TEMPO: Detecting Pathway-Specific Temporal Dysregulation of Gene Expression in Disease

Christopher Michael Pietras
Tufts University
Medford, Massachussets
christopher.pietras@tufts.edu

Faith Ocitti
Tufts University
Medford, Massachussets
faith.ocitti@tufts.edu

Donna K. Slonim
Tufts University
Medford, Massachussets
slonim@cs.tufts.edu

ABSTRACT

While many transcriptional profiling experiments measure dynamic processes that change over time, few include enough time points to adequately capture temporal changes in expression. This is especially true for data from human subjects, for which relevant samples may be hard to obtain, and for developmental processes where dynamics are critically important. Although most expression data sets sample at a single time point, it is possible to use accompanying temporal information to create a virtual time series by combining data from different individuals.

We introduce TEMPO, a pathway-based outlier detection approach for finding pathways showing significant temporal changes in expression patterns from such combined data. We present findings from applications to existing microarray and RNA-seq data sets. TEMPO identifies temporal dysregulation of biologically relevant pathways in patients with autism spectrum disorders, Huntington's disease, Alzheimer's disease, and COPD. Its findings are distinct from those of standard temporal or gene set analysis methodologies.

Overall, our experiments demonstrate that there is enough signal to overcome the noise inherent in such virtual time series, and that a temporal pathway approach can identify new functional, temporal, or developmental processes associated with specific phenotypes.

Availability: An R package implementing this method and full results tables are available at bcb.cs.tufts.edu/tempo/.

KEYWORDS

temporal modeling, developmental dysregulation, autism, Huntington's disease, time series gene expression analysis, RNA-seq analysis, COPD, Alzheimer's Disease

ACM Reference Format:
Christopher Michael Pietras, Faith Ocitti, and Donna K. Slonim. 2018. TEMPO: Detecting Pathway-Specific Temporal Dysregulation of Gene Expression in Disease . In *ACM-BCB'18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, August 29-September 1, 2018, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3233547.3233559

1 INTRODUCTION

Understanding the dynamic aspects of molecular processes is essential, especially for inherently temporal functions such as those involved in development, disease progression, or aging [46, 68]. Transcriptional profiling, whether by microarrays, RNA-seq, or other technologies, has proven useful for identifying temporal regulatory programs.

However, the collection of data from large numbers of time points has proven to be prohibitively expensive and fraught, particularly in cases involving human subjects [73]. Thus the number of available data sets that include sufficient temporal resolution to solve key problems of interest remains limited. In most available human data sets, samples are taken only during medically indicated procedures, often yielding a single time point per individual.

If temporal information is available, however, it is possible to combine multiple samples from individuals at different ages or times into a single virtual time-series. Here we describe a method using temporal models of expression and functional gene sets to identify how and why those models break down in disease states. We do this using existing data sets featuring a single time point per individual, and we demonstrate that by so doing we can learn new things about the temporal and developmental processes associated with specific phenotypes.

1.1 Previous Work

The analysis of time series is a well-established field of data science whose relevance to expression data analysis has long been known. Computational methods specifically developed for the analysis of time series expression data are the subject of many reviews [5, 7, 55]). For example, several approaches to clustering temporal gene expression profiles have been proposed (e.g., [2, 7, 24, 47]).

Other methods have been designed to detect significantly different temporal expression profiles across experimental groups, conditions, or phenotypes. Most methods that do so (e.g., [6, 17, 56]) use similar paradigms: each gene in each condition has an expression profile that is modeled as a function of time. A score is generated for each gene, capturing the difference between the models for the different conditions; genes are then ranked by their scores.

Most effective approaches, such as those cited here, were designed specifically for time series expression data sets, which typically include only small numbers of samples for each condition and few time points. Notably, none of these methods explicitly scores gene *sets* or pathways, though it would be possible to adapt any of them to do so by using the gene scores as ranks and assessing gene set enrichment among the ranked gene list.

However, of the methods we surveyed, only maSigPro [17] provides publicly released code and has properties suitable for use with

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233559

virtual time-series. Specifically, because virtual time series combine the availability of data from whatever time points appear in the static source data, they rarely feature matched case and control samples taken at consistent time points. This property rules out straightforward utilization of time series analysis methods that require the same set of time points across both conditions, that don't allow for missing data, or that don't allow for multiple samples at the same time point. How to adjust such methods or their input data to allow their use with virtual time series is not readily apparent.

1.2 Our Contributions

Here, we introduce an approach we call TEMPO (TEmporal Modeling of Pathway Outliers) to identify pathways or gene sets that show phenotype-associated temporal dysregulation. Given a gene expression data set where each sample is characterized by an age or time point as well as a phenotype (e.g. control or disease), and a collection of gene sets or pathways, TEMPO includes the following steps. First, for each set of genes in the gene set collection, it builds a partial least squares model to predict the age of the control samples as a function of the expression of the genes in that gene set. Prediction accuracy in controls is assessed by cross validation. It then uses the same model, trained on all the control samples, to predict age in the samples with the phenotype of interest. The gene sets are ranked by a scoring function that prioritizes models that predict age well in the controls but poorly in the disease samples, suggesting temporal dysregulation. We assess the significance of the observed scores via permutation.

Note that finding models that perform well in control samples but break down in other conditions is the underlying theme of several existing outlier detection methods, including our own [44, 45]. Such strategies have therefore been widely used in a variety of contexts. However, this is the first application of this methodology to temporal models of transcriptional profiles.

We compare the ranked lists of gene sets output by TEMPO to those from two other analyses of the same data sets: Gene Set Enrichment Analysis (GSEA) [57], a standard gene-set enrichment approach to differential expression analysis that makes no explicit use of temporal information, and maSigPro [17], the only comparator method whose use on virtual time series data is straightforwardly feasible for the reasons indicated above. Still, because maSigPro itself does not look at functional enrichment, we need to translate its results at the gene level to the level of gene sets. To do so, we rank the genes by their maSigPro scores and then use GSEA to identify functional enrichment in the ranked list.

We demonstrate TEMPO's utility on four previously published expression data sets, three of which examine peripheral blood in patients with neurological conditions. The first of these is a developmental microarray data set comparing gene expression in children with or without autism spectrum disorders. The next two data sets examine neurodegenerative disorders whose progression correlates with age: a microarray data set measuring expression in the blood of people with or without Alzheimer's disease; and an RNA-seq data set that measures gene expression in adults of different ages with Huntington's disease, either before or after the onset of symptoms, or in controls. The fourth data set looks at expression in airway epithelial cells of smokers with and without COPD.

We initially chose Gene Ontology (GO) Biological Process terms [3] as our gene set collection for the experiments described here. However, for the autism data set, we augmented the GO annotations with annotations from the DFLAT project, which incorporates additional developmentally relevant annotations into the GO framework [65].

Comparing the output of different analytical methods can be complex, because related functional terms often involve similar groups of genes, so the gene sets are not independent of each other. For example, if one method implicates "neuron apoptotic process" and another "regulation of neuron death," two terms that share a common parent in the GO hierarchy ("neuron death," GO:0070997), we would like to capture this relationship. We therefore use a measure based on semantic similiarity [48] to assess relationships between the top gene-set lists output by different analytical methods.

Our examples demonstrate that TEMPO can identify age- and phenotype-related changes in expression that differ from those found by either the static analysis of GSEA or the traditional temporal modeling analysis in maSigPro. Further, our work illustrates the power of combining existing static data into virtual time series to study pathway-related temporal changes in dynamic processes.

2 METHODS

2.1 TEMPO

2.1.1 Computational model to predict age.

For a gene set G, TEMPO trains a partial least squares regression (PLSR) model [66], using the pls package in R, to predict age as a function of gene expression of all genes in G. Ages for all the control samples $C = \{S_1, S_2, ...S_j\}$ are predicted in leave-one-out cross-validation using j separate PLSR models $M_1, M_2, ...M_j$ (Figure 1). PLSR models with up to 10 components were built for each gene set; we then chose the most accurate of these models in leave one out cross-validation on the control samples, and used that model for predicting ages in the test samples. (Note that this step is not illustrated in Figure 1 to improve readability.) The best *single* size is chosen and used to train one final model M_{j+1} on the control samples $C = \{S_1, \ldots, S_j\}$. Then ages for disease samples $D = \{S_{j+1}, \ldots, S_k\}$ are predicted using M_{j+1}. We also considered using other regression models in place of PLSR (see Appendix A), but we found PLSR to be most effective.

2.1.2 Scoring gene sets by performance on cases and controls.

For each gene set G, we have a set of age predictions for all control samples C and all disease samples D. We obtain a vector of prediction errors, the differences between the predicted ages for G and the actual ages. We call this vector of prediction errors E_G, where $E_{G,s}$ is the prediction error for sample s under gene set G. Using these errors, we determine the degree to which G is temporally dysregulated by calculating a score that quantifies the accuracy of the predictions for the control samples and the *in*accuracy of the predictions for the disease samples.

If our data sets behave as expected, these errors can be assumed to be normally distributed (although we assess and relax this assumption in Appendix B). Let μ_G and σ_G be the mean and standard deviation of the observed prediction errors on the control samples for gene set G, and let $\mathcal{N}_G(x)$ be the probability of seeing an error

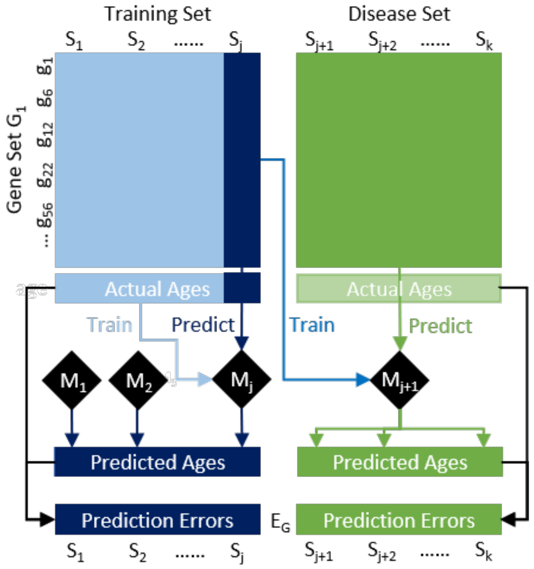

Figure 1: PLSR prediction for an arbitrary gene set G_1. For j training samples, ages are predicted using j PLSR models in cross-validation. For the $k - j$ disease samples, ages are predicted using a single PLSR model trained on all training samples. The difference between the predicted and actual ages for sample S_i is the prediction error E_{G,S_i}.

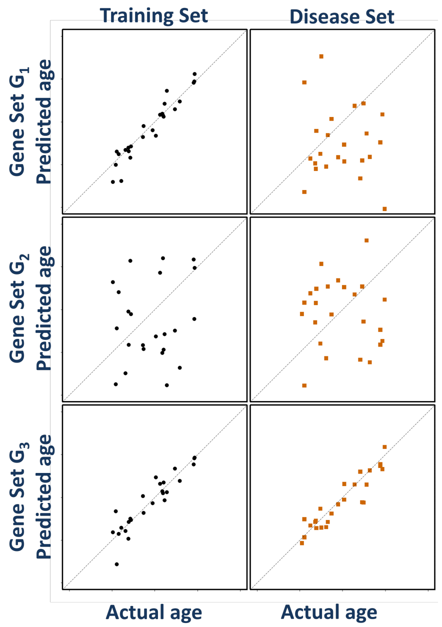

Figure 2: Predicted age v. actual age for hypothetical gene sets G_1, G_2, and G_3 for control (left) and disease (right) samples. Gene sets like G_1 have higher scores (Eq. 1).

at least as large as x under the normal distribution with mean μ_G and standard deviation σ_G.

We then calculate the following score for gene set G, control sample set C, and disease sample set D:

$$Score(G) = \frac{|C| \sum_{s \in D} -log(\mathcal{N}_G(E_{G,s}))}{|D| \sum_{s \in C} -log(\mathcal{N}_G(E_{G,s}))} \times L_1 \times L_2 \qquad (1)$$

This is essentially a normalized ratio of the average "surprisal" score [52] of the disease samples to that of the control samples. It is highest when the disease sample predictions are surprisingly bad, using an accurate model trained on the controls.

This score also captures our criteria for interesting gene sets. In gene sets where a reliable temporal pattern of expression in the controls breaks down in disease, we would be able to build a regression model that accurately predicts age in the control samples, but is unable to predict age accurately in disease, yielding many samples with improbable prediction errors and a high score. In gene sets where this is not the case, the regression model will have the same predictive power regardless of class label, yielding low scores (Figure 2).

The $\frac{|C|}{|D|}$ factor normalizes the score for the size of the control and disease sample sets, allowing meaningful comparison of results across experiments. Yet the ratio of the average surprisal scores alone can be sensitive to differences in the age distribution between cases and controls, a common confounding factor in many data sets. Such differences can result in situations where even extremely poor models of age as a function of gene expression would be reported as significant, as in the hypothetical example in Figure 3.

To avoid detecting such spurious hits, we multiply the surprisal ratio by two coefficients L_1 and L_2, both obtained from a new linear model M'_G of the predicted ages for G as a function of the true ages on the control data. The L_1 term models how linear the resulting fit is; we set $L_1 = 1$ minus the p-value associated with the linear model M'_G. L_2 models how close the resulting fit is to the $x = y$ line; we use $L_2 = cos2\theta$, where θ is the angle of deflection of the best-fit line for M'_G from the $x = y$ line. Together, these coefficients down-weight any potentially high scoring models whose high scores appear to be spurious.

2.1.3 Significance of Observed Scores. We estimate statistical significance via a permutation testing procedure. Specifically, we generate a set of 500 random permutations of phenotype labels. For each permutation P in this set and each gene set G, we build a new temporal model on the new set of nominal "control" samples and recompute the score of the gene set for that permutation (we call this $Score(G, P)$). The reported p-value for G is simply the percentage of all permutations where $Score(G, P) \geq Score(G)$. To account for multiple hypothesis testing, we calculate false discovery rates using the Benjamani-Hochberg procedure [9].

We report results for gene set G only if Score(G) ≥ 1, raw $p \leq 0.05$, and FDR ≤ 0.25. Since this method is primarily intended for hypothesis generation, we might still be interested in gene sets with a false discovery rate this large; this is the default cutoff for the GSEA software as well [57]. Scores below 1 mean that the average surprisal of the controls is larger than the average surprisal of patients with the phenotype. All three of these values (score, raw p, and FDR) are reported in our full results tables online.

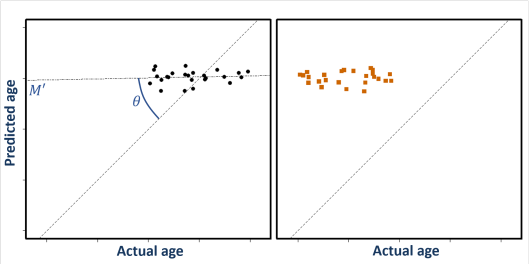

Figure 3: Predicted age v. actual age for hypothetical gene set, for control samples on the left and disease samples on the right. PLSR has no true predictive power in this gene set; it predicts almost the same age regardless of input. However, due to the different age distributions in the control and disease sets, the average surprisal ratio term of Equation 1 (without the L_1 and L_2 terms) is relatively high, because the control predictions are close to the ideal $x = y$ line, while the disease predictions are farther from it. The $L_2 = cos2\theta$ term down-weights such results. The L_1 term is correlated with the closeness of the fit around M'.

2.2 Expression data sets

Autism spectrum disorders: The autism data set, referred to as ASD, is based on a study by Mark Alter, *et al.* [1], that includes expression microarray data from peripheral blood lymphocytes for 59 control patients and 72 patients with autism spectrum disorders, with ages ranging from two to fourteen years. The data are available as GSE25507 in the Gene Expression Omnibus (GEO) database [22]; from this data set, we used all the samples for which subject ages were available.

Alzheimer's Disease: The Alzheimer's disease data set, referred to as AD, is based on a subset of the data used in a study by Sood, *et al.* [54] from AddNeuroMed [37]. We include all samples from Batch 1 (available as GSE63060 on GEO) marked as "included in the case-control study," for a data set consisting of blood gene expression data for 49 samples from Alzheimer's patients and 67 from roughly similar-aged controls. All of these samples were annotated with patient ages in integer years.

Huntington's disease: The Huntington's disease data set, referred to as HD, includes normalized gene counts from an RNASeq experiment characterizing blood from Huntington's disease patients [40]. Its GEO accession number is GSE51779. The data set includes 33 control samples and 91 Huntington's disease carriers, 27 of whom are asymptomatic (defined as patients for whom the motor score component of the Unified Huntington's Disease Rating Scale [62] is 5 or less). All of these samples were annotated with patient ages in years to .01 precision, ranging from about 20 to 80 years.

COPD: The COPD data set is based on data from studies by Carolan, *et al.* [14] and Tilley, *et al.* [58], available as GSE5058 on GEO. This data set contains small airways gene expression data from 15 smokers with COPD and 12 smokers who are apparently healthy. Each patient has an integer age in years.

2.3 Gene set collections

For the HD, AD, and COPD data sets, we used Gene Ontology [3] (GO) Biological Process gene sets. However, for the ASD data set, we used a version of the GO collection augmented with additional developmentally relevant annotations from the DFLAT project [65]. Specifically, the Feburary 19, 2016 gene set gmt files were downloaded from the DFLAT web site (dflat.cs.tufts.edu). The Gene Ontology collection, generated at the same time as the DFLAT gene sets, was obtained from the same web site. Both the DFLAT and GO collections were filtered to remove all gene sets of size greater than 500 or less than 5, resulting in a total of 8416 DFLAT gene sets and 6484 GO gene sets.

2.4 Comparator methods

2.4.1 GSEA. To account for differences in expression that are not related to age or time, we compare to Gene Set Enrichment Analysis [57]. GSEA ranks gene sets by how represented genes from a given gene set are at the top (or bottom) of the list of all genes ranked by differential expression between two conditions. In this mode, GSEA does not account for any differences in expression dependant on time. GSEA also accepts pre-generated rankings of genes in GSEA-preranked, which we use with maSigPro below.

2.4.2 maSigPro. We first translated each of our static expression datasets into a suitable time series data set, with the number of replicates equal to the number of patients and each with a single time point.

We used the R package released with maSigPro [17] to generate scores for each of the genes measured in each of our data sets. We then needed to extend these results to identify implicated *gene sets* rather than individual genes. We therefore used the "preranked" option in GSEA, with the rankings corresponding to the maSigPro scores, to identify differentially-expressed gene sets. It is worth noting that with preranked data, GSEA assesses significance by permuting gene sets, since it cannot permute class labels.

2.5 Comparing gene set lists

Semantic similarity: To compare the similarity of the top-scoring gene sets from different analyses, exact-match methods are insufficient, because different analyses may find different but related terms; one may discover "apoptotic process" while another may discover "neuron apoptotic process." To capture these semantic relationships, we use pairwise Resnik semantic similarity scores [48]. All scores were calculated using the GoSemSim [69] R package. Though GoSemSim offers tools for calculating semantic similarity between sets of GO terms, we found these numbers difficult to assess in absolute terms.

To address this, we instead examine which terms have significantly similar matches in the other term set. That is, given two collections of terms T_1 and T_2, for each term t_i in T_1, we want to know if there exists a semantically similar term t_j in T_2. Given the distribution of pairwise Resnik similarity scores involving t_i, we say a term t_j is semantically similar to t_i if Resnik(t_i,t_j) is above some chosen cutoff c. For our experiments here, we chose $c = 0.6$, which corresponds to approximately the top 0.3% of pairwise Resnik scores between all biological process gene sets, and compare between collections of gene sets of size 40. The number 40 was chosen arbitrarily to represent the variety of top functions in the output.

This choice of parameters means that there exists some likelihood that some number of gene sets in each collection are semantically similar by chance alone. We quantify this likelihood by permutation. For each permutation i, we generate two random collections of 40 terms and determine the number of terms n_i from the first collection that have semantically similar terms in the second. We do this 500 times, and compare the the number of similar terms n from T_1 to T_2 to this distribution to obtain the likelihood of seeing as much similarity by chance; this is simply the fraction of permutations where $n_i \geq n$. For $|T_1| = |T_2| = 40$, we found that an overlap of at least 8 semantically similar gene sets is required for the likelihood of seeing such overlap by chance to be below 0.05.

Correlation: We also consider the Spearman's rank-correlations between full gene set lists from two different analyses. While such an approach penalizes changes in the rankings of even insignificant gene sets, it has the advantage that it involves all gene sets equally. We do not consider correlations between either GSEA or maSigPro and TEMPO in this way; while high TEMPO and high GSEA scores denote something comparable, low TEMPO scores denote a complete lack of pattern of temporal expression and low GSEA scores can indicate enrichment in the control condition. Using the absolute value of the GSEA score might be more appropriate, but again it is not clear that such values would be comparable with rankings by other methods.

3 RESULTS AND DISCUSSION

3.1 TEMPO finds unique temporal dysregulation in disease classes

In all four data sets, TEMPO identifies pathways that are known to change with development or age, but whose normal temporal trajectory is disrupted in disease. The observed temporal dysregulation is in many cases consistent with prior knowledge, and sometimes consistent with identified or proposed therapeutic targets for the indicated disease. Thus, novel findings from this approach may be suggestive of new targets or interventions.

The TEMPO results differ in many respects from the gene sets returned by comparator methods GSEA and maSigPro. *No exactly identical* gene sets appeared in the top 40 listed in any TEMPO analysis and any comparator method. Table 1 shows the number and significance of *semantically similar* gene sets observed between TEMPO and the comparator methods. Only between TEMPO and GSEA on the COPD data set do we see a *nearly*-significant number of semantically similar gene sets between the two lists. Furthermore, in several cases, either GSEA or maSigPro does not identify *any* significant gene sets. In these cases, we nonetheless compared the semantic similarity of the top 40 highest-scoring gene sets from each method to the TEMPO results.

The differences between the TEMPO and GSEA results are not unexpected. Gene sets with high GSEA scores will not necessarily have high TEMPO scores, because gene sets where there is no pattern of expression as a function of time will not be scored highly by TEMPO regardless of any time-independent differential expression that may exist.

Table 1: Number of semantically similar gene sets (with significance) from the top 40 TEMPO results for the data set in each row and the top 40 results of the indicated comparator method run on the same data set.

TEMPO on	GSEA	maSigPro+GSEA
Autism	3 (0.666)	0 (1.000)
Huntington's	2 (0.838)	4 (0.468)
Alzheimer's	1 (0.958)	5 (0.286)
COPD	7 (0.074)	3 (0.666)

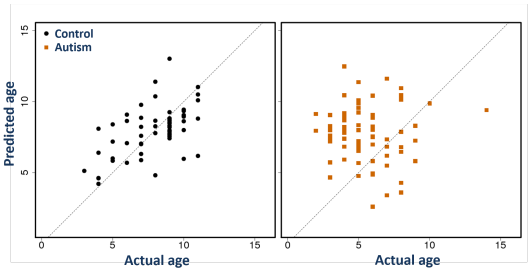

Figure 4: Predicted age vs. actual age for both control (black circles, left) and autistic (orange squares, right) subjects for genes with the annotation "Regulation of serotonin uptake" on the ASD data. Each dot represents one patient. Predicted ages are those produced by TEMPO using the model built from the controls, based only on the expression values of genes in the regulation of serotonin uptake pathway.

For space reasons, full results tables and scatter plots for all methods and data sets are available online at bcb.cs.tufts.edu/tempo/. However, we reproduce a part of the TEMPO results table for the ASD data set in this manuscript as an example.

3.2 Developmental dysregulation in Autism: neurotransmitters and inflammation

In the ASD expression data, TEMPO identified 591 significant gene sets. A selection of the highest-scoring of these is shown in Table 2. Common themes in this list include inflammation, angiogenesis, *PTEN* activity, apoptosis, developmental processes, and neurotransmitter signaling.

Results from both a static GSEA analysis and the maSigPro-plus-enrichment analysis on the same data set are also available on the TEMPO web site. Neither GSEA nor maSigPro analysis returns *any* gene sets with FDR $\leq .25$, though both have several hundred gene sets with raw $p \leq .05$.

The role of serotonin and other neurotransmitters in the etiology of ASD has long been investigated [18, 49]. Although serotonin activity is evident very early in human development [43], the nature and expression of serotonin response pathways change considerably during both childhood and adolescence [19], consistent with observations that children and adults respond differently to drugs targeting this system [63]. Further, while SSRIs are often used to treat ASD patients, there is considerable evidence of increased adverse events in the pediatric autistic population, suggesting increased care is needed in the use of these drugs [32]. Understanding specifically how the expression of serotonin-related genes is expected to change with age in the neurotypical population, and

Table 2: **A selection of high-scoring gene sets ins ASD ranked by TEMPO score.**

Rank	Gene Set	Score	p-value	FDR
1	regulation of cell migration involved in sprouting angiogenesis	3.82	0.01	0.11
2	secretion by cell	2.27	0.00	0.07
5	myeloid leukocyte activation	2.01	0.00	0.07
6	pos. reg. of sequence-specific DNA binding transcription factor activity	1.93	0.00	0.07
7	positive regulation of NF-kappaB transcription factor activity	1.89	0.00	0.07
9	inflammatory response	1.87	0.00	0.07
10	central nervous system neuron differentiation	1.86	0.00	0.07
14	phosphatidylinositol biosynthetic process	1.82	0.00	0.07
15	phosphatidylinositol metabolic process	1.80	0.00	0.07
17	myeloid dendritic cell activation	1.77	0.00	0.07
20	facial nerve development	1.72	0.01	0.12
22	nuclear division	1.72	0.00	0.07
24	regulation of sequence-specific DNA binding transcription factor activity	1.72	0.00	0.07
32	pos. reg. of cysteine-type endopeptidase activity involved in apoptotic process	1.65	0.00	0.08
33	negative regulation of neurotransmitter uptake	1.65	0.00	0.07
34	regulation of serotonin uptake	1.65	0.00	0.07

how autistic patients differ from these expectations, may be key to the better prediction of tolerance and appropriate dosage in this population.

Figure 4 plots the actual and TEMPO-predicted ages for the gene set "regulation of serotonin uptake." The plot on the left shows the relatively accurate predictive age models in the controls, while that on the right show how the developmental program of the genes in the pathway breaks down in the group of subjects with ASD.

Inflammatory pathways have also been linked to ASD [20], and an increase in the circulating frequency of myeloid dendritic cells, which modulate immune response, has been observed in children with ASD compared to controls [12]. NFKB signaling has been implicated as well [72], possibly contributing to the dysregulation of inflammatory cytokines [35].

Programmed cell death is known to play a key role in normal brain development [67]. Disruption of apoptotic pathways has been shown to contribute to the development of ASD and to symptoms suggestive of it in animal models [38, 64]. It has been suggested that abnormal *PTEN* function, which has been documented in a subset of the autism patients [34], may contribute to apoptosis in neural development by regulating PI3K / AKT signaling [64, 71]. Both *PTEN* (also known as "phosphatase and tensin homolog") and *PI3K* ("phosphatidylinositol-3-kinase") are involved in phosphatidylinositol metabolism; this pathway has even been suggested as a possible therapeutic target for autism [23]. *PTEN* has also been shown to regulate angiogenesis [16], which has itself been implicated in autism spectrum disorders [4] and which is the most significant process identified by TEMPO in Table 2.

3.3 Temporal dysregulation of apoptosis in Alzheimer's disease

In the Alzheimer's data set, TEMPO identified 2772 significant gene sets. The pathways implicated include several processes known to have relevance in Alzheimer's disease, including apoptosis, neurological development, immunity, the DNA damage response, and regulation of phosphorylation.

Amyloid beta plaques have been observed to induce apoptosis in Alzheimer's disease (AD) patients [28]. Intrinsic apoptosis through altered mitochondrial permeability, triggered by accumulations of amyloid beta precursor protein, has been proposed as the mechanism by which amyloid plaques induce mitochondrial oxidative stress in AD [8]. Five of the top 40 gene sets identified by TEMPO in the Alzheimer's population are related to apoptosis, including "regulation of apoptotic signaling pathway," "intrinsic apoptotic signaling pathway," and "regulation of intrinsic apoptotic signaling pathway," with scores ranking 7th, 20th, and 39th, and all with FDR ≤ 0.006.

Neurodevelopmental pathways have been shown to play a role in signaling in AD [29]. Dysregulation in the expected age-related changes in the "central nervous system development" and "brain development" gene sets was identified by TEMPO. Possible consequences of functional changes in developmental signaling pathways in AD patients include protein hyperphosphorylation [11]. Hyperphosphorylation of tyrosines in the tau protein is thought to contribute to aggregation of atypically-folded tau into neurofibrillary tangles in the brain [39]. Thus, dysregulation of neurodevelopmental signaling may also explain the top TEMPO gene set, "peptidyl-tyrosine phosphorylation," having an atypical profile in the AD patients.

Finally, the substantial role of the immune system and cytokine signaling in Alzheimer's is well explored [50], and has been proposed as the basis of new immunotherapeutic approaches [41]. Previous work has shown changes in immune processes and signaling in healthy aging. T-cell populations change and pro-inflammatory cytokine signaling increases with age [27, 61]. TEMPO's identification of cytokine signaling and T-cell activation pathways again confirms that it is finding likely pathways that have a predictable age-related pattern that breaks down in disease, and that may suggest therapeutic targets.

3.4 Age-related expression dysregulation in pre-symptomatic HD patients

Huntington's disease (HD) is known to be caused by a trinucleotide repeat expansion of the *huntingtin* (*HTT*) gene. However, many other genes have been found to modify the effects of these expansions, reflecting age of onset, severity, and specific characteristics of the disorder [42]. Such modifiers are actively sought as potential therapeutic targets for patients.

For this data set, 414 gene sets met the significance criteria, suggesting that there are age-specific expression patterns for many biological processes that are disrupted in the disorder. In contrast, neither Gene Set Enrichment Analysis nor maSigPro returns any significant gene sets for this data set. The full results for all these analyses are available at the TEMPO web site.

Two of the ten highest scoring gene sets in the TEMPO analysis are "regulation of ERBB signaling pathway" and "regulation of EGFR signaling pathway." Prior evidence has implicated the ERBB pathway and EGFR signaling in the pathogenesis of HD [31, 36]. Mechanistic studies suggest that mutant *HTT* interferes with EGFR

Table 3: Number of semantically similar or identical gene sets (with significance) from the top 40 TEMPO results for the Huntington's subset in each row to the top 40 TEMPO results for the subset in the corresponding column.

	All	Symptomatic	Asymptomatic
All	-	28 (0.00)	21 (0.00)
Symptomatic	29 (0.00)	-	21 (0.00)
Asymptomatic	18 (0.00)	18 (0.00)	-

signaling, and ERBB signaling defects have been implicated in other neurodegenerative diseases including Alzheimer's [13].

Another notable observation is the relative temporal dysregulation of telomere maintenance genes in HD, consistent with the second-ranked TEMPO hit "telomere maintenance via recombination." Telomere length in HD has recently been verified to be shorter than in controls, and more so than in other forms of dementia [33]. This process is known to reflect aging in general, but identifying further disruption of the normal aging patterns in HD represents an important finding with potential therapeutic implications.

These results are based on comparing controls to both symptomatic and pre-symptomatic patients together, but many of the observations hold when symptomatic and pre-symptomatic patients are considered separately. The pairwise Spearman correlations between the TEMPO scores for just symptomatic, just pre-symptomatic, and the combined data set are all extremely high (≥ 0.99). The top-scoring gene sets returned by TEMPO for each of these three comparisons are also very similar, with a minimum of 18 out of the top 40 gene sets being semantically similar or identical in each pairing, as can be seen in Table 3. In general, there is more significant disruption of age-specific patterns in the symptomatic patients, but such disruptions are still detectable when comparing the pre-symptomatic patients to the controls (see e.g. Figure 5). Considered on their own, gene set scores in asymptomatic patients do not meet FDR significance criteria, but the close correlation with scores for symptomatic patients suggests these scores are in fact meaningful.

Our results suggest a pattern of expression disruption for many of these gene sets that is detectable before disease onset. This is perhaps not surprising; prior imaging work has identified differential aging in a transgenic rat model of HD [10], even before the onset of symptoms, and some expression changes have even been documented in pre-symptomatic human HD patients [15]. Still, the presence of a coherent change in age-related regulation prior to symptom onset may yield novel therapeutic insights.

3.5 Age-dependent dysregulation of immune pathways in COPD

In the COPD data set, TEMPO identified 3085 significant gene sets whose predictable age-related expression relationships in airway epithelial cells from healthy smokers are disrupted in COPD.

Previous work has shown increased expression of pro-inflammatory cytokines and decreased NK cell activity in asymptomatic smokers [70]. Increased inflammatory signaling correlates with pack-years, and therefore with age, even in smokers without apparent disease [30]. Unexpectedly early aging-like changes in vascular smooth muscle cells have been correlated with the inflammatory cytokines and oxidative stress likely to result from smoking [60].

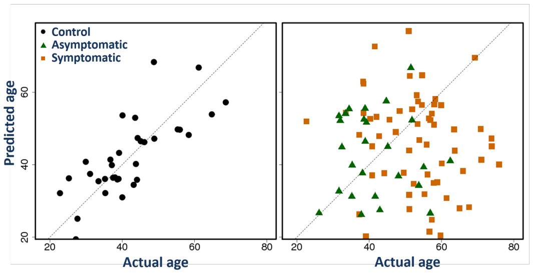

Figure 5: Predicted age vs. actual age for control (left), pre-symptomatic, and symptomatic Huntington's (right) subjects for a top-scoring pathway on the Huntington's Disease data, "Regulation of ERBB signaling pathway." Each dot represents one patient in the control group; squares represent HD patients, and triangles represent HD patients who are still pre-symptomatic. Predicted ages are those produced by TEMPO using the model built from the controls, based only on the expression values of genes in the EERB signaling pathway.

It is not therefore surprising that we find excellent predictive models of age using the GO gene sets "positive regulation of interferon-gamma secretion," "T-helper 17 cell lineage commitment," "vascular smooth muscle cell development," and "regulation of interleukin-13 secretion," that break down in COPD. These gene sets are ranked 2, 7, 9, and 10 of those identified in the COPD data (ranking of gene set G is by Score(G)); the raw p-values for all of these are 0.002, and the corresponding adjusted FDR is less than 0.02.

Age related changes specific to alanine and glutamine transport have been observed in rat blood cells [26]. Although there is little data describing amino acid transport with age in human airway epithelial cells, it is intriguing that exactly these two have been observed with disrupted age related patterns in the COPD patients. Other immune and inflammatory gene sets, including "negative regulation of type 2 immune response," "neutrophil homeostasis," and "negative regulation of activin receptor signaling pathway," are found to be significantly disrupted in the COPD data as well.

3.6 Modest similarities among neurological disorders

The number of semantically similar gene sets between the top TEMPO results in each of the three data sets from peripheral blood in neurodevelopmental or neurodegenerative disorders were not individually significant, with a maximum of 5 (0.286) similar between Autism and Huntington's. However, all pairwise comparisons have at least one semantically similar term related to regulation of transcription or transcription factor activity. Furthermore, the full TEMPO gene set lists still show modest rank-correlation, with the highest observed correlation, 0.39, between the Huntington's and Alzheimer's disease results (see Table 4). This makes sense because both are based on models of normal age-related processes in blood samples that are degraded over time in a neurodegenerative disease state.

Table 4: Spearman's rank correlations between TEMPO scores for all shared gene sets on each pair of data sets.

	HD	AD	COPD
ASD	0.1619	0.2713	0.0290
HD	-	0.3891	0.0506
AD	-	-	0.0893

We note that the neurological conditions all show lower similarity to the COPD data set, with zero observed semantically similar gene sets among the top hits, and lower correlations across the board. Although the COPD controls are in a similar-aged population to the AD and HD controls, we note that the COPD controls are all smokers, the samples are measuring expression in small airway epithelial cells rather than blood, and COPD is not a neurodegenerative disorder. All of these points likely contribute to explain the observed results.

4 CONCLUSION

Many studies have focused on identifying dynamic expression changes in temporal processes. Most of these, however, use either static or traditional time series analyses on self-contained temporal data sets with a limited number of time points [73]. Generating additional time points for such analyses involves a cost-benefit tradeoff that has recently been explored [51]. Although there is typically greater benefit from adding time points at the expense of replicates, the costs of sampling adequately to identify medically-relevant changes in temporal dynamics may be prohibitive, especially when the dynamic processes are not already well understood.

We have therefore suggested integrating temporal information across static data sets to create a virtual time series, and we introduced an approach based on outlier detection to identify functional pathways or gene sets in which the temporal pattern of expression is disrupted. It is perhaps somewhat surprising that the temporal signal in disease can be strong enough to overcome the noise inherent in combining data points from different subjects, but that observation emphasizes the power to be gained by using an explicit temporal model. Such an approach to data integration will be increasingly valuable as the collections of usable static data in public repositories continue to grow.

This approach may also be applied to any continuous variable, not just time or age, that characterizes high-dimensional data that likely reflects categorical phenotypes or sample characteristics. Potential applications are many, but it seems particularly likely that these methods could be of value in gaining a better mechanistic understanding of developmental disorders or issues in geriatric medicine.

Diseases involving progressive decline or loss of function represent another important application area. Although here we have focused on age as the relevant temporal variable, a more appropriate temporal annotation might be time since diagnosis, or time since some other clinically-defined criterion, rather than age *per se*. It is not then obvious what the appropriate temporal annotation to measure in the control patients would be, but meaningful solutions could be derived for individual use cases. Such an approach might help identify early degenerative or compensatory signals in the course of disease, with potential implications for treatment.

At present, TEMPO identifies only dysregulation in predefined sets of genes. Another important direction for future work is identification of de novo gene sets. Such a method could use the TEMPO dysregulation score as a fitness metric in an optimization algorithm, expanding or recombining known dysregulated gene sets to identify new ones.

Finally, many expression data sets include expression data taken from the same individual at a small number of time points. An interesting and important question for future work is to develop methods for integrating such short time series with static data in an intelligent way. Specifically, the method should make use of dependencies between samples from the same individuals while allowing the use of unrelated samples to learn more about the temporal or age-related expression variation. Doing so will enable better exploitation of existing repositories of transcriptional data for novel discovery.

ACKNOWLEDGEMENTS

We thank Diana Bianchi, Jill Maron, and Lystra Hayden for valuable comments on an earlier draft of this manuscript, and members of the Tufts BCB research group for helpful discussions.

FUNDING

This work was supported by NIH R01HD076140.

REFERENCES

[1] Mark D Alter, Rutwik Kharkar, Keri E Ramsey, David W Craig, Raun D Melmed, Theresa A Grebe, R Curtis Bay, Sharman Ober-Reynolds, Janet Kirwan, Josh J Jones, et al. 2011. Autism and increased paternal age related changes in global levels of gene expression regulation. *PloS one* 6, 2 (2011), e16715.

[2] IP Androulakis, E Yang, and RR Almon. 2007. Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annual review of biomedical engineering* 9 (2007), 205.

[3] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25–29.

[4] E.C. Azmitia, Z.T. Saccomano, M.F. Alzoobaee, M. Boldrini, and P.M. Whitaker-Azmitia. 2016. Persistent Angiogenesis in the Autism Brain: An Immunocytochemical Study of Postmortem Cortex, Brainstem and Cerebellum. *J Autism Dev Disord.* 46, 4 (2016), 1307–18.

[5] Z. Bar-Joseph. 2004. Analyzing time series gene expression data. *Bioinformatics* 20, 16 (Nov 2004), 2493–2503.

[6] Ziv Bar-Joseph, Georg Gerber, Itamar Simon, David K Gifford, and Tommi S Jaakkola. 2003. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proceedings of the National Academy of Sciences* 100, 18 (2003), 10146–10151.

[7] Ziv Bar-Joseph, Anthony Gitter, and Itamar Simon. 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics* 13, 8 (2012), 552–564.

[8] M. G. Bartley, K. Marquardt, D. Kirchhof, H. M. Wilkins, D. Patterson, and D. A. Linseman. 2012. Overexpression of amyloid-ÃŐÂš protein precursor induces mitochondrial oxidative stress and activates the intrinsic apoptotic cascade. *J. Alzheimers Dis.* 28, 4 (2012), 855–868.

[9] Y. Benjamini and Y. Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. *J R Stat Soc B* 57 (1995), 289âĂŞ300.

[10] I. Blockx, N. Van Camp, M. Verhoye, R. Boisgard, A. Dubois, B. Jego, E. Jonckers, K. Raber, K. Siquier, B. Kuhnast, F. DollÃľ, H.P. Nguyen, S. Von HÃűrsten, B. Tavitian, and A. Van der Linden. 2011. Genotype specific age related changes in a transgenic rat model of Huntington's disease. *Neuroimage* 58, 4 (15 Oct 2011), 1006–16.

[11] M. Bothwell and E. Giniger. 2000. Alzheimer's disease: neurodevelopment converges with neurodegeneration. *Cell* 102, 3 (Aug 2000), 271–273.

[12] E. Breece, B. Paciotti, C. W. Nordahl, S. Ozonoff, J. A. Van de Water, S. J. Rogers, D. Amaral, and P. Ashwood. 2013. Myeloid dendritic cells frequencies are increased

in children with autism spectrum disorder and associated with amygdala volume and repetitive behaviors. *Brain Behav. Immun.* 31 (Jul 2013), 69–75.

[13] E. M. Bublil and Y. Yarden. 2007. The EGF receptor family: spearheading a merger of signaling and therapeutics. *Curr. Opin. Cell Biol.* 19, 2 (Apr 2007), 124–134.

[14] Brendan J Carolan, Adriana Heguy, Ben-Gary Harvey, Philip L Leopold, Barbara Ferris, and Ronald G Crystal. 2006. Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers. *Cancer research* 66, 22 (2006), 10729–10740.

[15] K. H. Chang, Y. C. Chen, Y. R. Wu, W. F. Lee, and C. M. Chen. 2012. Downregulation of genes involved in metabolism and oxidative stress in the peripheral leukocytes of Huntington's disease patients. *PLoS ONE* 7, 9 (2012), e46492.

[16] S. Choorapoikayil, B. Weijts, R. Kers, A. de Bruin, and J. den Hertog. 2013. Loss of Pten promotes angiogenesis and enhanced VEGFA expression in zebrafish. *Dis Model Mech.* 6, 5 (2013), 1159–66.

[17] Ana Conesa, María José Nueda, Alberto Ferrer, and Manuel Talón. 2006. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* 22, 9 (2006), 1096–1102.

[18] EH Cook, Rachel Courchesne, Catherine Lord, Nancy J Cox, Shuya Yan, Alan Lincoln, Richard Haas, Eric Courchesne, and Bennett L Leventhal. 1997. Evidence of linkage between the serotonin transporter and autistic disorder. *Molecular psychiatry* 2 (1997), 247–250.

[19] F. Crews, J. He, and C. Hodge. 2007. Adolescent cortical development: a critical period of vulnerability for addiction. *Pharmacol. Biochem. Behav.* 86, 2 (Feb 2007), 189–199.

[20] Jan Croonenberghs, Eugene Bosmans, Dirk Deboutte, Gunter Kenis, and Michael Maes. 2002. Activation of the inflammatory response system in autism. *Neuropsychobiology* (2002).

[21] Harris Drucker, Chris JC Burges, Linda Kaufman, Alex Smola, Vladimir Vapnik, et al. 1997. Support vector regression machines. *Advances in neural information processing systems* 9 (1997), 155–161.

[22] Ron Edgar, Michael Domrachev, and Alex E Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* 30, 1 (2002), 207–210.

[23] L. Enriquez-Barreto and M. Morales. 2016. The PI3K signaling pathway as a pharmacological target in Autism related disorders and Schizophrenia. *Mol Cell Ther* 4 (2016), 2.

[24] Jason Ernst and Ziv Bar-Joseph. 2006. STEM: a tool for the analysis of short time series gene expression data. *BMC bioinformatics* 7, 1 (2006), 191.

[25] H.A. Farahani, A. Rahiminezhad, L. Same, and K. Immannezhad. 2010. A Comparison of Partial Least Squares (PLS) and Ordinary Least Squares (OLS) regressions in predicting of couples mental health based on their communicational patterns. *Procedia Social and Behavioral Sciences* 5 (2010), 1459âĂŞ63.

[26] A. Felipe, O. Vinas, and X. Remesar. 1992. Changes in alanine and glutamine transport during rat red blood cell maturation. *Biosci. Rep.* 12, 1 (Feb 1992), 47–56.

[27] S. K. Garg, C. Delaney, H. Shi, and R. Yung. 2014. Changes in adipose tissue macrophages and T cells during aging. *Crit. Rev. Immunol.* 34, 1 (2014), 1–14.

[28] S. Ghavami, S. Shojaei, B. Yeganeh, S. R. Ande, J. R. Jangamreddy, M. Mehrpour, J. Christoffersson, W. Chaabane, A. R. Moghadam, H. H. Kashani, M. Hashemi, A. A. Owji, and M. J. ?os. 2014. Autophagy and apoptosis dysfunction in neurodegenerative disorders. *Prog. Neurobiol.* 112 (Jan 2014), 24–49.

[29] M. Grilli, G. Ferrari Toninelli, D. Uberti, P. Spano, and M. Memo. 2003. Alzheimer's disease linking neurodegeneration with neurodevelopment. *Funct. Neurol.* 18, 3 (2003), 145–148.

[30] S. S. Hacievliyagil, L. C. Mutlu, and I. Temel. 2013. Airway inflammatory markers in chronic obstructive pulmonary disease patients and healthy smokers. *Niger J Clin Pract* 16, 1 (2013), 76–81.

[31] Ravi Kiran Reddy Kalathur, Miguel A Hernández-Prieto, and Matthias E Futschik. 2012. Huntington's Disease and its therapeutic target genes: a global functional profile based on the HD Research Crossroads database. *BMC neurology* 12, 1 (2012), 1.

[32] A. Kolevzon, K. A. Mathewson, and E. Hollander. 2006. Selective serotonin reuptake inhibitors in autism: a review of efficacy and tolerability. *J Clin Psychiatry* 67, 3 (Mar 2006), 407–414.

[33] L.N. Kota, S. Bharath, M. Purushottam, N.S. Moily, P.T. Sivakumar, M. Varghese, P.K. Pal, and S. Jain. 2015. Reduced telomere length in neurodegenerative disorders may suggest shared biology. *J Neuropsychiatry Clin Neurosci.* 27, 2 (2015), e92–6.

[34] S. Kyrylenko, M. Roschier, P. Korhonen, and A. Salminen. 1999. Regulation of PTEN expression in neuronal apoptosis. *Brain Res. Mol. Brain Res.* 73, 1-2 (Nov 1999), 198–202.

[35] T. Lawrence. 2009. The nuclear factor NF-kappaB pathway in inflammation. *Cold Spring Harb Perspect Biol* 1, 6 (Dec 2009), a001651.

[36] Devys D Liu YF, Deth RC. 1997. SH3 domain-dependent association of huntingtin with epidermal growth factor receptor signaling complexes. *J Biol Chem.* 272, 13 (1997), 8121–4.

[37] Simon Lovestone, Paul Francis, Iwona Kloszewska, Patrizia Mecocci, Andrew Simmons, Hilkka Soininen, Christian Spenger, Magda Tsolaki, Bruno Vellas, Lars-Olof Wahlund, et al. 2009. AddNeuroMedâĂŤthe European collaboration for the discovery of novel biomarkers for Alzheimer's disease. *Annals of the New York Academy of Sciences* 1180, 1 (2009), 36–46.

[38] R. L. Margolis, D. M. Chuang, and R. M. Post. 1994. Programmed cell death: implications for neuropsychiatric disorders. *Biol. Psychiatry* 35, 12 (Jun 1994), 946–956.

[39] L. Martin, X. Latypova, C. M. Wilson, A. Magnaudeix, M. L. Perrin, C. Yardin, and F. Terro. 2013. Tau protein kinases: involvement in Alzheimer's disease. *Ageing Res. Rev.* 12, 1 (Jan 2013), 289–309.

[40] Anastasios Mastrokolias, Yavuz Ariyurek, Jelle J Goeman, Erik van Duijn, Raymund AC Roos, Roos C van der Mast, GertJan B van Ommen, Johan T den Dunnen, Peter AC't Hoen, and Willeke MC van Roon-Mom. 2015. HuntingtonâĂŹs disease biomarker progression profile identified by transcriptome sequencing in peripheral blood. *European Journal of Human Genetics* 23, 10 (2015), 1349–1356.

[41] A. Monsonego, A. Nemirovsky, and I. Harpaz. 2013. CD4 T cells in immunity and immunotherapy of Alzheimer's disease. *Immunology* 139, 4 (Aug 2013), 438–446.

[42] I. Munoz-Sanjuan and G. P. Bates. 2011. The importance of integrating basic and clinical research toward the development of new therapies for Huntington disease. *J. Clin. Invest.* 121, 2 (Feb 2011), 476–483.

[43] L. C. Murrin, J. D. Sanders, and D. B. Bylund. 2007. Comparison of the maturation of the adrenergic and serotonergic neurotransmitter systems in the brain: implications for differential drug effects on juveniles and adults. *Biochem. Pharmacol.* 73, 8 (Apr 2007), 1225–1236.

[44] K. Noto, C. Brodley, and D. Slonim. 2010. Anomaly Detection Using an Ensemble of Feature Models. *Proc IEEE Int Conf Data Min* (Dec 2010), 953–958.

[45] K. Noto, C. Brodley, and D. Slonim. 2012. FRaC: a feature-modeling approach for semi-supervised and unsupervised anomaly detection. *Data Min Knowl Discov* 25, 1 (2012), 109–133.

[46] T. M. Przytycka, M. Singh, and D. K. Slonim. 2010. Toward the dynamic interactome: it's about time. *Brief. Bioinformatics* 11, 1 (Jan 2010), 15–29.

[47] M. F. Ramoni, P. Sebastiani, and I. S. Kohane. 2002. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. U.S.A.* 99, 14 (Jul 2002), 9121–9126.

[48] P. Resnik. 1999. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research (JAIR)* 11 (1999), 95–130.

[49] E. R. Ritvo, A. Yuwiler, E. Geller, E. M. Ornitz, K. Saeger, and S. Plotkin. 1970. Increased blood serotonin and platelets in early infantile autism. *Arch. Gen. Psychiatry* 23, 6 (Dec 1970), 566–572.

[50] J. M. Rubio-Perez and J. M. Morillas-Ruiz. 2012. A review: inflammatory process in Alzheimer's disease, role of cytokines. *ScientificWorldJournal* 2012 (2012), 756357.

[51] E. Sefer, M. Kleyman, and Z. Bar-Joseph. 2016. Tradeoffs between Dense and Replicate Sampling Strategies for High-Throughput Time Series Experiments. *Cell Syst* 3, 1 (Jul 2016), 35–42.

[52] C.E. Shannon. 1948. A mathematical theory of communication (Part I). *Bell Syst Tech J* 27 (1948), 379–423.

[53] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[54] Sanjana Sood, Iain J Gallagher, Katie Lunnon, Eric Rullman, Aoife Keohane, Hannah Crossland, Bethan E Phillips, Tommy Cederholm, Thomas Jensen, Luc JC van Loon, et al. 2015. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome biology* 16, 1 (2015), 185.

[55] D. Spies and C. Ciaudo. 2015. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput Struct Biotechnol J* 13 (2015), 469–477.

[56] Oliver Stegle, Katherine J Denby, Emma J Cooke, David L Wild, Zoubin Ghahramani, and Karsten M Borgwardt. 2010. A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology* 17, 3 (2010), 355–367.

[57] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102, 43 (2005), 15545–15550.

[58] Ann E Tilley, Ben-Gary Harvey, Adriana Heguy, Neil R Hackett, Rui Wang, Timothy P O'connor, and Ronald G Crystal. 2009. Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine* 179, 6 (2009), 457–466.

[59] Randall D. Tobias. 1995. An introduction to partial least squares regression. In *SUGI: Proceedings of the 20th Annual SAS User's Group International meeting.* Orlando, Florida, 1250–7.

[60] M. Trindade, W. Oigman, and M. Fritsch Neves. 2017. Potential Role of Endothelin in Early Vascular Aging. *Curr Hypertens Rev* 13, 1 (2017), 33–40.

[61] K. S. van der Geest, W. H. Abdulahad, S. M. Tete, P. G. Lorencetti, G. Horst, N. A. Bos, B. J. Kroesen, E. Brouwer, and A. M. Boots. 2014. Aging disturbs the balance between effector and regulatory CD4+ T cells. *Exp. Gerontol.* 60 (Dec 2014), 190–196.

[62] Erik van Duijn, Elisabeth M Kingma, Reinier Timman, Frans G Zitman, Aad Tibben, Raymund AC Roos, and Rose C van der Mast. 2008. Cross-sectional study on prevalences of psychiatric disorders in mutation carriers of Huntington’s disease compared with mutation-negative first-degree relatives. *The Journal of clinical psychiatry* 69, 11 (2008), 1–478.

[63] A. L. Varigonda, E. Jakubovski, M. J. Taylor, N. Freemantle, C. Coughlin, and M. H. Bloch. 2015. Systematic Review and Meta-Analysis: Early Treatment Responses of Selective Serotonin Reuptake Inhibitors in Pediatric Major Depressive Disorder. *J Am Acad Child Adolesc Psychiatry* 54, 7 (Jul 2015), 557–564.

[64] Hongen Wei, Ian Alberts, and Xiaohong Li. 2014. The apoptotic perspective of autism. *International Journal of Developmental Neuroscience* 36 (2014), 13–18.

[65] Heather C Wick, Harold Drabkin, Huy Ngu, Michael Sackman, Craig Fournier, Jessica Haggett, Judith A Blake, Diana W Bianchi, and Donna K Slonim. 2014. DFLAT: functional annotation for human development. *BMC bioinformatics* 15, 1 (2014), 45.

[66] Herman Wold. 1985. Partial least squares. *Encyclopedia of statistical sciences* (1985).

[67] W. Yeo and J. Gautier. 2004. Early neural cell death: dying to become neurons. *Dev. Biol.* 274, 2 (Oct 2004), 233–244.

[68] N. Yosef and A. Regev. 2011. Impulse control: temporal dynamics in gene transcription. *Cell* 144, 6 (Mar 2011), 886–896.

[69] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26, 7 (2010), 976–978.

[70] A. Zeidel, B. Beilin, I. Yardeni, E. Mayburd, G. Smirnov, and H. Bessler. 2002. Immune response in asymptomatic smokers. *Acta Anaesthesiol Scand* 46, 8 (Sep 2002), 959–964.

[71] J. Zhou and L. F. Parada. 2012. PTEN signaling in autism spectrum disorders. *Curr. Opin. Neurobiol.* 22, 5 (Oct 2012), 873–879.

[72] M. N. Ziats and O. M. Rennert. 2011. Expression profiling of autism candidate genes during human brain development implicates central immune signaling pathways. *PLoS ONE* 6, 9 (2011), e24691.

[73] G. E. Zinman, S. Naiman, Y. Kanfi, H. Cohen, and Z. Bar-Joseph. 2013. ExpressionBlast: mining large, unstructured expression databases. *Nat. Methods* 10, 10 (Oct 2013), 925–926.

APPENDIX A

We suspected that, compared to linear regression, PLSR would be better able to handle the dimensions and redundancy of gene expression data [59]. We also considered using Support vector regression (SVR) models with either a linear or radial kernel function [21, 53]. On both the ASD autism data set and on an additional developmental data set from GEO (GSE32472), we evaluated the predictive performance of linear regression (LR), PLSR, and SVR models, trained on all control samples using all genes in leave-one-out cross validation. We implemented both LR and SVR in R, the former via the `lm()` function from the `stats` package, the latter in the `e1071` package with default settings.

As hypothesized, linear regression predicted ages less accurately than other methods. On both data sets, PLSR models had lower Mean Squared Error than either SVR model (see Table 5). However, we did not explore the space of possible parameters for SVR.

Table 5: Mean squared errors for PLSR, SVR with both linear and radial kernels, and linear models on all control samples in two data sets.

Method	ASD MSE	GSE32472 MSE
PLSR	3.65	2.33
SVR (linear kernel)	4.12	4.26
SVR (radial kernel)	4.29	4.25
Linear Regression	710.74	32.72

APPENDIX B

In Section 2.1.2, we model prediction errors using a normal distribution. To test this assumption, we assessed normality using the Shapiro-Wilk normality test in R. On many data sets, we found that the observed error distributions on the control set are in fact normal for nearly all gene sets. However, on some data sets, a substantial fraction of the gene sets have slightly skewed error distributions that do not pass the criteria for normality. We believe that such skewing arises from a lack of uniformity in the age distribution of the control samples.

Some regression models rely on the assumption of normality. However, PLSR is considered relatively robust to data that do not fit this assumption [25]. We found that, although there is a modest negative correlation (-0.31) between the normality of the residuals for a gene set and the accuracy of that gene set’s model on the training samples (Figure 6), there are many high-quality models with non-normal residuals.

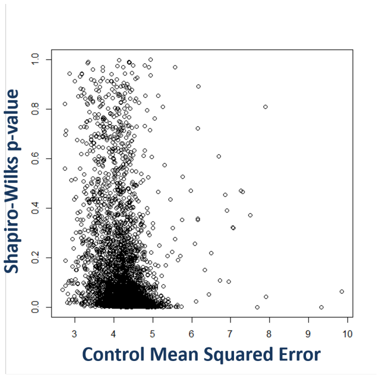

Figure 6: Plot of model quality (control mean squared error) vs. model normality (p-value from the Shapiro-Wilk test, with low p-values *rejecting* normality). Each dot represents a gene set; dots with low MSE and low p-values represent relatively accurate models that fail to meet the criteria for normally distributed residuals.

We emphasize that even in these cases, the non-normality of the prediction errors does not appreciably affect our results. This is because the scoring function does not make use of any specific properties of the normal distribution.

To verify this, we assessed performance of an alternative, non-parametric scoring function:

$$Score'(G) = \frac{|C| \sum_{s \in D}(E_{G,s})^2}{|D| \sum_{s \in C}(E_{G,s})^2} \times L_1 \times L_2 \qquad (2)$$

This score is the ratio of the mean squared errors for the disease and control sets. Using this scoring function returns nearly identical ranked lists of gene sets (Spearman rank correlation ≥ .99 between *Score* and *Score′* on all data sets used in this manuscript). We conclude that even if the error distributions are somewhat skewed, the surprisal probabilities are close enough to those expected from the normal distribution that our scoring function is capturing the intended relationship between prediction accuracies in the control and disease sample sets.