Tradict enables accurate prediction of eukaryotic transcriptional states from 100 marker genes

Surojit Biswas[1,*], Konstantin Kerner[2], Paulo José Pereira Lima Teixeira[3,4], Jeffery L. Dangl[3-7,†], Vladimir Jojic[8,†], Philip A. Wigge[9,†,*]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[2] Botanical Institute, Biocenter, University of Cologne, D-50674 Cologne, Germany.
[3] Howard Hughes Medical Institute and [4] Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.
[5] Carolina Center for Genome Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[6] Department of Microbiology and Immunology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[7] Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[8] Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[9] Sainsbury Laboratory, University of Cambridge, Cambridge CB2 1LR, UK

[†] Contributed equally
* Correspondence: surojitbiswas@g.harvard.edu, Philip.Wigge@slcu.cam.ac.uk

Abstract

Transcript levels are a critical determinant of the proteome and hence cellular function. Because the transcriptome is an outcome of the interactions between genes and their products, it may be accurately represented by a subset of transcript abundances. We developed a method, Tradict (<u>tra</u>nscriptome pre<u>dict</u>), capable of learning and using the expression measurements of a small subset of 100 marker genes to predict transcriptome-wide gene abundances and the expression of a comprehensive, but interpretable list of transcriptional programs that represent the major biological processes and pathways of the cell. By analyzing over 23,000 publicly available RNA-Seq datasets, we show that Tradict is robust to noise and accurate. Coupled with targeted RNA sequencing, Tradict may therefore enable simultaneous transcriptome-wide screening and mechanistic investigation at large scales.

Introduction

As the critical determinant of the proteome and therefore cellular status, the transcriptome represents a key node of regulation for all life[1]. Transcriptional control is managed by a finely tuned network of transcription factors that integrate environmental and developmental cues in order to actuate the appropriate responses in gene expression[2–4]. Importantly, the transcriptomic state space is constrained. Pareto efficiency constraints suggest that no gene expression profile or phenotype can be optimal for all tasks, and consequently, that some expression profiles or phenotypes must come at the expense of others[5,6]. Furthermore, across all major studied kingdoms of life, cellular networks demonstrate remarkably conserved scale-free properties that are topologically characterized by a small minority of highly connected regulatory nodes that link the remaining majority of sparsely connected nodes to the network[7–9]. These theories suggest that the effective dimension of the transcriptome should be far less than the total number of genes it contains. If true to a large enough extent, it may be possible to faithfully compress and prospectively summarize entire transcriptomes by measuring only a small, carefully chosen subset of it.
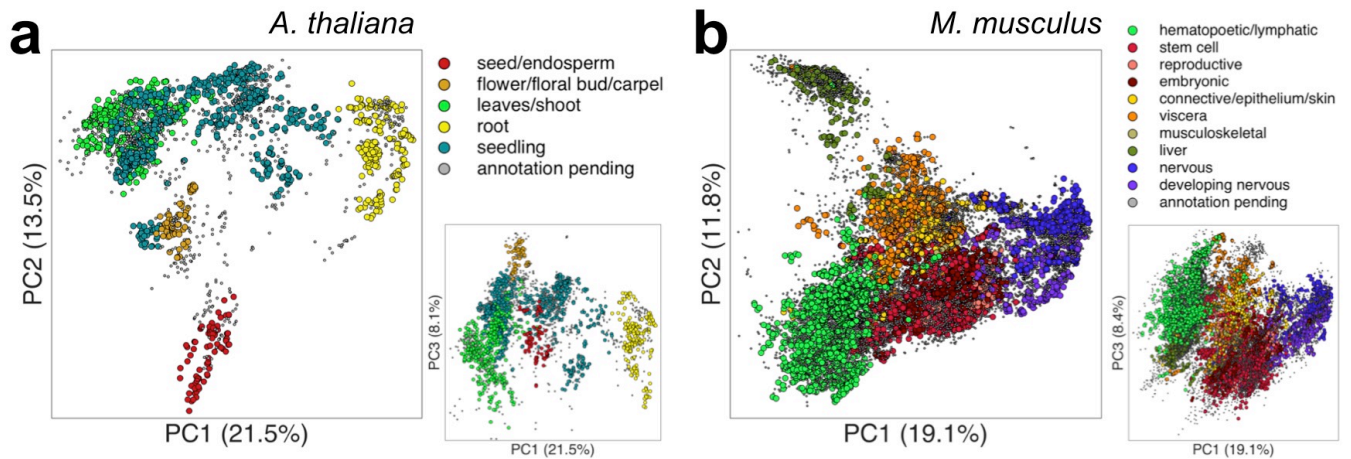
Figure 1. The primary drivers of variation in our training transcriptome collection are developmental stage and tissue. a) *A. thaliana*, b) *M. musculus*. Also shown are plots of PC3 vs. PC1 to provide another perspective.

51 Indeed, previous studies have exploited this reduced dimensionality to perform gene expression
52 imputation for microarray data for missing or corrupted values[10–12]. Others have extended these
53 intuitions to predict expression from probe sets containing a few hundred genes[13,14]. However,
54 prediction accuracies have been modest and usually limited to 4,000 target probes/genes. Recently,
55 several studies examined the transcriptomic information recoverable by shallow sequencing especially
56 as it applies to single-cell experiments[15–18]. Jaitin *et al.* (2014) and Pollen *et al.* (2014) demonstrated
57 that only tens of thousands of reads are required per cell to learn and classify cell types *ab initio*[16,18].
58 Heimberg *et al.* (2016) extended these findings and demonstrated that the major principal components
59 of a typically sequenced mouse bulk or single-cell expression dataset may be estimated with little error
60 at even 1% of the depth[15]. Though these approaches advance the notion of strategic transcriptome
61 undersampling, they only recover broad transcriptional states and are restricted to measuring only the
62 most abundant genes. During sample preparation -- arguably the most expensive cost of a multiplexed
63 sequencing experiment -- shallow sequencing-based approaches still utilize protocols meant for
64 sampling the entire transcriptome and therefore consume more resources than necessary.
65 Furthermore, given that the expression of even the most abundant genes is highly skewed, sequencing
66 effort is wastefully distributed compared to an approach that chooses which genes to measure more
67 wisely. Finally, it is still not clear from sample sizes and biological contexts previously studied whether
68 the low dimensionality of the transcriptome may be leveraged unconditionally (or nearly so) across
69 organism and application.
70 In this work, we introduce Tradict (<u>tra</u>nscriptome pre<u>dict</u>), a robust-to-noise and probabilistically
71 sound algorithm, for inferring gene abundances transcriptome-wide, and predicting the expression of a
72 transcriptomically comprehensive, but interpretable list of *transcriptional programs* that represent the
73 major biological processes and pathways of the cell. Tradict makes its predictions using only the
74 expression measurement of a single, context-independent, machine-learned subset of 100 marker
75 genes. Importantly, Tradict's predictions are formulated as posterior distributions over unmeasured
76 genes and programs, and therefore simultaneously provide point and credible interval estimates over
77 predicted expression. Using a representative sampling of over 23,000 publicly available, transcriptome-
78 wide RNA-Seq datasets for *Arabidopsis thaliana* and *Mus musculus*, we show Tradict prospectively
79 models program expression with striking accuracy. Our work demonstrates the development and large-
80 scale application of a probabilistically reasonable multivariate count/non-negative data model, and
81 highlights the power of directly modeling the expression of a comprehensive list of transcriptional
82 programs in a supervised manner. Consequently, we believe that Tradict, coupled with targeted RNA
83 sequencing[19–24], can rapidly illuminate biological mechanism and improve the time and cost of
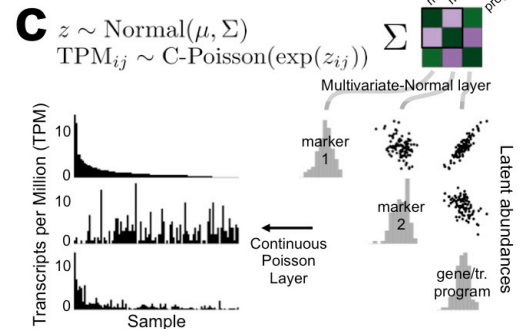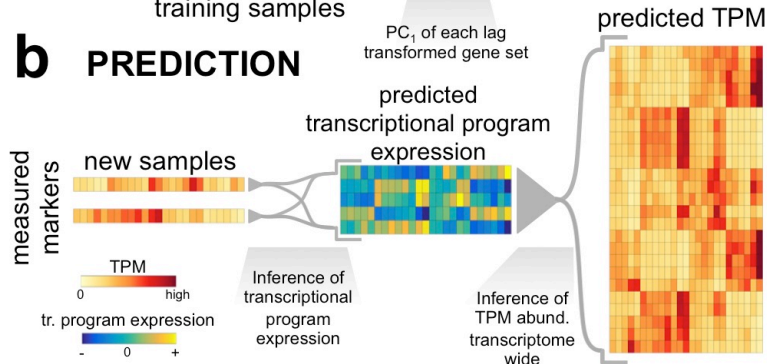84 performing large forward genetic, breeding, or chemogenomic screens.
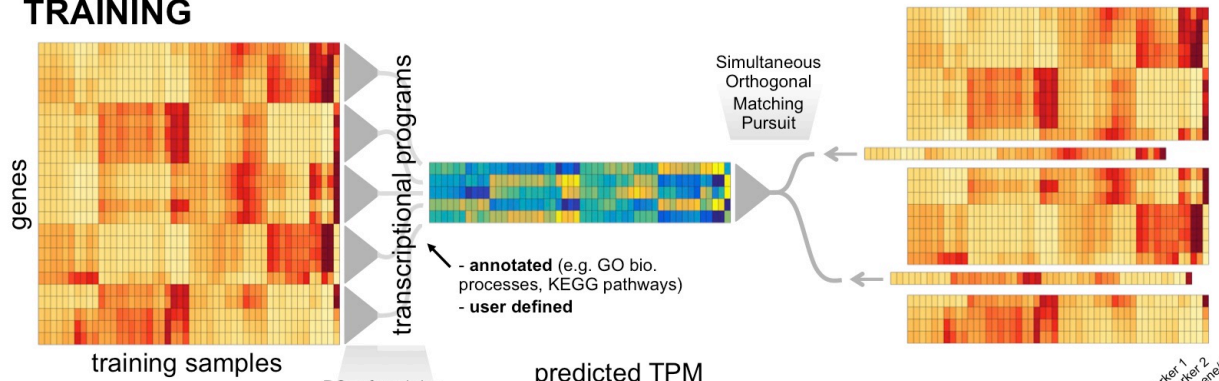
Figure 2. Tradict's algorithmic workflow. a) During training, the transcriptome is first quantitatively summarized in terms of a collection of a few hundred, biologically comprehensive transcriptional programs. These are then decomposed into a subset of marker genes using an adaptation of the Simultaneous Orthogonal Matching Pursuit algorithm. A Multivariate Normal Continuous-Poisson hierarchical model is used as a predictive model to capture covariance relationships between markers, transcriptional programs, and all genes. b) During prediction, Tradict predicts the expression of transcriptional programs and all genes in the transcriptome using the expression measurements of the marker genes. c) The Multivariate Normal Continuous-Poisson hierarchy enables Tradict to efficiently model statistical coupling between the non-negative expression measurements typical of sequencing experiments. This is done by assuming that associated with each observed, noisy TPM measurement, there is an unmeasured (denoised), latent abundance the logarithm of which comes from a Multivariate Normal distribution over all genes and transcriptional programs.

Results

Assembly of a comprehensive training collection of transcriptomes

We downloaded all available Illumina sequenced publicly deposited RNA-Seq samples (transcriptomes) on NCBI's Sequence Read Archive (SRA). Among samples with at least 4 million reads, we successfully downloaded and quantified the raw sequence data of 3,621 and 27,450 transcriptomes for *A. thaliana* and *M. musculus,* respectively. After stringent quality filtering, we retained 2,597 (71.7%) and 20,847 (76.0%) transcriptomes comprising 225 and 732 unique SRA submissions for *A. thaliana* and *M. musculus*, respectively. An SRA 'submission' consists of multiple, experimentally linked samples submitted concurrently by an individual or lab. We defined 21,277 (*A. thaliana*) and 21,176 (*M. musculus)* measurable genes with reproducibly detectable expression given our tolerated minimum sequencing depth and mapping rates (see Supplemental Information "Materials and Methods" for further information regarding data acquisition, transcript quantification, quality filtering, and expression filtering). We hereafter refer to the collection of quality and expression filtered transcriptomes as our *training transcriptome collection*.

In order to assess the quality and comprehensiveness of our training collection, we performed a deep characterization of the expression spaced spanned by these transcriptomes. We found that the transcriptome of both organisms was highly compressible and that the primary drivers of variation were

104   tissue and developmental stage (Figure 1a-b, Figure S1), with many biologically realistic trends within
105   each cluster (Supplemental Analysis 1). We additionally examined the distribution of submissions
106   across the expression space, compared inter-submission variability within and between tissues,
107   inspected expression correlations among genes with well-established regulatory relationships, and
108   assessed the evolution of the expression space across time. These investigations revealed our training
109   collection is of high and reproducible technical quality, reflective of known biology, and stable
110   (Supplemental Analysis 1, Figures S2-S4). Given additionally the diversity of tissues, genetic
111   perturbations, and environmental stimuli represented in the SRA, these results, taken together, suggest
112   that our training collection is an accurate and representative sampling of the transcriptomic state space
113   that is of experimental interest for both organisms.
114
Tradict - algorithm overview
116   Given a training transcriptome collection, Tradict encodes the transcriptome into a single subset
117   of globally representative *marker* genes and learns their predictive relationship to the expression of a
118   comprehensive collection of transcriptional programs (e.g. pathways, biological processes) and to the
119   rest of the genes in the transcriptome. Tradict's key innovation lies in using a Multivariate Normal
120   Continuous-Poisson (MVN-CP) hierarchical model to model marker latent abundances -- rather than
121   their measured, noisy abundances -- jointly with the expression of transcriptional programs and the
122   abundances of the remaining non-marker genes in the transcriptome. In so doing, Tradict is able to 1)
123   efficiently capture covariance structure within the non-negative, right-skewed output typical of
124   sequencing experiments, and 2) perform robust inference of gene set and non-marker expression even
125   in the presence of significant noise.
126   Figure 2 illustrates Tradict's general workflow. Estimates of expression are noisy, especially for
127   low to moderately expressed genes. Given samples are often explored unevenly and that the *a priori*
128   abundance of each gene differs, the level of noise in a gene's measured expression for a given sample
129   varies, but it can be estimated. Therefore, during training, Tradict first learns the log-latent, denoised
130   abundances for each gene in every sample in the training collection using the lag transformation[25]. It
131   then collapses this latent transcriptome into a collection of predefined, comprehensive collection of
132   *transcriptional programs* that represent the major processes and pathways of the cell related growth,
133   development, and response to the environment (Supplemental Tables 3-4). In this work, we focus on
134   creating a Gene Ontology derived panel of transcriptional programs, in which the first principal
135   component of all genes contained within an appropriately sized and representative GO term is used to
136   define an accordingly named transcriptional program[26,27]. The expression values of these programs are
137   then encoded using an adapted version of the Simultaneous Orthogonal Matching Pursuit algorithm
138   into a small subset of marker genes selected from the transcriptome[28,29]. Tradict finally stores the mean
139   and covariance relationships between the log-latent expression of the selected markers, the
140   transcriptional programs, and the log-latent expression of the remaining non-marker genes at the
141   Multivariate Normal layer of the MVN-CP hierarchical model (Figure 2a).
142   Prospectively, only the expression of these marker genes needs to be measured and the
143   expression of genes and/or transcriptional programs can be inferred as needed. During prediction,
144   Tradict uses the observed marker measurements as well as their log-latent mean and covariance
145   learned during training, to estimate -- via MCMC sampling -- the posterior distribution over the log-latent
146   abundances of the markers[30]. Though a simply a consequence of proper inference of our model, this
147   denoising step adds considerable robustness to Tradict's predictions. From this estimate, Tradict uses
148   covariance relationships learned during training to estimate the conditional posterior distributions over
149   the remaining non-marker genes and transcriptional programs (Figure 2b). From these distributions,
150   the user can derive point estimates (e.g. posterior mean or mode), as well as measures of confidence
151   (e.g. credible intervals). A complete description of the entire algorithm can be found in the "Tradict
152   algorithm" section of the Materials and Methods in the Supplemental Information.
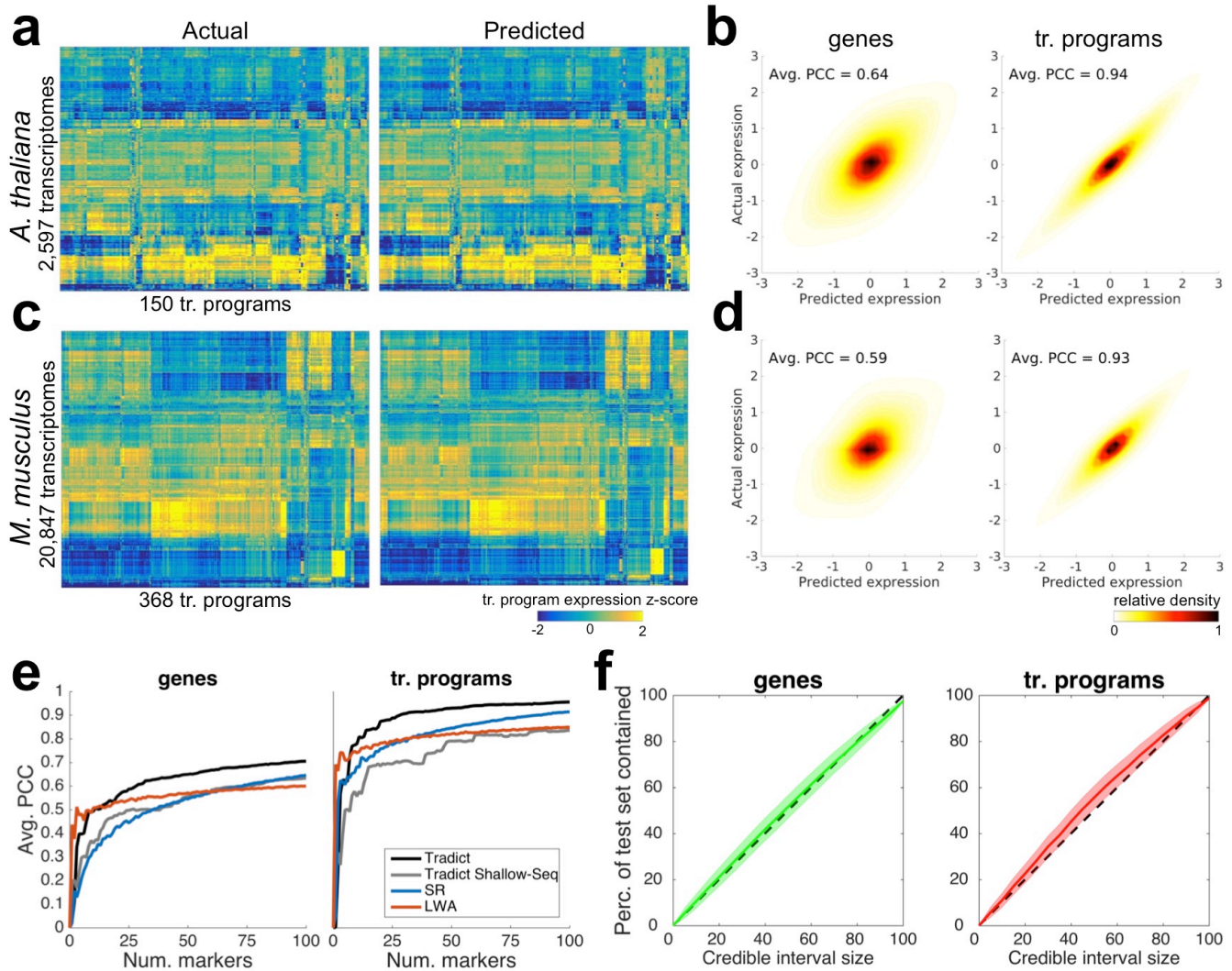153

154

Figure 3. Tradict prospectively predicts gene and transcriptional program expression with superior and robust accuracy. Tradict's prospective prediction accuracy during 20-fold cross validation of the entire training collection for both organisms. a) Heatmaps illustrating test-set reconstruction performance of all transcriptional programs for *A. thaliana*. Shown is the reconstruction performance for all samples in our transcriptome collection when they were in the test-set. b) Density plots of predicted vs. actual test-set expression for all genes (left) and transcriptional programs (right) for *A. thaliana*, after controlling for inter-sumbission biological signal. The intra-submission expression of each gene and transcriptional program was z-score transformed to make their expression comparable. c & d) Same as a & b, but for *M. musculus*. e) Comparison of Tradict's performance vs. several baselines: SR (structured regression), LWA (locally weighted averaging), and Tradict Shallow-Seq. f) A posterior predictive check illustrating the concordance between Tradict's posterior predictive distribution and the distribution of test-set expression values for genes (left) and transcriptional programs (right). Plotted is the percent of test-set observations contained within a credible interval vs. the size of the credible interval. A unique credible interval is derived for each gene/program. The "x=y" line is illustrated as a dotted black line. Shaded error bands depict the sampling distribution of this analysis across test-sets from a 20-fold cross validation on the *A. thaliana* dataset.

155
156 **Tradict prospectively predicts gene and transcriptional program expression with superior and**
157 **robust accuracy**
158   To understand Tradict's prospective predictive performance, we performed 20-fold cross
159 validation on the training transcriptome collections for both *A. thaliana* and *M. musculus* and evaluated
160 Pearson correlation coefficients (PCC) between predicted and actual expression for each fold when the
161 remaining 95% of folds were used for training. To make this experiment as reflective of reality as
162 possible, folds were divided by submission so that samples from the same set of experiments would not

163 appear both in training and test sets. Because submissions to the SRA span a broad array of biological
164 contexts, the total biological signal contained in any given test set likely exceeds that of what would be
165 expected for typical application, which in turn would lead to overly optimistic estimates of prediction
166 accuracy. We therefore evaluated *intra-submission accuracy*, in which PCC calculations are performed
167 on 'submission-adjusted' expression values. To do this, for each gene and program, each test-set
168 submission's mean expression was subtracted from the expression values for all samples associated
169 with that submission. In effect, this regresses out between-submission effects, and allows us to assess
170 Tradict's predictive performance one experiment at a time, as one would do in practice.
171   Figures 3a and 3c illustrate that the reconstruction performance for transcriptional programs in
172 both organisms is strikingly accurate across all collected submissions. Quantitatively speaking, the
173 average intra-submission PCCs for transcriptional programs are 0.94 and 0.93 for *A. thaliana* and *M.*
174 *musculus*, respectively. This is despite lower prediction performance on gene expression (Figures 3b
175 and 3d). Intuitively, this is because transcriptional programs are measured as linear combinations of the
176 log-latent TPMs of the genes that comprise them, effectively smoothing over the orthogonal noise
177 present in each gene's expression prediction.
178   We also found Tradict's performance to be superior to several baselines. These include two
179 successful approaches developed in Donner *et al.* (2012) for microarray[14], and a version of Tradict that
180 uses the 100 most abundant genes as its selected markers (Figure 3e, Figure S5, Supplemental
181 Analysis 2). The former baselines rely on structured regression (SR) and locally weighted averaging
182 (LWA), linear/parametric and non-parametric methods, respectively. The latter baseline examines the
183 utility of simple shallow sequencing by using the most abundant genes as markers for Tradict (see
184 Supplemental Analysis 2 for a more detailed description). Figure 3e illustrates test-set intra-submission
185 performance of each method as a function of the number of markers entered into the model. LWA
186 demonstrates the quickest performance gain, but then saturates after 10 markers. This is likely
187 because a non-linear kernel based approach makes the most efficient use of a few markers, but is
188 adversely impacted by the curse of dimensionality as more markers are added. The parametric
189 methods (Tradict, SR) navigate this dimensionality increase more efficiently and ultimately realize
190 better performance for a still reasonable number of markers. Tradict outperforms SR and Tradict
191 Shallow-Seq, ultimately obtaining a PCC between predicted and actual expression of 0.71 for genes
192 and 0.96 for transcriptional programs. This suggests Tradict's probabilistic framework is more
193 reasonable than SR's and that Tradict's marker selection is more optimal than picking the most
194 abundant genes. We additionally found Tradict's predictions were robust to noise in the form of low
195 sequencing depth and/or corrupt marker measurements (Figure S5 and S10, Supplemental Analysis 2
196 and 5), which we attribute to its probabilistic framework, in which training and prediction are performed
197 in the space of denoised latent abundances.
198   To further assess the validity of Tradict's modeling assumptions, we examined how Tradict's
199 posterior predictive distribution matched the distribution of test-set gene and program expression
200 values. Specifically, we performed a posterior predictive check in which we asked what percent of test-
201 set gene or program expression values fall within an X% credible interval, where a unique interval is
202 defined for each gene/program[30]. If Tradict's posterior predictive distribution is reasonable then X% of
203 the true expression values should fall within this interval for any X. Figure 3f illustrates the results of this
204 analysis as performed on disjoint test-sets from a 20-fold cross validation on the *A. thaliana* dataset. On
205 average, the X% credible interval captures X% of test-set observations for any choice of X. The
206 posterior predictive distribution for transcriptional programs may be slightly too wide at moderate
207 interval sizes (30-70%), which would make Tradict more conservative (higher type II error rate) than it
208 should be. However, in practice it is accurate (p-value = 0.24, *t*-test) for larger, more standard interval
209 sizes (e.g. 95%). We conclude that Tradict's probabilistic modeling assumptions capture unseen data
210 well.
211   We next characterized Tradict's limitations through error, power, program annotation robustness
212 analyses, and a timing and memory analysis (Supplemental Analysis 3-4). These analyses revealed
213 that training-set expression variance and mean abundance correlated positively with both program and

214   gene expression prediction performance (Supplemental Analysis 3.I, Supplemental Tables 3-4).
215   Combined with program size as another predictor, these variables could account for most of the error
216   (60% of total variance) in program expression prediction. A power analysis revealed that for both *A.*
217   *thaliana* and *M. musculus*, 1,000 samples -- comprising approximately 100 submissions -- is sufficient
218   for optimum performance (Supplemental Analysis 3.II). An examination of how gene-to-program mis-
219   annotation rates influence predictive performance revealed that program expression prediction
220   perfromance was robust up to a 20% mis-annotation rate and that gene expression prediction
221   performance was completely robust to any level of mis-annotation. The latter result is a consequence of
222   a statistical decoupling between gene and program expression prediction (Supplemental Analysis 3.III).
223   Finally, Tradict's training time and peak memory requirements scaled linearly with training set size, and
224   increased 0.26 seconds and 1.1 Mb per sample. Prediction time was limited by MCMC sampling from
225   the conditional posterior distributions of gene and program expression, and required 3.1 seconds per
226   sample on average (Supplemental Analysis 4). Taken together, we conclude that the causes of
227   Tradict's errors are well understood and intuitive, and that Tradict's sample requirements are
228   reasonable, especially for major model organisms (Supplemental Table 5, Supplemental Analysis 3.II).
229   Furthermore, Tradict is robust to noisily defined transcriptional programs, and its computational
230   requirements scale well to large datasets.
231
232   **Case studies reveal the power of predicting and studying transcriptional program expression**
233   To demonstrate how Tradict may be applied in practice, we focused on two case studies related
234   to innate immune signaling -- one performed using bulk *A. thaliana* seedlings (detailed below), and the
235   other using primary immune *M. musculus* cell lines (detailed in Supplemental Analysis 5; Figure S10).
236   We trained Tradict on our full collection of training transcriptomes for each organism to produce two
237   organism-specific Tradict models. Each was based on the selection of 100 markers learned from the
238   full training transcriptome collection (Supplemental Tables 7-8) that we assert are globally
239   representative, and context-independent. The case study samples do not, of course, appear in the
240   collection of training transcriptomes.
241
242   **Tradict accurately predicts temporal expression patterns for a diverse panel of *A. thaliana***
243   **immune signaling mutants under different hormone perturbations -** The hormones salicylic acid
244   (SA) and jasmonic acid (JA) play a major, predominantly antagonistic regulatory role in the activation of
245   plant defense responses to pathogens. Yang *et al.* (2016) investigated the effect of a transgenically
246   expressed bacterial effector, HopBB1, on immune signaling in *A. thaliana*[31]. In their study, they
247   performed a time course experiment, treating plants with MeJA (a JA response inducer), BTH (an SA
248   mimic and SA response inducer), or mock buffer and monitored the transcriptome of bulk seedlings at 0
249   hr, 1 hr, 5 hr, and 8 hr post treatment. These experiments included several immune signaling mutants
250   with differing degrees of response efficiency to MeJA and BTH treatment. Among other findings, they
251   conclude that HopBB1 enhances the JA response, thereby repressing the SA response and facilitating
252   biotrophic pathogen infection.
253   We asked to what extent strategic undersampling of the transcriptome and application of Tradict
254   could quantitatively recapitulate the findings of Yang *et al.* (2016). Given Tradict's near perfect accuracy
255   on predicting the expression of transcriptional programs, we took a top down, but hypothesis driven
256   approach to our analysis which first examined the expression of all transcriptional programs. Figure 4a
257   illustrates the actual and predicted expression of all transcriptional programs in *A. thaliana* as a function
258   of time and treatment. Here, Tradict reconstructs the expression of all transcriptional programs with an
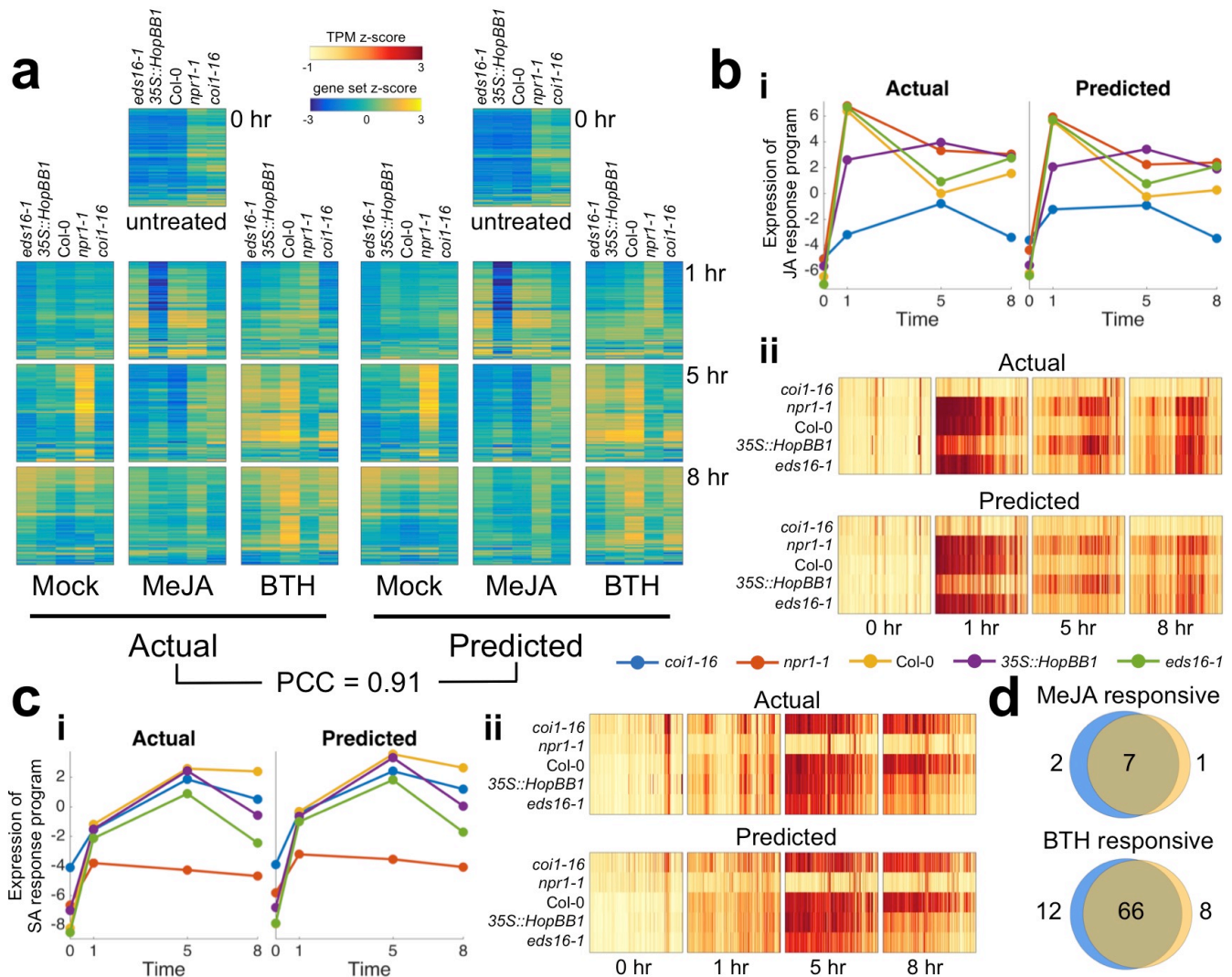259   average PCC of 0.91.

Figure 4. Tradict accurately predicts transcriptional responses across time in response to hormone perturbation in an A. thaliana innate immune signaling dataset. After being trained on the full *A. thaliana* training transcriptome collection, the selected set of 100 globally representative and context-independent markers were used to predict the expression of transcriptional programs and all genes for the transcriptomes presented in Yang *et al.* (2016). a) Actual vs. predicted heatmaps for the expression of all 150 transcriptional programs in *A. thaliana* across genotype, time, and hormone treatment. b) Predicted vs. actual expression of i) the JA response transcriptional program, and ii) the genes involved in the JA response program. c i-ii) Same as b, but for the SA response transcriptional program. d) Hypothesis free, differential transcriptional program expression analysis as performed on the actual expression of transcriptional programs vs those predicted by Tradict. Blue circles represent the actual and orange represent the predicted. All heatmaps are clustered in the same order across time, treatment, genotype, and between predicted and actual.

260   Recall that the genes participating in each of our transcriptional programs are pre-defined, in
261 this work, by a carefully chosen, interpretable, but maximally representative set of GO biological
262 processes. Therefore, given the goals of this study, we next examined the expression of the "response
263 to jasmonic acid" and "response to salicylic acid" transcriptional programs. Figure 4b shows the
264 expression behavior for the "response to jasmonic acid" transcriptional program across all the
265 genotypes and time points upon MeJA treatment. More specifically, part (i) shows that the predicted
266 expression and actual expression are qualitatively and quantitatively in agreement, both in magnitude
267 and rank across the different genotypes. For example, as expected, *coi1-16*, which cannot sense JA,
268 does not respond to the MeJA stimulus, while wildtype Col-0 does. However, even more subtle
269 expression dynamics are captured by Tradict's predictions. For example, *eds16-1* and *npr1-1* -- slightly

270 and strongly impaired SA responders, respectively -- are slightly and strongly hyper-responsive to
271 MeJA, respectively -- just as expected from the JA-SA antagonism. Furthermore, as demonstrated in
272 Yang *et al.* (2016), the *35S::HopBB1* transgenic line exhibits a prolonged and sustained JA response
273 for both the actual and predicted expression for this transcriptional program. Part (ii) of Figure 4b
274 illustrates the expression of all the MeJA responsive genes in this transcriptional program. Again
275 Tradict's predictions are in good agreement with actuality, achieving a PCC of 0.72, and it's visually
276 clear that the expression magnitude of these genes positively correlates with the registered expression
277 magnitude of the "response to jasmonic acid" transcriptional program. Figure 4c parts (i) and (ii) are
278 presented in the same light as Figure 4b, but are instead illustrated for the SA response transcriptional
279 program and constituent genes under BTH treatment. Again predictions match actuality, and the
280 observed trends make biological sense[32].
281   In order to illustrate Tradict's use in hypothesis-free investigation, we performed a differential
282 transcriptional program expression analysis for transcriptional programs affected by MeJA or BTH
283 treatment (Figure 4d, see Methods). Differentially expressed transcriptional programs based on
284 Tradict's predictions versus actual measurements were highly concordant and biologically reasonable.
285 Transcriptional programs differentially expressed with respect to MeJA treatment included "response to
286 bacterium," "defense response to fungus", "response to wounding," and "response to jasmonic acid" as
287 expected. Transcriptional programs differentially expressed with respect to BTH treatment included
288 various abiotic stress responses, "defense response to fungus", "response to jasmonic acid" (via
289 antagonism), and "response to salicylic acid," again, as expected.
290
291 **Discussion**
292   Tradict is an accurate, robust-to-noise method for predicting the expression of a comprehensive,
293 but interpretable list of transcriptional programs that represent the major biological processes and
294 pathways of the cell. Given the comprehensiveness, stability, and exponentially growing size of the
295 training datasets we have assembled from publicly available sources, and as evidenced by our cross
296 validation experiments, which were performed over thousands of samples deposited over the span of
297 six years, the 100 markers Tradict learns are likely to be predictive independent of most contexts and
298 applications. As illustrated through our case studies, examining the expression of these predicted
299 transcriptional programs makes intuitive sense and provides a neat summary of underlying gene
300 expression patterns.
301   Tradict additionally provides expression predictions for all genes in the transcriptome. Though
302 its current gene expression prediction accuracy is less than ideal for more sensitive applications,
303 Tradict's performance is superior to previous efforts and is improving logarithmically in the number of
304 samples. We attribute Tradict's performance gains over previous methods first to improved
305 measurement technology. Previous methods were developed for microarray, a substantially more noisy
306 technology than RNA-Sequencing[10–14]. Consequently, training efficiency was lower as well as
307 measurement accuracy of true expression, thus leading to modest prediction accuracy. By contrast,
308 Tradict is meant to interface with sequencing based readouts of gene expression, a data type that is
309 popular and proliferating exponentially as the time and price of sequencing continues to fall. Second,
310 we believe Tradict's probabilistic framework goes a step beyond previous efforts by modeling marker-
311 gene and marker-program relationships not at the level of measured abundances, which are noisy, but
312 at the level of latent abundances. Working in this denoised space naturally improves accuracy and
313 affords robustness.
314   Taken together, we believe that Tradict coupled with target RNA sequencing can enable
315 transcriptome-wide screening cheaply and at scale. Well-established commercial[19,20] and non-
316 commercial[21,33] methods exist for targeted RNA sequencing in a multiplexed manner, and they are able
317 to measure the expression of 10's-100's of genes with accuracy, making their use immediately
318 compatible with Tradict. One method in particular, RASL-Seq[22–24], does so cheaply with high precision
319 and multiplexibility by directly probing total RNA and making efficient use of dual-indexing. We estimate
320 that Tradict coupled with a time and resource efficient targeted RNA-sequencing protocol such as

RASL-Seq could bring the cost of obtaining actionable transcriptome-wide information simultaneously for thousands to tens of thousands of samples to close to $1 per sample.

This scale could greatly benefit high-throughput breeding and screening applications. Forward genetic screens in most eukyarotic organisms require assaying 10^3-10^4 mutants. Small molecule, or more generally chemogenomic, drug screens often require screening thousands of molecules against multiple cell lines in multiple conditions. Agricultural screens -- whether for breeding or field phenotyping -- also require measuring thousands of individuals. Though in these cases a screen is made cheap and scalable by monitoring an easily selectable phenotype, new phenotyping architectures must be developed and optimized for each new screen (e.g. reporter lines, imaging hardware/software). Given the ubiquity of RNA, a transcriptome-wide screening approach would not suffer from such a drawback. Furthermore, and more importantly, though quickly interpretable, the phenotype being screened for is usually a uni-dimensional datum that offers little immediate insight into mechanism. In contrast, using Tradict to help perform transcriptome-wide screening could couple the process of hypothesis generation and mechanistic investigation. Here, we argue that the scalable monitoring of the expression of a comprehensive list of just a few hundred transcriptional programs affords an attractive balance of nuance and interpretability. Consequently, this efficient investigation, largely facilitated by Tradict, could greatly accelerate the pace of genetic dissection, breeding, and drug discovery.

Code availability
A MATLAB implementation of Tradict is available at https://github.com/surgebiswas/tradict. All code to perform data downloads, analysis, and generate figures are available at https://github.com/surgebiswas/transcriptome_compression. Note that in order to make Tradict open-source and reach a wider audience we are developing a user-friendly, unit-tested Python version of Tradict that will be made available before publication.

References
1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* **3,** 318–356 (1961).
2. Kaufmann, K., Pajoro, A. & Angenent, G. C. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nat. Rev. Genet.* **11,** 830–842 (2010).
3. Mitchell, P. J. & Tjian, R. Transcriptional Regulation in Mammalian Cells by DNA Binding Proteins. *Science (80-.).* **245,** 371–378 (1989).
4. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34,** (2003).
5. Hart, Y. *et al.* Inferring biological tasks using Pareto analysis of high-dimensional data. *Nat. Methods* **12,** (2015).
6. Shoval, O. *et al.* Evolutionary Trade-Offs, Pareto Optimality, and the Geometry of Phenotype Space. *Science (80-.).* **336,** 1157–1160 (2012).
7. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, a L. The large-scale organization of metabolic networks. *Nature* **407,** 651–654 (2000).

8.   Albert, R., Lee, J. H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406,** 378–382 (2000).

9.   Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science (80-.).* **286,** 509–513 (1999).

10. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. **17,** 520–525 (2001).

11. Liew, A. W., Law, N. & Yan, H. Missing value imputation for gene expression data : computational techniques to recover missing data from available information. **12,** 498–513 (2010).

12. Celton, M., Malpertuy, A., Lelandais, G. & Brevern, A. G. De. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. (2010).

13. Ling, M. H. T. & Poh, C. L. A predictor for predicting Escherichia coli transcriptome and the effects of gene perturbations. *BMC Bioinformatics* **15,** 140 (2014).

14. Donner, Y., Feng, T., Benoist, C. & Koller, D. Imputing gene expression from selectively reduced probe sets. *Nat. Methods* **9,** (2012).

15. Heimberg, G., Bhatnagar, R., El-samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst.* **2,** 239–250 (2016).

16. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** (2014).

17. Kliebenstein, D. J. Exploring the shallow end; estimating information content in transcriptomics studies. *Front. Plant Sci.* **3,** 1–10 (2012).

18. Jaitin, D. A. *et al.* Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science (80-.).* **343,** 776–779 (2014).

19. ThermoFisher Scientific. Targeted RNA Sequencing by Ion Torrent Next-Generation Sequencing. at <http://www.thermofisher.com/us/en/home/life-science/sequencing/rna-sequencing/targeted-rna-sequencing-ion-torrent-next-generation-sequencing.html>

20. Illumina. TruSeq Targeted RNA Expression Kits. at <http://www.illumina.com/products/truseq-targeted-rna-expression-kits.html>

21. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9,** 989–1009 (2014).

22. Li, H., Qiu, J. & Fu, X. RASL-seq for Massively Parallel and Quantitative Analysis of Gene Expression. *Curr. Protoc. Mol. Biol.* 1–9 (2012). doi:10.1002/0471142727.mb0413s98

23. Larman, H. B. *et al.* Sensitive, multiplex and direct quantification of RNA sequences using a modified RASL assay. *Nucleic Acids Res.* 1–12 (2014). doi:10.1093/nar/gku636

24. Scott, E. R. *et al.* RASLseqTools: open-source methods for designing and analyzing RNA-mediated oligonucleotide Annealing, Selection, and, Ligation sequencing (RASL-seq) experiments. *bioRxiv* (2016).

25. Biswas, S. The latent logarithm. *arXiv* 1–11 (2016).

26. Ma, S. & Kosorok, M. R. Identification of differential gene pathways with principal component analysis. *Bioinformatics* **25,** 882–889 (2009).

27. Fan, J. *et al.* Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* **13,** 241–244 (2016).

28. Tropp, J. a & Gilbert, A. C. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Trans. Inf. Theory* **53,** 4655–4666 (2007).

29. Tropp, J. a., Gilbert, A. C. & Strauss, M. J. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing* **86,** 572–588 (2006).

30. Gelman, A. *et al. Bayesian Data Analysis.* (Chapman & Hall, 2013).

31. Yang, L. *et al.* The Pseudomonas syringae type III effector HopBB1 fine tunes pathogen virulence by gluing together host transcriptional regulators for degradation. *Submitted* (2016).

32. Jones, J. D. G. & Dangl, J. L. The plant immune system. *Nature* **444,** 323–9 (2006).
33. Fu, G. K. *et al.* Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. **111,** 1891–1896 (2014).