset relevance. We demonstrate that our method outperforms established techniques and significantly improves the predictive power of rank-based aggregations. We show that our method, which does not require an external gold standard, provides reliable estimates of dataset relevance and allows the same set of data to be integrated differently depending on the specific signal of interest.

Availability: The method is implemented in R and is freely available at http://www.pitt.edu/~mchikina/BIRRA/

Contact: mchikina@pitt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 21, 2014; revised on June 16, 2014; accepted on July 24, 2014

1 INTRODUCTION

The availability of high-throughput experimental technologies has greatly increased the rate of data generation. However, the reliability of individual high-throughput datasets is often uncertain, and the development of techniques to integrate the results across different studies to improve the validity and reproducibility of findings has been an important area of research.

In samples studying disease-related expression changes, for example, sample size is a common limitation, and data integration by statistical meta-analysis has been used repeatedly to boost statistical power and find reliable signals. Formal meta-analyses rely on the statistical properties of each independent dataset, and common strategies involve combining effect sizes and *P*-values [see (Tseng *et al.*, 2012) for a comprehensive review]. However, the fundamental limitation of this approach is that statistical properties may not correlate with how well a particular biological question is addressed. A single study may show large effect sizes and small *P*-values but can be biologically invalid owing to design flaws. For example, a study purporting to find large expression differences in ethnic groups (Spielman *et al.*, 2007) was later shown to be confounded by batch effect (Akey *et al.*, 2007).

An alternative data integration approach that is robust to discordance between statistical power and biological validity is Bayesian integration with an external gold standard. This approach does not use any internal dataset statistics and has the additional advantage that it can be used to perform integration tasks that cannot even in principle be addressed with statistical meta analysis, such as integrating data from completely unrelated experimental platforms. In the Bayesian integration approach, the accuracy of each dataset is measured in terms of how well it is able to reproduce known information in the gold standard. For example, in integrating genomic datasets for functional network predictions, each dataset is evaluated against a gold standard of functionally interacting proteins (Huttenhower *et al.*, 2009; Lee *et al.*, 2004). In this case, datasets only contribute a set of final predictions (which may be categorical or real valued) whose accuracy is judged independently of the underlying data structure.

The limitation of the Bayesian approach is that it requires the existence of an acceptable gold standard against which individual predictions are evaluated. For some integration tasks, such as disease gene expression studies, such a standard is typically not available. Moreover, even well-established standards of knowledge, such as the Gene Ontology or KEGG, are only limited to capturing accuracy relative to what is already known. Thus, if a dataset captures correct but completely novel biological information, its value is not recognized by the comparison with a gold standard lacking any sampling of this signal, and its contribution will be discarded.

A practical alternative to formal statistical meta-analysis or Bayesian analysis is combining datasets directly without any treatment of the underlying statistics or any reference to the external standard. Such methods can be applied in a variety of contexts where statistics and external standards may be unavailable or prone to producing misleading quality assessments. In this integration approach, each dataset contributes equally to the consensus signature. Without loss of generality, the approach

*To whom correspondence should be addressed.

can be viewed as an instance of rank aggregation where each dataset supplies a prediction list (that is not necessarily complete) in order of differential expression, correlation or some other metric of relevance. Unfortunately treating all datasets equally can make these methods unacceptably sensitive to noise and outliers.

To address these limitations, we propose a new data integration method that combines the rank aggregation framework with the power of Bayesian reasoning. Our method assumes that there is a ground truth to be discovered and this is represented with varying fidelity in the datasets to be integrated. Starting with an equally weighted consensus generated by rank averaging, we iteratively fit Bayes Factors to each dataset and recompute the integration result. Thus, this approach combines the advantages of being agnostic to dataset-specific statistics and external standards while using the robustness of the Bayesian framework to handle outliers and inconsistencies. We demonstrate that this approach has favorable theoretical properties on an established simulation pipe-line and improves over other rank aggregation methods in real data integration tasks. The method is computationally simple, yet significantly increases the predictive power of rank-based aggregations. Additionally it provides a reliable estimate of dataset relevance and can be used to produce context-sensitive integrations that focus in on specific areas of interest. Because it requires only ranked lists, it is readily applicable to many integration problems.

2 APPROACH

Our method is inspired by Latent Class Analysis (LCA), a statistical technique widely used in social science and medical diagnostics to analyze a series of related measurements by assuming that they capture an underlying latent trait. For example, in the absence of definitive knowledge regarding disease status, LCA can be used to infer the sensitivities and specificities of a series of diagnostic tests. In essence, the technique works by consensus, if multiple diagnostic tests report a positive result the individual is more likely to have the disease. The entire combination of diagnostic test-specific false-positive and false-negative rates and the disease status probability of individual patients can thus be estimated by maximum likelihood.

While the underlying problem appears to be similar to genomic data integration and a related approach has been successful in integrating binary interaction networks (Weile *et al.*, 2012), the methodology is not directly transferable to arbitrary genomic data. LCA operates on categorical data, while genomic data are usually continuous valued. Moreover, the assumption of the existence of a few discrete latent classes is often incorrect and genomic datasets can be extremely discordant, resulting in many solutions with similar likelihoods. For example, in a case of large disagreements, solutions dictating that one dataset is right and all others are wrong are all equally valid. We address these limitations with several heuristic techniques, which give good results in practice. Our methods works by iteratively computing an aggregated ranking from dataset-specific Bayes factors and using the ranking to update the Bayes factor calculation. We refer to our method as Bayesian Iterative Robust Rank Aggregation (or BIRRA).

While the method can be applied to any data integration task, for the purpose of illustration we can assume that the input to our method is a set of datasets that rank genes in order of disease relevance, for example, a set of differential expression studies. If we had a representative set of genes that were known to be differentially expressed in the disease, we could infer the probability of disease relevance (DR) for the rest of the genes given the set of evidence (gene's ranks within a dataset) $E_{1...n}$ via the standard Bayesian approach

$$P(DR|E_1, E_2, \ldots E_n) = C \cdot \prod_{i=1}^{n} P(E_i|DR) \quad (1)$$

Where C is a normalization constant and $P(E_i|DR)$ is calculated from the set of known positive examples. This approach naturally captures the fact that individual datasets vary in how well they are able to discriminate disease-relevant genes and their relative contribution to the final ranking should be adjusted accordingly. Our method uses the same principle without relying on a set of positive examples to calculate $P(E_i|DR)$ but rather estimating it from the structure of the data.

Require: input data to be integrated for n genes, rank ordered by some measure of relevance (such as differential expression P-values or correlation). Missing values in each dataset are assigned the lowest rank P a prior probability for the 'positive class' B number of bins.

1: Compute initial aggregate ranking by mean ranks

2: Compile a working standard by assigning the top np of aggregated predictions to be the positive class and the rest of the predictions to the negative class

3: Discretize the data into B bins

4: **repeat**

5: Compute bin-wise cumulative Bayes factors for each dataset by comparing with the working standard

6: Aggregate the data by naive Bayesian integration according to Equation 1

7: Update working standard with new rankings obtained in step 6

8: **until** rankings unchanged or max iterations exceeded

9: **return** Final ranking

In principle, for complex genomic datasets, this kind of procedure can be unstable because in cases of large disagreements or low signal, multiple solutions are equally valid. We implement several heuristic techniques to make this approach more robust. In particular, we compute Bayes factors cumulatively starting from the top bin and enforce a monotone decrease. Thus, the minimum Bayes factor is limited by the prior probability, and each piece of data can only contribute positive evidence. Secondly, we use a winzorization step whereby for each bin we decrease the maximum Bayes factor to that of the second maximum. This ensures that at least two datasets are always combined and prevents the appearance of singleton solutions whereby only one datasets is predicted to have any signal.

While we find that our experiments with real data were robust to variations in bin size, we intentionally leave this as free parameters because the optimal choice is likely to be data dependent. Because the algorithm proceeds entirely by bootstrapping any noise in Bayes factor, estimation may be propagated and thus choice of bin size must strike a balance between using all the available signal while minimizing the influence of random fluctuations. We find that for integrations of data from mammalian genomic datasets, a choice of 50 bins was appropriate. Our implementation includes a plotting functionality that tracks the progress of the Bayes factor estimates (as in Fig. 2) and can be used to investigate the influence of different binning. We also find that in all our analyses, using real data convergence was achieved in $\leq$10 steps (which is the default number of maximum iterations).

3 RESULTS

3.1 Simulations

While using a Bayesian framework, our approach is still most accurately classified as rank aggregation because it requires only ranked lists and the conditional probabilities are bootstrapped from the data itself. Rank aggregation methods are attractive data integration tools because they require no prior knowledge and are thus widely applicable and easy to implement. However, they are fundamentally limited by their ability to handle noise and outliers because there is no mechanism to prevent a large number of noisy datasets from overwhelming the results. We expect that the Bayesian framework should perform well in this regard and we evaluate the theoretical properties of our method with a simulation strategy similar to that outlined in Kolde *et al.* (2012).

We simulate a differential expression meta-analysis with 1000 genes where 5% of the genes are truly differentially expressed and the differential expression is assayed in independent experiments. To create a single dataset, we draw values form a $\mathcal{N}(1,1)$ distribution for the differentially expressed genes, and for the remaining 95% of background genes, we drawn values from a $\mathcal{N}(0,1)$ distribution. The final values are subsequently rank transformed. Using these distribution parameters produces individual datasets that rank the true differentially expressed genes with an average Area Under the Curve (AUC) of 0.75 and we refer to these as 'signal' datasets. We also introduce 'noise' datasets where the genes are ranked randomly.

We use the complete set of signal and noise datasets to evaluate different rank aggregation approaches [mean ranks, Robust Rank Aggregation (RRA) (Kolde *et al.*, 2012), Stuart (Stuart *et al.*, 2003), and our method BIRRA]. Aggregating 10 signal and 30 noise datasets with several different methods we find that even on a dataset with 75% uninformative data all methods tested produce results that outperform individual signal datasets (Fig. 1) and the computationally sophisticated RRA and Stuart methods surpass mean ranks at low sensitivity. However, RRA outputs a *P*-value of 1 for almost 80% of the genes and thus has no discriminative power at higher recall, producing a total AUC that is lower than mean ranks. Overall, the more sophisticated methods improve only slightly if at all over the simple mean ranks, demonstrating that these methods provide little additional robustness. The BIRRA method, on the other hand, is specifically designed to address robustness to noise, as it explicitly estimates the information content of individual datasets down-weighting the uninformative datasets directly.

Figure 2 shows a sample run of the BIRRA method that highlights this key advantage of our approach. We plot the estimated log Bayes factors against bin number for each iteration of our algorithm and we observe that the pure noise datasets (gray lines) and signal datasets (red lines) separate quickly, with the noise datasets converging to lower Bayes factors. Of course, if we had a representative gold standard to evaluate the conditional probabilities, we would find that the noise datasets contain no information and the corresponding log Bayes factors should be 0. While in the absence of an independent standard the noise datasets are not discounted completely, they are still significantly down-weighted, producing a result whose performance falls between various single-step aggregation methods and optimal naive Bayes (Fig. 1).

Varying the number of non-informative datasets we find that BIRRA maintains an advantage across a large range of added noise and produces the most improvement for integration tasks that contain between 50 and 85% uninformative datasets (Fig. 3). BIRRA also performs comparably with other methods for values outside this range. When little or no noise is present,

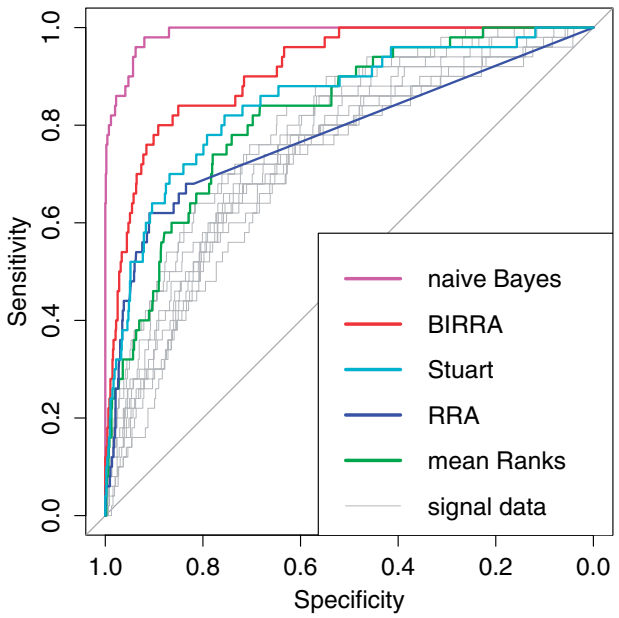

Fig. 1. We simulate a noisy rank aggregation task with two types of datasets. For 10 'signal' datasets, the values for 50 differentially expressed genes are drawn from a $\mathcal{N}(1,1)$ distribution, while the values for 950 background genes are drawn from a $\mathcal{N}(0,1)$ distribution. For 30 'noise' datasets, the values are drawn from the same distribution. We aggregate the data using different rank aggregation methods and compare the results to those obtained with an optimal naive Bayes (i.e. using the exact conditional distributions). The BIRRA algorithm outperforms other aggregation methods producing results that are between the optimal naive Bayes and established rank aggregation methods. AUC values: Mean Ranks 0.82, RRA 0.78, Stuart 0.85, BIRRA 0.91, Naive Bayes 0.99

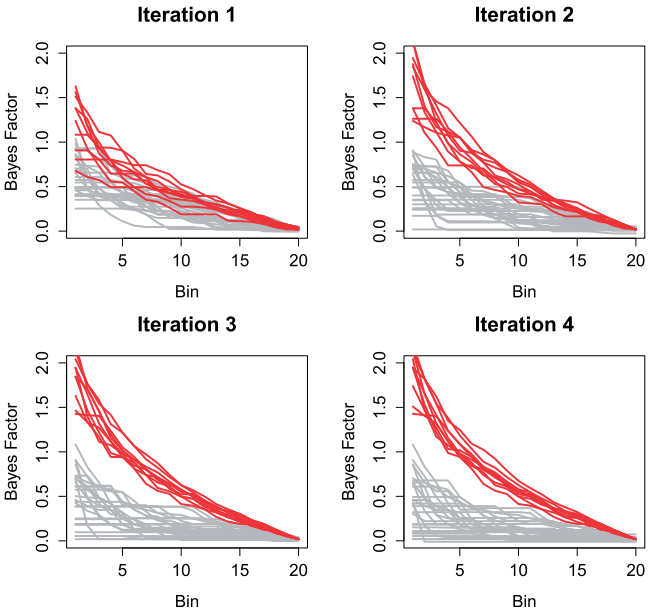

Fig. 2. Evolution of BIRRA computed Bayes factors for the dataset evaluated in Figure 1. At each iteration, the BIRRA algorithm computes dataset-specific Bayes factors against the current working standard. Bayes factors are plotted in black for the 'signal' datasets and in gray for the 'noise' datasets. In just a few iterations, BIRRA successfully down-weights the noise datasets

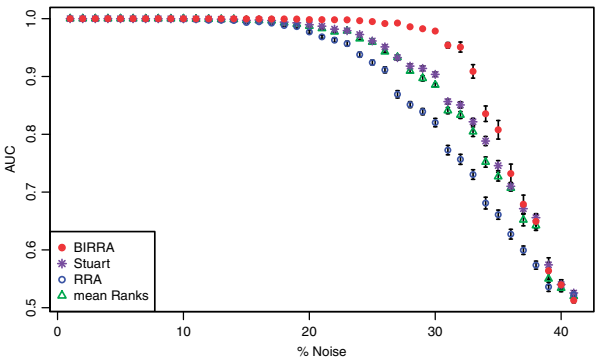

Fig. 3. Varying the fraction of 'noise' datasets we find that BIRRA provides the most performance gains when the fraction of uninformative datasets is large and it performs as well as other methods when all the datasets have signal

BIRRA is effectively equivalent to mean ranks, suggesting that it can be applied out-of-the-box to a variety of integration tasks.

3.2 Disease biomarker meta-analysis

One of the key motivations for the development of BIRRA is the need to integrate disease-related expression datasets. In these cases, it is difficult to evaluate performance, as typically no universally accepted standards of disease-related expression changes exist. However, one important source of validation can be reproducibility, and we can use this criterion to evaluate our method. We have compiled a set of 14 related gene expression datasets

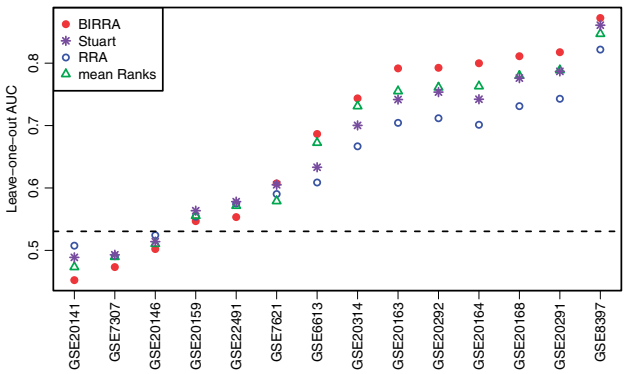

Fig. 4. Comparison of rank aggregation methods using a compendium of PD datasets. While there is no gold standard for gene expression changes associated with PD, we can judge the aggregation results based on how well the aggregated ranking reproduces the result of an independent study. To simulate this, we have taken a leave-one-out approach where we aggregate all but one of the studies and use the remaining study as a gold standard (top 500 genes are considered positive) to evaluate the aggregation results. We plot the resulting AUCs for different aggregation methods. The horizontal line represents the 99% confidence threshold for AUC being >0.5. We observe that leave-one-out aggregation is predictive for most of the datasets tested, resulting in significant AUCs. We also observe that in most cases BIRRA produces superior results. In particular, it is able to improve our ability to predict blood expression changes (GSE6613) from the remaining datasets that use brain tissue

that compare tissue from Parkinson's disease (PD) patients with controls. Given these datasets we can ask how well the results of a single one of them can be predicted by integrating the remaining 13. We evaluate our method alongside mean ranks and the more sophisticated RRA and Stuart rank aggregation methods.

Figure 4 shows that for a number of datasets their results could not be predicted by any of the rank aggregation methods, demonstrating that they capture entirely different signals. However, for those datasets for which mean rank aggregation successfully captures significant signal, BIRRA was able to improve on that result in 9 of 11 cases, producing the best overall performance of the four methods tested. Importantly, the two datasets that were better recapitulated with mean ranks than with BIRRA were difficult to predict with all methods, and the performance difference was small; on the other hand, the datasets for which integration was a good predictor (>0.7 AUC in mean ranks), our method produced a substantial improvement. We also investigated how well the relative ranking within the top 500 genes was predicted by different aggregation approaches and find that BIRRA showed an improvement on this metric as well (Supplementary Fig. S1).

An additional advantage of our BIRRA approach is that it not only outputs aggregated results but also the computed Bayes factors, which can serve as an alternative measure of dataset quality because these values capture how well each dataset reproduces the consensus signal. When datasets are combined by BIRRA, these Bayes factors determine the relative contribution of each. On the other hand, in formal statistical meta-analyses, the relative contribution is determined by the statistical properties of the underlying data. It is natural to ask how the statistical quantities that are intrinsic to each dataset and the BIRRA

quality estimates that are computed from the consensus are related. While there is no single metric that can describe dataset quality in a statistical sense, we focus on two parameters that would be of relevance in various meta-analysis approaches. Common meta-analysis strategies include directly merging raw data and combining effect sizes or *P*-values (Tseng *et al.*, 2012). Thus, one relevant parameter is the sample size of the study, which is an a priori measure of the expected statistical power. We also choose the maximum *P*-value for the top 50 differentially expressed genes to represent the actual observed magnitude of the overall disease effect. Plotting the sum of bin-wise Bayes factors against these statistical quantities we find that there is no relationship between these measures of intrinsic dataset quality and quality as judged by replication (Fig. 5).

For example, the dataset with the seemingly largest effect (GSE22491) had a poor BIRRA quality score, revealing that it produced differential expression rankings that were not supported by other datasets, while for the dataset that BIRRA judged to be most informative (GSE20291) the *P*-value for the first 50 DE genes was greater than would be expected from a uniform distribution. This analysis demonstrates the great discordance between intrinsic statistical properties and the ability of datasets to capture reproducible biological signal and underscores the potential pitfalls of relying exclusively on dataset statistics when performing meta-analysis.

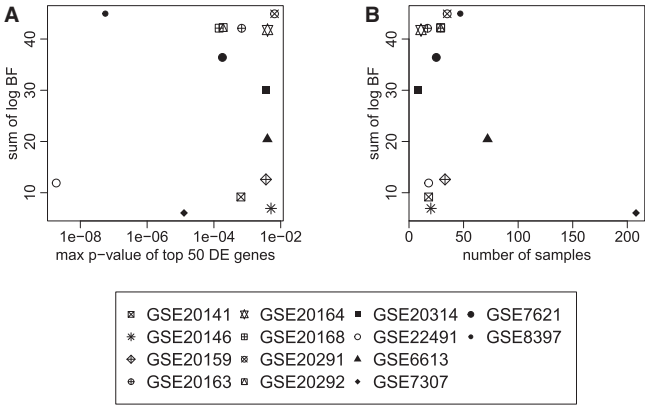

Fig. 5. Evaluating BIRRA estimated dataset relevance against dataset statistical properties. We plot the sum of log Bayes factors against maximum *P*-value of top 50 DE genes (**A**) and study sample size (**B**). We find that the BIRRA-estimated quality is unrelated to dataset intrinsic statistical properties

3.3 Predicting transcription factor-target networks

To evaluate how well our method performs in a real data integration task for which an external source of validation is available, we follow the approach described in Kolde *et al.* (2012). The goal of this integration task is to predict transcription factor (TF) targets from correlation between the TFs and other genes in several gene expression datasets. We perform the analysis for 14 TFs involved in embryonic stem-cell (ES) renewal and assayed by ChIPseq in Chen *et al.* (2008) and 12 ES related microarray datasets. We compute correlations between the TF gene and all other genes in each dataset, and the resulting ranked lists serve as the input to various rank aggregation methods. The individual datasets and results of four aggregation approaches (mean ranks, BIRRA, RRA and Stuart) are evaluated for correctness against the ChIPseq dataset that directly measures physical binding (Fig. 6). We find that on this integration task, both the Stuart and RRA methods typically outperformed mean ranks but not the best individual dataset. On the other hand, our method BIRRA outperformed the best dataset in 10 of the 14 cases, a substantial improvement over alternative rank aggregation techniques. Importantly, BIRRA showed the largest improvement in cases where the individual datasets varied most with respect to how well they were able to capture the relevant signal, for example, Myc and Smad1, demonstrating that BIRRA correctly identifies the more relevant datasets without relying on an external gold standard.

Using the independent gold standard of ChIPseq interactions we can not only evaluate the predictive power of our final rank aggregation but also the estimates of dataset quality that are a byproduct of our approach. We compare the BIRRA-estimated Bayes factors with the values obtained by directly plugging in the true-positive and false-positive counts obtained from the ChIPseq gold standard (as in standard naive Bayes analysis).

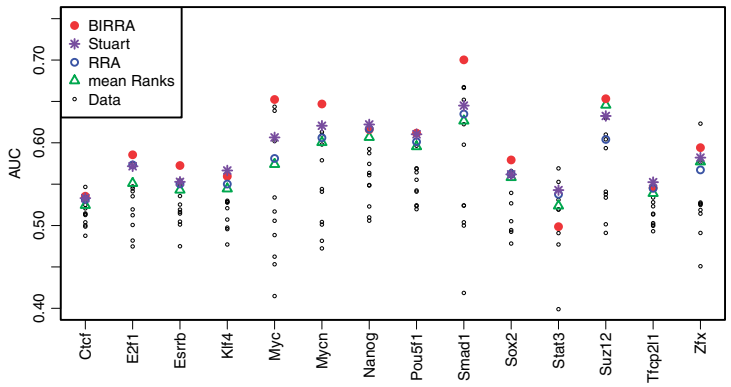

Fig. 6. TF targets predicted from expression correlation. We aggregate the results of 10 different ES expression datasets to predict TF target interaction and evaluate the result using a ChIPseq dataset. We find that for most TFs tested BIRRA aggregation produced outperformed other methods and the best individual dataset

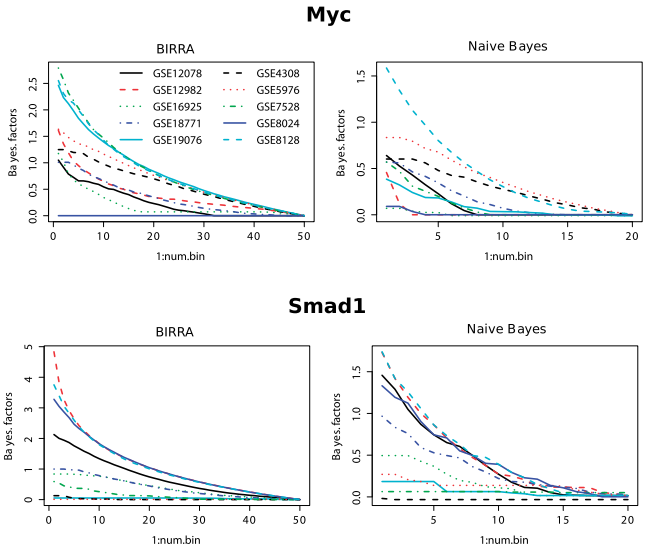

Fig. 7. Comparing BIRRA estimates to Bayes factors computed using the independent ChIPseq dataset we find that the dataset weights as estimated by BIRRA are in agreement with their ability to recapitulate the ChIPseq signal. Importantly, datasets varied widely in their ability to recapitulate the ChIPseq interactions for different TFs, and the BIRRA-computed Bayes factors were consistent in each case

We plot the dataset-specific Bayes factors derived from the two methods (BIRRA and naive Bayes) side by side in Figure 7 and find that the relative dataset quality as judged by the self-contained BIRRA approach is in fairly good agreement with that obtained by using an independent standard. Importantly the true dataset quality (as estimated by direct comparison with ChIPseq data) varied for different TFs, and BIRRA was able to correctly identify which specific datasets were informative in each case. For example, based on comparisons with the gold standard, GSE4308 (black dashed lines) performs well at predicting targets for Myc, yet is completely uninformative for Smad1, and this trend if fully captured by BIRRA (Fig. 7).

While the ChIPseq-based validation relies only on the correlations between TFs and other genes, we can use the BIRRA-computed Bayes factors to produce integrations over all gene pair correlations. Importantly, however, focusing on a specific TF produces different relevance estimates for the same dataset, and the relative contribution of each dataset to the final result varies substantially (Fig. 7). Consequently, when extended to the complete correlation network, our method produces different networks depending on the TFs of interest. Thus, unlike other rank integration methods that can only output a single result, BIRRA can produce context-sensitive rank aggregations that maximize the agreement over a specific subset of the biological signal.

4 DISCUSSION

We present a method that implements rank aggregation via Bayesian reasoning. Our method differs from previous rank aggregation approaches in that it directly estimates the reliability of individual datasets, weighing them accordingly. We have shown that the BIRRA method outperforms established approaches on simulations and performs well in real data integration scenarios. However, the method will likely not be applicable to all data integration tasks without some further developments. For example, as it operates entirely by bootstrapping, it is even more susceptible to deviations from the independence assumption underlying the naive Bayes calculations. In general, the Bayes factors and probabilities calculated by our method are always inflated relative to those that are calculated with an external gold standard, and the gap widens when dataset dependence is present. BIRRA performance on the stem cell expression integration, which has considerable dataset dependence, demonstrates that the method is effective despite the independence assumption. However, dependence will have to be addressed directly for integrations that involve a larger number of datasets. It is likely that in those cases, the deleterious effects of data dependence can be negated using regularization techniques commonly applied to standard-based Bayesian integration (Huttenhower *et al.*, 2009; Lee *et al.*, 2008), and finding a suitable regularization method will be an important direction for future research.

5 METHODS

5.1 Rank aggregation methods

The BIRRA method was programmed in R, and the source code is available in the supplement. The RRA and Stuart methods were called from the RobustRankAggreg R package.

5.2 TF target networks

Data for 14 ES-related studies were downloaded from InSilicoDB (Coletta *et al.*, 2012), which provides Robust Multi-array Average (RMA)-processed Affymetrix files mapped to gene symbols. As in (Kolde *et al.*, 2012) we created a standard for TF target interaction using the ChIPseq-determined TF-promoter mapping provided by (Chen *et al.*, 2008). Following (Kolde *et al.*, 2012), we considered all genes with a score of $\geq$0.6 as targets. We noted that for two of the TFs considered in the ChIPseq study, Smad1 and Suz12, the majority of individual expression datasets were predictive of their targets only when anti-correlation was considered, suggesting that these TFs act as repressors. Suz12 is known to be directly involved in gene repression (Cao and Zhang, 2004), and though Smad1 is not a dedicated repressor, in ESs it interacts with Nanog to repress its targets (Suzuki *et al.*, 2006). Therefore, for these two TFs, we ranked genes in order of anti-correlation. For the remaining TFs, correlation was used.

5.3 PD meta-analysis

We collected the majority of the datasets used in a previously reported Parkinson's meta-analysis (Zheng *et al.*, 2010). For Affymetrix datasets, we downloaded the data from InSilicoDB. For Illumina datasets, the data was taken from the Gene Expression Omnibus database (GEO) (Edgar *et al.*, 2002) and processed with background correction using the 'bg.adjust' function in the 'affy' Bioconductor package, log transformation and quantile normalization. Datasets were filtered by discarding the bottom 20% of low expression and low variance genes, and processed for differential expression with respect to disease status, including additional covariates, such as gender and brain region, whenever those were applicable and available in GEO.

Funding: Funded by NIH Contract HHSN272201000054C and a grant from the Michael J Fox Foundation.

Conflict of interest: none declared.

REFERENCES

Akey,J.M. *et al.* (2007) On the design and analysis of gene expression studies in human populations. *Nat. Genet.*, **39**, 807–808; author reply 808–809.

Cao,R. and Zhang,Y. (2004) SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol. Cell*, **15**, 57–67.

Chen,X. *et al.* (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**, 1106–1117.

Coletta,A. *et al.* (2012) InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. *Genome Biol.*, **13**, R104.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Huttenhower,C. *et al.* (2009) Exploring the human genome with functional maps. *Genome Res.*, **19**, 1093–1106.

Kolde,R. *et al.* (2012) Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, **28**, 573–580.

Lee,I. *et al.* (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.

Lee,I. *et al.* (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. *Nat. Genet.*, **40**, 181–188.

Spielman,R.S. *et al.* (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.*, **39**, 226–231.

Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.

Suzuki,A. *et al.* (2006) Nanog binds to Smad1 and blocks bone morphogenetic protein-induced differentiation of embryonic stem cells. *Proc. Natl Acad. Sci. USA*, **103**, 10294–10299.

Tseng,G.C. *et al.* (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, **40**, 3785–3799.

Weile,J. *et al.* (2012) Bayesian integration of networks without gold standards. *Bioinformatics*, **28**, 1495–1500.

Zheng,B. *et al.* (2010) PGC-1, a potential therapeutic target for early intervention in Parkinson's disease. *Sci. Transl. Med.*, **2**, 52ra73.