Pacific Symposium on Biocomputing 25:683-694(2020)

aTEMPO: Pathway-Specific Temporal Anomalies for Precision Therapeutics

Christopher Michael Pietras[†1], Liam Power[2], and Donna K. Slonim[1,2]

1. Computer Science, Tufts University,
Medford, MA 02155, USA
2. Genetics, Sackler School of Graduate Biomedical Sciences
Tufts University School of Medicine
Boston, MA 02111, USA
[†]*E-mail: christopher.pietras@tufts.edu*

Dynamic processes are inherently important in disease, and identifying disease-related disruptions of normal dynamic processes can provide information about individual patients. We have previously characterized individuals' disease states via pathway-based anomalies in expression data, and we have identified disease-correlated disruption of predictable dynamic patterns by modeling a virtual time series in static data. Here we combine the two approaches, using an anomaly detection model and virtual time series to identify anomalous temporal processes in specific disease states. We demonstrate that this approach can informatively characterize individual patients, suggesting personalized therapeutic approaches.

Keywords: Anomaly detection, precision medicine, temporal expression modeling

1. Introduction

Predictable dynamic processes characterize many biological states, from development to the cell cycle to healthy aging. We are interested in using these processes to identify unusual disease-related temporal dynamics that may offer potentially actionable information about individual patients. The most prevalent data source, although by no means the only possible one, for such a project is bulk expression data, which is readily available for a variety of dynamic disease states. Transcriptomic profiles need not explicitly include time series; instead, we can create virtual time series from static data, so long as age or temporal information is available. We apply this approach to both developmental and degenerative disorders in which there are normal age-related changes to discover patient-specific temporal disruptions of expression that are associated with disease.

This problem is related to *anomaly detection*,[1] a machine learning paradigm in which rare anomalous events are detected using only normal (or mostly normal) training samples. Common applications of such methods include fraud, intrusion, or spam detection. In a biological context, anomaly detection can be used to characterize specific patients with rare or heterogeneous disorders.

However, the dimensions of transcriptomic data sets - typically at least tens of thousands of features, representing genes or transcripts, but much smaller numbers of samples - often make

traditional anomaly detection approaches perform poorly. Furthermore, typically only a small fraction of genes provide phenotypic information about the underlying samples. Fortunately, gene expression also has a great deal of underlying structure. We can thus use known domain information, such as previously-defined sets of functionally related genes, to make learning on these data sets possible.

There has been substantial prior work applying anomaly detection to bioinformatics problems. Previous efforts relating to expression data have explored identifying differentially expressed genes or gene sets.[2–5] Only recently have other methods begun to explore outliers in the context of characterizing individual samples,[6,7] but none of these methods considers outlier individuals in a temporal context.

Our group previously introduced Feature Regression and Classification (FRaC),[8] a robust feature prediction approach for the general anomaly detection problem, and showed that it is more robust to irrelevant variables than top competing methods.[9] We then used FRaC as a component of CSAX,[10] a pathway-based method for identifying and interpreting anomalies in individual gene expression samples. CSAX addresses the issue that to learn something clinically useful about an individual patient, is often not enough to simply know *that* a given sample is anomalous - we also need to know *why*. Such information might distinguish different disease subtypes, provide insight into the mechanisms of disease, or provide additional information about a particular patient's condition.

Several additional methods characterize individual samples via enrichment using molecular signatures, including ssGSEA (single sample gene set enrichment analysis),[11] GSVA (gene set variation analysis),[12] PLAGE (pathway level analysis of gene expression),[13] and combining z-scores.[14] However, none of these identifies breakdowns in expected temporal dynamics.

For this purpose, we previously developed TEMPO,[15] a temporal modeling method for characterizing time-related expression dysregulation in disease. TEMPO works by finding gene sets where there is a good predictive model of age or time as a function of pathway-specific gene expression in healthy patients that breaks down consistently in patients with a disease or phenotype of interest.

Here we introduce anomaly-TEMPO (aTEMPO), an extension of TEMPO for *anomaly detection* incorporating temporal dysregulation in disease. Using insights from Cousins *et al.*,[16] in which we demonstrated that it is possible to perform similarly accurate anomaly detection to FRaC while modeling only a small subset of the features, we use aTEMPO to model time using expression of limited numbers of functionally related genes. Modeling time in this way removes the scaling dependency on the size of the feature space, making aTEMPO practical to run on large data sets. This allows for the significant runtime improvements of Cousins *et al.*[16] while retaining CSAX's ability to interpret anomalies in individual patients' samples. We compare aTEMPO to original TEMPO as well as to general-purpose anomaly detection algorithms and to previous anomaly detection methods designed for expression data, and we show that in its specific domain, aTEMPO is consistently better at anomaly detection than comparator algorithms. We further show how aTEMPO uncovers medically relevant insights about individual patient samples that other non-temporal methods do not.

2. Methods

2.1. *aTEMPO*

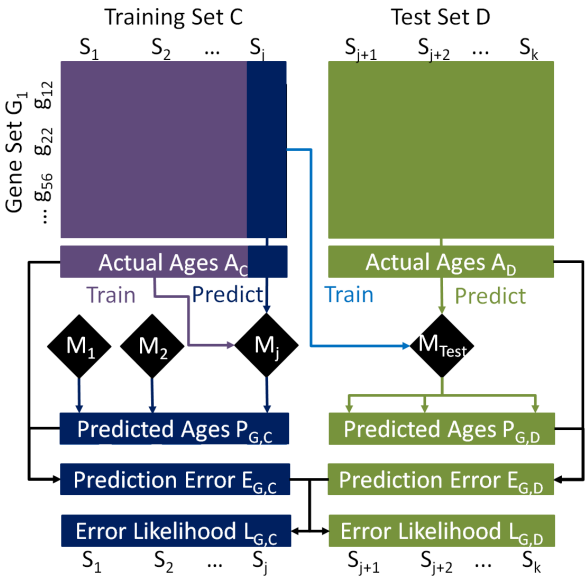

Fig. 1. PLSR prediction for an arbitrary gene set G_1. Error likelihoods $L_{G_1,D}$ are used as anomaly scores for gene set G_1.

As in TEMPO,[15] for a gene set G, aTEMPO trains a partial least squares regression (PLSR) model,[17] using the *pls* package in R, to predict age as a function of the 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), identically to the procedure used by TEMPO. Ages for test samples $D = \{S_{j+1}, \ldots, S_k\}$ are predicted using M_{Test}, trained on all the control samples.

We obtain a vector of prediction errors for G, 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. 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 at least as large as x under the normal distribution with mean μ_G and standard deviation σ_G. The error likelihood for gene set G and sample s is then:

$$L_{G,s} = -log(\mathcal{N}_G(E_{G,s})) \tag{1}$$

Rather than using these error likelihoods to compute a dysregulation score for a gene set, we can instead consider it as an anomaly score. Given a collection of gene sets, we will obtain a matrix of anomaly scores with one entry for each gene set and test set sample. To obtain a single anomaly score for sample s, we simply sum over all gene sets:

$$A_s = \sum_{i=1}^{n} L_{G_i,s} \tag{2}$$

where a larger A_s indicates a higher likelihood that the sample is anomalous. We note that A_s could be modified by changing which gene sets are included in the summation, perhaps by restricting to known disease-associated gene sets, or to only those gene sets with significantly predictive models under the criteria used in TEMPO.

2.2. *Single Sample aTEMPO*

We evaluate aTEMPO's single sample characterization by constructing a training set of all normal samples and a test set of all disease samples. aTEMPO produces an anomaly score for each disease sample and gene set, which we use as the single-sample score.

2.3. *Other Methods*

We compared aTEMPO to top-performing general anomaly detection methods Local Outlier Factor[18] and one-class support vector machines,[19] as well as gene expression specific methods FRaC and CSAX. We use the same implementations of FRaC, LOF, and one-class SVMs as described in Noto et al. (2015).[10] The CSAX implementation is largely identical as well, although minor changes were made to prevent extraneous bagging in instances where additional bagging would make no significant difference to the final anomaly scores (results not shown).

To further test whether considering gene sets in a temporal context improves accuracy over considering only individual genes, we also compared to a version of FRaC that used only the anomaly score for the age feature.

We compared aTEMPO's single sample output to ssGSEA, PLAGE, GSVA, and z-scores using the GSVA package in R. However, as all four of these packages produce enrichment scores - where both high and low values also indicate an enrichment - scores produced by these methods are not directly comparable to aTEMPO scores, where a large value indicates an anomaly and a low value does not. In these cases, we compare gene sets with negative enrichment scores and gene sets with positive enrichment scores to the corresponding aTEMPO scores separately.

Connectivity mapping[20] is a technique by which gene expression changes due to a disease or phenotype can be matched to drugs that reverse observed expression changes in a set of drug-treated cells. To evaluate patient-specific observations derived from aTEMPO, we queried the Clue interface[21] for drug connectivity using the 150 most up- and down-regulated genes in the given sample or in the consensus signature for a set of samples.

2.4. *Expression Data Sets*

We evaluate each algorithm on expression data sets for human diseases where age is a reported and disease-relevant variable, the first four of which were described in Pietras *et al.*:[15]

Autism spectrum disorder (ASD): Based on a study by Alter, *et al.*[22] (GSE25507 in the Gene Expression Omnibus (GEO) database[23]), this data set includes expression microarrays characterizing peripheral blood lymphocytes of 72 children with autism spectrum disorders and 59 controls, with ages ranging from 2 to 14 years.

Huntington's disease (HD): This data set includes normalized gene counts from an

RNASeq experiment characterizing blood from 91 Huntington's disease carriers, 27 of whom are pre-symptomatic, and 33 similar-aged controls[24] (GSE51779 in GEO). Samples are annotated with patient ages in years to .01 precision.

Alzheimer's disease (AD): We include all samples from Batch 1 marked as "included in the case-control study" in an Illumina beadchip data set by Sood, *et al.*[25] from the AddNeuroMed consortium[26] (GSE63060 in GEO). It contains 49 samples characterizing peripheral blood in Alzheimer's patients and 67 from similar-aged controls, with ages in integer years.

Chronic obstructive pulmonary disease (COPD): This data set includes small airway gene expression microarray data from 15 smokers with COPD and 12 smokers who are apparently healthy, derived from studies by Carolan, *et al.*[27] and Tilley, *et al.*[28] (GSE5058 in GEO). Each patient has an integer age in years.

Bronchopulmonary dysplasia (BPD): Drawn from a study on preterm birth complications by Pietrzyk, *et al.*[29] (GSE32472 in GEO), this data set includes microarray data profiling peripheral blood of 66 five-day-old infants born preterm and later diagnosed with BPD, plus 35 controls who did not develop BPD but may have had other complications. Gestational ages at birth range from 22 to 33 weeks.

2.5. *Gene Set Collections*

For aTEMPO and CSAX, we used the same Gene Ontology[30] gene sets used in Pietras *et al.*[15] for the four data sets originally described in that paper. This collection excludes all gene sets of size greater than 500 or less than 5, resulting in a total of 6079 gene sets. For BPD, we used DFLAT[31] biological process gene sets, generated October 2017, excluding gene sets of size greater than 500 or less than 10, resulting in 4917 gene sets.

2.6. *Anomaly Detection Tasks*

We evaluate aTEMPO and each of the comparator methods on five replicates of "semi-supervised" anomaly detection tasks, where we train on normal samples only and evaluate on normal and abnormal test data. Specifically, for each data set and replicate, we generate a training set consisting of a randomly-selected $\frac{2}{3}$ of the known normal data, and a test set consisting of the remaining $\frac{1}{3}$ of the normals and all of the anomalies (corresponding to disease states). Each anomaly detection algorithm is given access to the expression data and ages for each sample, and asked to produce an anomaly score for each test set sample. We calculate the Area Under the ROC Curve (AUC) for each replicate and algorithm to assess each method's ability to distinguish normals from anomalies.

3. Results

Full results for all data sets and methods on anomaly detection and single sample tasks, as well as several supplemental tables, are available on the project website at bcb.cs.tufts.edu/atempo.

3.1. *Anomaly Detection*

On the semi-supervised anomaly detection task, aTEMPO outperformed competitors on all five data sets (Table 1). While there is often a close competitor, no other individual method

is consistently strong on all five data sets. Performance is best on the COPD or BPD data sets for all methods. This observation likely partially reflects the fact that COPD is the data set analyzing the most disease-relevant tissue (small airway cells rather than blood).

Table 1. Average anomaly detection AUCs over five replicates, with standard deviations, for each method. Top scores per data set are in bold.

	aTEMPO	FRaC	FRaC, age feature	CSAX	LOF	One-class SVM
ASD	**0.692**	0.531	0.683	0.534	0.533	0.500
	(0.025)	(0.063)	(0.034)	(0.077)	(0.060)	(0.000)
AD	**0.615**	0.586	0.493	0.507	0.605	0.526
	(0.030)	(0.050)	(0.019)	(0.044)	(0.047)	(0.025)
COPD	**0.987**	**0.987**	0.710	0.854	0.707	0.500
	(0.030)	(0.018)	(0.147)	(0.130)	(0.042)	(0.000)
HD	**0.622**	0.468	0.616	0.491	0.550	0.506
	(0.070)	(0.045)	(0.038)	(0.034)	(0.056)	(0.034)
BPD	**0.845**	0.758	0.816	0.714	0.729	0.686
	(0.025)	(0.062)	(0.039)	(0.080)	(0.033)	(0.048)

We had additional phenotypic information for two data sets. Patients with the Huntington's genotype were either symptomatic or asymptomatic at the time of the study, and BPD patients were designated as having either mild, moderate, or severe BPD. We additionally evaluated anomaly detection to identify only members of these subgroups (full results online). Here, aTEMPO was still the best-performing method overall, though again there was often a close competitor and aTEMPO was slightly outperformed by another method in two cases. We also found that aTEMPO's ability to identify disease cases as anomalous increased with disease severity, a pattern that holds for all five comparator methods as well. However, no method performed well at identifying pre-symptomatic Huntington's patients as atypical when trained on normal controls.

Nonetheless, aTEMPO and CSAX can potentially mitigate this problem by using only a subset of the gene-set-specific anomaly scores for each patient. Both aTEMPO and CSAX compute a single anomaly score for each sample that is the sum of anomaly scores for each gene set and sample. Instead of using *all* gene sets, we could instead compute an anomaly score using only a subset of the gene sets - ideally those which we know to be related to the disease of interest.

For example, on the HD data set, semi-supervised anomaly detection of asymptomatic Huntington's disease is poor, with the most accurate method obtaining an AUC of only 0.55. However, by using only those gene sets found to be dysregulated by TEMPO when comparing the normal patients in each replicate to the symptomatic patients, we are able to obtain far better performance even on the asymptomatic patients. The average AUC of aTEMPO improves from 0.520 to 0.727 using this approach, and the average performance of CSAX

changes from 0.469 to 0.610. We conclude that in domains where we have some prior knowledge about the types of anomalies we are looking for - for example, early detection of a known disease - we can likely increase the predictive power of aTEMPO.

3.2. *Single-sample aTEMPO for precision medicine*

3.2.1. *Individual aTEMPO results differ from TEMPO*

There is significant variation in suprisal values amongst individual patients with a disease. While TEMPO reports a single dysregulation score for each gene set that captures the overall trend of dysregulation in that gene set, the suprisal value for an individual patient might be quite different. Supplemental Table 3 shows the average of the rank correlations between the TEMPO dysregulation scores for each gene set and each disease sample's anomaly scores for each gene set. There is generally a mildly positive rank correlation (0.07 - 0.43) between the typical sample's results and the TEMPO results across all samples, but the variance is quite high. The correlation can even be substantially negative for individual samples (e.g. below -0.5). Cases in which individual patients' dysregulation in a particular gene set differs substantially from an overall disease average might offer especially valuable opportunities for precision therapeutics.

3.2.2. *aTEMPO differs from non-temporal single-sample methods*

Overall, correlations between the scores for either the positively or negatively enriched gene sets from any of the comparator methods for a given patient and the aTEMPO scores for the corresponding gene sets for that patient are low, as can be seen in Table 2 (correlations shown are for BPD - tables for other data sets are similar, and available online). Correlations of results between the different enrichment-based methods are more substantial, with the exception of ssGSEA. This illustrates that temporal modeling produces novel and distinct results from those found by single-sample enrichment methods.

Table 2. Average correlations between single sample score vectors for the same patient across methods in BPD. Correlations between aTEMPO and enrichment-based methods include only gene sets that are either positively or negatively enriched in the comparator method.

	ssGSEA	Plage	Zscore	aTEMPO (+)	aTEMPO (−)
GSVA	0.102	-0.212	0.829	0.062	-0.052
ssGSEA		-0.029	0.105	0.043	-0.08
Plage			-0.156	0.018	-0.008
Zscore				-0.004	-0.016

3.2.3. *Sample heterogeneity in bronchoplumonary dysplasia (BPD)*

For BPD, we used subjects' disease severity and sex to examine how single sample aTEMPO characterizes sample heterogeneity and disease subtypes. To consider broader patterns of heterogeneity that may be related to clinical properties of the BPD patients, we identified gene sets defining large distinct subgroups within the aTEMPO results.

Specifically, we consider gene sets where the anomaly score for that gene set is one of the 100 highest for at least a fifth of the samples, and where it is *not* one of the 100 highest for at least another fifth of the samples, suggesting that there are sub-populations of BPD cases with different patterns. There are 123 such gene sets in the aTEMPO results, many of which are related to vascular development, neurodevelopment, and hearing. Fifty three of these gene sets are significantly related to BPD severity (ANOVA p-values $\leq .05$). These include two vascular gene sets ("vasculogenesis" and "aorta morphogenesis"), six neurodevelopmental gene sets, and four hearing-related gene sets (including "cochlea development" and "inner ear morphogenesis"). These associations were not found by any of the static single sample comparator methods.

While BPD severity has previously been associated with male sex,[32] only a single one of the 123 gene sets is significantly associated with sex. However, average anomaly scores for males are higher than those for females (5.07 vs 4.67) and the average intra-class rank correlations for the anomaly score vectors are lower in males than females (0.17 vs 0.25). This suggests that the primary difference in between sexes might not be any specific gene set, but rather a greater lack of uniformity in progression of BPD in males.

The observed dysregulation of vasculogenesis and its correlation with BPD severity is particularly interesting, because vascular development is known to be disordered in infants with BPD, but whether this is a cause or consequence of the disorder, or of common co-morbidities such as sepsis or retinopathy of prematurity, is the subject of active debate.[33] BPD is also known to correlate with neurodevelopmental delays in the toddler and preschool years,[34,35] but the cause of this correlation is again not known, and it is thought that infection might play a role in causing both. Hearing loss has also been associated with BPD,[36] but the correlation has been putatively attributed to ototoxicity from postnatal antibiotic use. The association of these functions at five days of life with the severity of disease in a *future* diagnosis of BPD hints at other possible and perhaps actionable causes.

3.2.4. *Huntington's disease anomalies suggest precision therapeutics*

The gene set with the highest average aTEMPO anomaly score in HD patients is "negative regulation of DNA recombination," for which three patients had surprisal scores more than three standard deviations above the mean. The gene set with the third highest average suprisal is "positive regulation of sodium ion transmembrane transport." The aTEMPO scores for these gene sets are moderately correlated (0.49) across samples, even though the gene sets are completely distinct.

Recombination is invaluable in repairing the double stranded breaks (homologous recombination) implicated in HD and other trinucleotide repeat disorders. Acidosis, thought to be a product of impaired energy metabolism in HD brains that leads to CNS lactate build up

and disruption of acid-sensing ion channels, has been observed in HD models in vivo and in vitro as well as in patient brains.[37] These findings suggest that modulation of ion channels may be of therapeutic benefit in the subset of patients with high surprisal scores in both DNA recombination and sodium ion transport pathways, as temporally dysregulated ion channels may have an influence on regulation of recombination and DNA repair pathways.

To assess precision medicine opportunities for those pathways suggested by aTEMPO to be developmentally anomalous, we queried the Broad's connectivity database (`https://clue.io/`). Connectivity mapping in these three individual patients indeed suggests drugs affecting ion transport. ATPase inhibitors are also over-represented among the top hits in several of these patients. Sodium-related ATPase activity is known to be disrupted in Huntington's patients, likely causing aberrant mitochondrial function.[38]

3.2.5. *Alzheimer's disease, tyrosine phosphorylation, and cholinergic balance*

The gene sets with the highest average surprisal in the aTEMPO results from AD patients were "peptidyl-tyrosine modification" and "phosphatidylcholine metabolic process." In particular, there were six AD patients for whom "phosphatidylcholine metabolic process" and related gene sets were highly anomalous (aTEMPO scores of at least ten).

Connectivity mapping queries for the signatures of these patients reveal acetylcholine receptor antagonists among the top suggested drugs for four of the six patients, including the single most-connected drug for one patient. Acetylcholine receptors regulate phosphotidylcholine levels; there is a balance in the cholinergic system that needs to be maintained.[39] Acetylcholine receptor antagonists have previously been proposed as AD therapeutics.[40]

In addition, retinol, which has also been suggested as a means to restore cholinergic balance in AD,[41,42] was the drug with the highest average negative connectivity score across these six patients, with strongly negative scores in four of the six. Again, the aTEMPO results seem to be identifying specific patients as candidates for suggested therapeutic approaches that have not yet proven to be universally effective.

Further, analyses of post-mortem AD brain tissue has revealed elevated levels of phosphotyrosine protein and reduced specific activity of protein tyrosine kinases. These studies have suggested that tyrosine and phosphatase systems may be important in AD pathogenesis.[43,44] The results of aTEMPO support these hypotheses, suggesting that the systems of tyrosine phosphorylation and modification are dysregulated in some patients.

The intracellular neurofibrillary tangles involved in AD are made of the microtubule protein *tau* that is abnormally tyrosine phosphorylated and interacts with tyrosine kinases. The results from aTEMPO suggest that the subset of patients with the highest surprisal scores in these pathways may benefit from therapies that target the phosphorylation of the *tau* protein with kinase inhibitors.[45,46] Few such therapeutics appear in the public connectivity database, so it was not possible to evaluate this hypothesis using connectivity queries. However, given the prior connectivity results, the hypothesis that aTEMPO is identifying candidates for such *tau*-targeting compounds seems plausible.

4. Discussion

Anomaly detection, which can functionally characterize how a sample's expression or other genomic data differs from a set of normal control samples, is proving to be a promising paradigm for precision medicine. We introduced a model that finds predictable age-related pathway expression patterns and that identifies anomalous cases in which those patterns break down. The highlighted pathways differ from those found by static single-sample methods or by the original TEMPO approach. Note that in principle, there is no reason the temporal data needs to represent patient age - it could be time since study inception, or even some completely different ordinal variable, provided it is relevant to both the control and disease states.

We acknowledge several limitations of our approach. Crucially, we have no independent clinical data from these patients, so we have no way of experimentally validating that aTEMPO gene sets truly suggest effective precision therapeutics in these studies. The connectivity data that we used to provide some concurring evidence is derived primarily from cancer cells, whose relevance to disease-relevant cells in the diseases we consider is variable. We also note that we used the same expression data for the connectivity queries and the aTEMPO analysis, so this confirmation is not truly independent. However, the data are used in a completely different way, so finding suggested therapeutics that reflect the temporal dysregulation patterns identified by aTEMPO is still somewhat confirmatory. Nonetheless, true experimental validation awaits future work.

Simulation would also provide a method for experimentally validating aTEMPO results. However, while there are established and accepted methods for generating simulated static expression data sets (e.g., Law et. al.[47]), simulating data with time-related patterns for aTEMPO while remaining a fair comparison for static methods is nontrivial. Development of such simulation methods is an important direction for future work.

The PLSR models currently constructed by aTEMPO are currently extremely simple. This is necessary to reduce risk of overfitting when training models of up to 500 features with as few as fifteen training samples. Massive multi-omics data sets might allow training of more complicated and more accurate models, potentially incorporating other clinical variables. Such data sets may also make libraries of pre-trained models practical, which would allow meaningful aTEMPO characterization to be obtained for data sets consisting of only a single patient, assuming data consistency issues could be addressed.

In the supplemental material, we explored omitting gene sets from the anomaly score if their models were not significantly predictive. While this particular technique did not significantly improve over normal aTEMPO, exploring other weighting schemes for A_s might prove a valuable direction for future work.

Finally, one advantage that enrichment-based single sample methods like GSVA and ssGSEA have over characterizing single samples via anomaly detection is that they allow for directionality. A high aTEMPO anomaly score for a gene set merely indicates that the gene set's normal developmental pattern is somehow dysregulated, but doesn't directly indicate how it is dysregulated. Future efforts should explore this issue as well.

Acknowledgements and funding

We thank Lenore Cowen, Faith Ocitti, Keith Noto, Cyrus Cousins, and other current and former members of the Tufts BCB Group for their contributions to and comments on this and related work. This project was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number R01HD076140 (DKS, CMP) and by NIH award 5T32GM008448 to Tufts University School of Medicine for MD/PhD student research support (LP).

References

1. V. Chandola, A. Banerjee and V. Kumar, *ACM Comput. Surv.* **41**, 1 (2009).
2. D. Ghosh, *Journal of biopharmaceutical statistics* **20**, 193 (2010).
3. S. Karrila, J. H. E. Lee and G. Tucker-Kellogg, *Cancer informatics* **10**, CIN (2011).
4. L. Li, A. Chaudhuri, J. Chant and Z. Tang, *Physiological genomics* **32**, 154 (2007).
5. J. P. Mpindi, H. Sara, S. Haapa-Paananen, S. Kilpinen, T. Pisto, E. Bucher, K. Ojala, K. Iljin, P. Vainio, M. Björkman *et al.*, *PloS one* **6**, p. e17259 (2011).
6. Y. Zeng, G. Wang, E. Yang, G. Ji, C. L. Brinkmeyer-Langford and J. J. Cai, *PLoS genetics* **11**, p. e1004942 (2015).
7. L. Jiang, H. Chen, L. Pinello and G.-C. Yuan, *Genome biology* **17**, p. 144 (2016).
8. K. Noto, C. Brodley and D. Slonim, in *2010 IEEE International Conference on Data Mining*, 2010.
9. K. Noto, C. Brodley and D. Slonim, *Data Min Knowl Discov* **25**, 109 (2012).
10. K. Noto, S. Majidi, A. G. Edlow, H. C. Wick, D. W. Bianchi and D. K. Slonim, *Journal of Computational Biology* **22**, 402 (2015).
11. D. A. Barbie, P. Tamayo, J. S. Boehm, S. Y. Kim, S. E. Moody, I. F. Dunn, A. C. Schinzel, P. Sandy, E. Meylan, C. Scholl *et al.*, *Nature* **462**, p. 108 (2009).
12. S. Hänzelmann, R. Castelo and J. Guinney, *BMC bioinformatics* **14**, p. 7 (2013).
13. J. Tomfohr, J. Lu and T. B. Kepler, *BMC bioinformatics* **6**, p. 225 (2005).
14. E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker and D. Lee, *PLoS computational biology* **4**, p. e1000217 (2008).
15. C. M. Pietras, F. Ocitti and D. K. Slonim, in *ACM-BCB*, 2018. Updated in bioXriv (2019): https://doi.org/10.1101/651018.
16. C. Cousins, C. M. Pietras and D. K. Slonim, in *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, 2017.
17. H. Wold, *Encyclopedia of statistical sciences* **6** (1985).
18. M. Breunig, H. Kriegel, R. Ng and J. Sander, *ACM SIGMOD Records* **29**, 93 (2000).
19. B. Scholkopf, A. Smola, R. Williamson and P. Bartlett, *Neural Computing* **12**, 1207 (2000).
20. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, *science* **313**, 1929 (2006).
21. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu *et al.*, *Cell* **171**, 1437 (2017).
22. M. D. Alter, R. Kharkar, K. E. Ramsey, D. W. Craig, R. D. Melmed, T. A. Grebe, R. C. Bay, S. Ober-Reynolds, J. Kirwan, J. J. Jones *et al.*, *PloS ONE* **6**, p. e16715 (2011).
23. R. Edgar, M. Domrachev and A. E. Lash, *Nucleic acids research* **30**, 207 (2002).
24. A. Mastrokolias, Y. Ariyurek, J. J. Goeman, E. van Duijn, R. A. Roos, R. C. van der Mast, G. B. van Ommen, J. T. den Dunnen, P. AC't Hoen and W. M. van Roon-Mom, *European Journal of Human Genetics* **23**, 1349 (2015).
25. S. Sood, I. J. Gallagher, K. Lunnon, E. Rullman, A. Keohane, H. Crossland, B. E. Phillips,

T. Cederholm, T. Jensen, L. J. van Loon *et al.*, *Genome biology* **16**, p. 185 (2015).
26. S. Lovestone, P. Francis, I. Kloszewska, P. Mecocci, A. Simmons, H. Soininen, C. Spenger, M. Tsolaki, B. Vellas, L.-O. Wahlund *et al.*, *Annals of the New York Academy of Sciences* **1180**, 36 (2009).
27. B. J. Carolan, A. Heguy, B.-G. Harvey, P. L. Leopold, B. Ferris and R. G. Crystal, *Cancer research* **66**, 10729 (2006).
28. A. E. Tilley, B.-G. Harvey, A. Heguy, N. R. Hackett, R. Wang, T. P. O'connor and R. G. Crystal, *American journal of respiratory and critical care medicine* **179**, 457 (2009).
29. J. J. Pietrzyk, P. Kwinta, E. J. Wollen, M. Bik-Multanowski, A. Madetko-Talowska, C.-C. Günther, M. Jagła, T. Tomasik and O. D. Saugstad, *PloS one* **8**, p. e78585 (2013).
30. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig *et al.*, *Nature genetics* **25**, 25 (2000).
31. H. C. Wick, H. Drabkin, H. Ngu, M. Sackman, C. Fournier, J. Haggett, J. A. Blake, D. W. Bianchi and D. K. Slonim, *BMC bioinformatics* **15**, p. 45 (2014).
32. Z. Zysman-Colman, G. M. Tremblay, S. Bandeali and J. S. Landry, *Paediatrics & child health* **18**, 86 (2013).
33. K. R. Stenmark and S. H. Abman, *Annu. Rev. Physiol.* **67**, 623 (2005).
34. B. Schmidt, E. V. Asztalos, R. S. Roberts, C. M. Robertson, R. S. Sauve and M. F. Whitfield, *JAMA* **289**, 1124 (Mar 2003).
35. L. J. Schlapbach, M. Adams, E. Proietti, M. Aebischer, S. Grunt, C. Borradori-Tolsa, M. Bickle-Graz, H. U. Bucher, B. Latal, G. Natalucci, G. Zeilinger, A. Capone, F. Steiner, S. Schulzke, P. Weber, G. P. Ramelli, M. Nelle, M. Steinlin, S. Grunt, R. Hassink, W. Bar, E. Keller, C. h. Killer, K. Fuhrer, J. F. Tolsa, M. Bickle-Graz, R. E. Pfister, P. S. Huppi, C. Borradori-Tolsa, T. M. Berger, T. Schmitt-Mechelke, V. Pezzoli, M. Ecoffey, A. Mueller, A. Malzacher, J. P. Micallef, C. h. Schaefer, M. von Rhein, R. Arlettaz Mieth, V. Bernet, B. Latal and G. Natalucci, *BMC Pediatr* **12**, p. 198 (2012).
36. S. Zanchetta, L. A. d. L. Resende, M. R. Bentlin, L. M. Rugulo and C. E. Trindade, *Early human development* **86**, 385 (2010).
37. H. K. Wong, P. O. Bauer, M. Kurosawa, A. Goswami, C. Washizu, Y. Machida, A. Tosaki, M. Yamada, T. Knöpfel, T. Nakamura *et al.*, *Human molecular genetics* **17**, 3223 (2008).
38. A. R. Kumar and P. A. Kurup, *Neurol India* **50**, 174 (Jun 2002).
39. E. Nizri, I. Wirguin and T. Brenner, *Drug news & perspectives* **20**, 421 (2007).
40. J. L. Hoskin, Y. Al-Hasan and M. N. Sabbagh, *Nicotine Tob. Res.* **21**, 370 (Feb 2019).
41. H.-P. Lee, G. Casadesus, X. Zhu, H.-g. Lee, G. Perry, M. A. Smith, K. Gustaw-Rothenberg and A. Lerner, *Expert review of neurotherapeutics* **9**, 1615 (2009).
42. K. Shudo, H. Fukasawa, M. Nakagomi and N. Yamagata, *Current Alzheimer Research* **6**, 302 (2009).
43. J. G. Wood and P. Zinsmeister, *Neuroscience letters* **121**, 12 (1991).
44. I. Shapiro, E. Masliah and T. Saitoh, *Journal of neurochemistry* **56**, 1154 (1991).
45. E. E. Congdon and E. M. Sigurdsson, *Nature Reviews Neurology* **14**, p. 399 (2018).
46. G. Lee, R. Thangavel, V. M. Sharma, J. M. Litersky, K. Bhaskar, S. M. Fang, L. H. Do, A. Andreadis, G. Van Hoesen and H. Ksiezak-Reding, *Journal of Neuroscience* **24**, 2304 (2004).
47. C. W. Law, Y. Chen, W. Shi and G. K. Smyth, *Genome biology* **15**, p. R29 (2014).