Mixtures of Latent Variable Models for Interpreting Gene Expression Covariation from Pathways to Transcriptomes

Thesis by
Nicholas Markarian

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy



CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2026
Defended December 11th, 2025

© 2026

Nicholas Markarian
ORCID: 0000-0003-1347-2392

All rights reserved

ACKNOWLEDGEMENTS

For everything you taught me and for taking a chance on me, thank you, Kathy, Steve, Dr. Waterman, Remo, Rob, Matt, Justin, Niles, Barbara, Lior, and Paul.

And thank you, Paul, for reminding me it can be fun to read.

For all the help you have given me, thank you, Chris, Katie, Calista, Han, Tsu-Pei, Brendan, Fei, Jared, Satya, Ann, Mark, Delaney, Sina, Kimberly, Dustin, Tremayne, Paul, Caleb, Ulrich, Peter, and Alex.

For the memories, the adventures, and for making my workplaces and apartments a home, thank you Ian, Zach, Ara, Kanhai, Collin, Stephanie, Sunny, Sam, Terence, Natalie, Thomas, George, Everett, Nick, Sebby, Mike, Ray, Haseeb, Joy, Glenda, Atishay, Jonathan, Blade, Nate, Claire, Joe, Trevor, Ale, David, James, Olivia, Hillel, Alka, Joani, Austin, Kevin, Ben, and Bentley.

For everything that family means, thank you, Mom, Dad, Seda, Emma, and Emily.

To the village that raised me.

ABSTRACT

This thesis centers on interpretable subspace learning and latent variable models for characterizing covariation modulated by categorical variables in the context of biology. First, it introduces a probabilistic model with ties to Principal Component Analysis and k-means clustering, k-spaces, which has implications across different biological analyses through its interpretations as a subspace learning technique, a latent variable model, and a dimension reduction technique. Second, it establishes the problem of simultaneously characterizing gene covariation and expression level in known pathways in human tissue samples and applies k-spaces to GTEx data to lay the foundations for this line of research. Finally, it outlines a path forward to being able to use such data as references for clinical samples from patients.

PUBLISHED CONTENT AND CONTRIBUTIONS

Markarian, Nicholas, Barbara E Engelhardt, et al. (Nov. 2025). "k-spaces: Mixtures of Gaussian latent variable models". en. In: *bioArxiv*. DOI: https://doi.org/10.1101/2025.11.24.690254. URL: https://www.biorxiv.org/content/10.1101/2025.11.24.690254v1.

NM concieved the project with PS and LP, generated the data, analyzed the results, and wrote the manuscript with the supervision of BE, NP, PS and LP. BE and LP concieved the orginal algorithm, which was developed by NM and LP.

Markarian, Nicholas, Kimberly M. Van Auken, Dustin Ebert, and Paul W. Sternberg (Mar. 2024). "Enrichment on steps, not genes, improves inference of differentially expressed pathways". en. In: *PLOS Computational Biology* 20.3. Ed. by Marc Robinson-Rechavi, e1011968. ISSN: 1553-7358. DOI: https://10.1371/journal.pcbi.1011968. URL: https://dx.plos.org/10.1371/journal.pcbi.1011968.

NM concieved the project with PS, generated the data, analyzed the results, and wrote the manuscript with the supervision of PS.

TABLE OF CONTENTS

Acknowledgements iii
Abstract v
Published Content and Contributions vi
Table of Contents vi
Chapter I: Introduction 1
1.1 The inner workings of a cell 1
1.2 The clinical perspective 3
1.3 Pathways and curation 3
1.4 Gene expression networks 4
1.5 The biological historical context 6
1.6 The computational context 14
1.7 Thesis structure and contributions 24
Chapter II: Enrichment on steps, not genes, improves inference of differentially expressed pathways 35
Chapter III: *k*-spaces: Mixtures of Gaussian latent variable models 67
Chapter IV: Low-dimensional subspaces reveal tissue-level diversity of gene covariance within pathways 144
Chapter V: Conclusion 192

Chapter 1

INTRODUCTION

This thesis is motivated by the need to understand gene expression covariation within known biological pathways but also develops a linear latent variable model with widespread applications in biological data analysis. It begins with revisiting enrichment analysis, a technique developed after the human genome project and before the advent of modern RNA sequencing for generating hypotheses about differential regulation of biological processes. This in turn inspired my desire to better characterize gene covariance so it can be studied as a potential signal of transcriptional regulation, the fourth and final chapter of this work. To do this, I needed to develop a novel statistical technique, the subject of Chapter Three. It took on a life of its own, and although this thesis is framed around the questions of pathways, it is an equally important contribution of this dissertation. As such, this introduction has two parts — the first introducing transcriptional regulation of cellular physiology and the historical context for my work on biological pathways and the second concerning the statistical learning foundations of least squares, k-means clustering, and principal components analysis, upon which our method, k-spaces clustering, is built.

1.1 The inner workings of a cell

The fundamental unit of life can differ depending on one's perspective, but most would agree that cells are the simplest functional living unit and that physiological processes are generally regulated at the cellular level whether or not the process itself is a so-called emergent property of a system. Many members of the eukaryotic kingdom of life are multicellular, and many of these multicellular organisms have specialized cell types organized into tissues and organs, in which collections of cells work together to perform some physiological function. Deoxyribonucleic acid (DNA) is the chemical material of genetic inheritance (Avery, MacLeod, and McCarty, 1944; Hershey and Chase, 1952; Watson and F. H. Crick, 1953), proteins are, in general, the actors that carry out cellular functions, and genes are the units of DNA that contain the instructions for producing proteins through ribonucleic acid (RNA) acting as a messenger to produce proteins from genes (Volkin, Astrachan, and Countryman, 1958; Hall and Spiegelman, 1961; Nirenberg and Matthaei, 1961;

Francis H. Crick et al., 1961).[1]

With very few exceptions such a multinucleated muscle cells, haploid[2] sperm and eggs, and the lack of nuclei in red blood cells, all cells in an organism contain the same DNA in the same quantity, yet have remarkably different physiology and appearance, which is due to differences in the abundance and chemical modification of specific proteins. The abundance of these proteins is controlled at the transcriptional level — the process of producing RNA from DNA, the translational level — the process of their production from an RNA template, and the post-translational level — the processes of protein degradation and activation or deactivation via chemical modification. These proteins chemically react with each other or with non protein molecules that they synthesize, degrade, or modify, and we generally subdivide the vast network of chemical reactions in a cell into pathways, or sequences of molecular interactions that are sufficient to perform some function. These pathways can be joined in series or parallel as part of a larger biological process required to sustain life, and we are typically interested in the differential regulation of these processes — much of research is focused on smaller functional subunits such as single genes but still within the broader context of their implications on physiology.

The activity of the pathways that make up these processes (though not all processes necessarily require a pathway — for example, the binding and release of oxygen from a red blood cell is performed by a single protein complex, hemoglobin) is accomplished by complex regulation. It spans transcription, translation, and protein activation, deactivation, and degradation along with control of flux and equilibrium by regulating the availability of inputs or the concentrations of outputs. Even transcriptional regulation itself is extremely complex, involving transcription factors that turn on one or more genes, RNA species that can increase or repress transcription, chemical modifications to DNA that can repress or activate transcription, and more (Lyon, 1961; The ENCODE Project Consortium, 2012; Quinodoz and Guttman, 2022; Guo and Guttman, 2022). Many aspects of regulation are likely still unknown, but it is clear from basic principles that proteins facilitating the steps in these pathways must be present for a pathway to be active and that the level of active

[1] This is often stated as the "central dogma" of biology, that information flow is from DNA to RNA to protein. However, the central dogma stated by Francis Crick and later clarified (F. Crick, 1970) was only that sequence information cannot flow from protein to RNA. Reverse transcription (Temin and Mizutani, 1970), or the production of DNA from RNA, is not an exception to the central dogma. Incidentally, the word "dogma" was chosen by Crick because of a misunderstanding that dogma meant a belief without evidence (Horace Freeland Judson, 1996).

[2] Haploid means only containing one copy versus two of each chromosome.

proteins for at least some subset of these steps controls the flux through those steps and by extension, the pathway. To accomplish their basic physiology, respond to stimuli from their environment, or to perform the functions that give them their cell type identity despite sharing the same DNA as nearly every other cell in the body, cells must regulate pathways. This regulation can be directly measured through biochemical assays for inputs, outputs, and intermediates, but for practical reasons they are usually indirectly inferred through parallelized 'omics assays, the most widely available being transcriptomics, or the measurement of RNA transcripts.

1.2 The clinical perspective

Clinically, physicians and their patients are interested in knowing how pathways are regulated in health and disease, not the expression levels of individual genes. To guide research and future treatment decisions, a typical first step is to generate a hypothesis about which biological process is involved. Even if the root cause is suspected with greater specificity than the pathway level (i.e., detected mutations in genes) the next question is how that DNA mutation leads to a difference in physiology. A typical line of inquiry may begin with RNA-sequencing of a patient sample, a technology that produces transcriptomic data, and proceeds with some hypothesis generation tool that aims to interpret the results in terms of prior biological knowledge — known pathways or processes. An example of this is enrichment analysis, which aims to statistically quantify the likelihood of overlaps between differentially expressed genes in the experiment with known, biologically relevant sets of genes in pathways or processes. These "results" are actually hypotheses; it is then up to the clinician or researcher to critically evaluate them in terms of their own knowledge and literature and decide whether to test those hypotheses. It would be enormously beneficial to have a model describing the physiological changes reflected from the gene expression changes, though this is still out of reach as the abundant data type is gene expression, not direct assays of pathway physiology.

1.3 Pathways and curation

Biological pathways can be thought of as functional subunits within physiological processes. They have historically been discovered through biochemical or genetic approaches or a combination of the two (e.g., (Han et al., 1993; Fire et al., 1998)). Their beginnings and endings are decided by consensus of the scientific community with general criteria that the sequence of pathway steps is sufficient to produce a meaningful output, and teams of expert curators have aggregated and organized this

knowledge in databases such as the KEGG, PANTHER, Reactome, and the Genome Alliance (Kanehisa and Goto, 1999; Thomas, Campbell, et al., 2003; Milacic et al., 2024; Thomas, Hill, et al., 2019). Not all sets of genes involved in a biological function are necessarily in a pathway together, however, and multiple scales of organization can be meaningful. For general biological information, we have the Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium et al., 2023), which hierarchically curates knowledge about biological processes, cellular components, and molecular functions and the genes in them. This work, beginning in 2000, predates much of the curated pathways and was the predominant resource for early enrichment analysis papers. Today, we have a new type of pathway representation, the Gene Ontology Causal Activity Model (GO-CAM), which organizes pathways as sequences of steps (GO molecular functions), and these steps are then annotated with the genes involved, their location (GO cellular component), and the higher order processes they are a part of (GO biological processes).

1.4 Gene expression networks

While pathways are functional units within the large network of biochemistry within cells, the modules within gene expression networks have are inferred effective units of transcriptional regulation. Pathways are grounded in a biochemical or molecular biology framework, while gene expression networks are information theoretic. Modules within gene expression networks, sometimes called gene expression programs (Kotliar et al., 2019), generally correspond to groups of genes that tend to increase or decrease in expression together. Early forms of this may have produced a simple factorization for the data (Alter, Brown, and Botstein, 2000) or obtained sets of genes that have higher Pearson correlation coefficients with each other than with genes in other modules (Langfelder and Horvath, 2008), but drop nonessential information such as how the modules correlate to each other, which genes are most correlated within modules, and the sign of these correlations as the principal goal is to simplify the *apparent* transcriptional regulation of the genome into units that we can try to interpret in terms of our biological knowledge. Today, gene network analysis has evolved, with types of networks that further aim to represent causal relationships between the genes (Jiang et al., 2023; Lu et al., 2021) and potentially predict the effects of perturbations on the network. In some cases, they could even be used to infer biological processes (Chapman et al., 2022; Kunes et al., 2024).

The validity of these outputs does not depend on molecular biology or biochemistry

or on units of physiological function, only their usefulness. In gene expression networks, not all members of a pathway will end up in the same module, most modules will incorporate not only several pathways but also several biological processes, and some modules may incorporate seemingly unrelated biological processes either because they are functionally related in ways we do not yet understand or merely as a consequence of the shared transcriptional machinery and genomic space that they occupy. Validity comes from an information perspective, but we hope that these simplifications of the transcriptome will be useful because their representations of the data will converge with known molecular biology and biochemistry, and studies typically aim to show just that by interpreting modules based on the overlaps with sets of genes in known biological pathways or processes.

Critically, in some clinical contexts, the *how* is not important, only the *what.* Abstracting the complexities of transcriptional, translational, and post-translational regulation and using the information theoretic perspective is sufficient for these. A statistical model that can generate good hypotheses about differential regulation of pathways is still useful for most translational research even if it is not a biophysical model built up from first principles with differential equations and kinetics. This thesis takes that perspective with the recognition that our basic knowledge of which genes function together for steps in pathways are curated and aggregated into cohesive resources, but regulation is frequently missing or non-quantitatively described in these models. Even when transcriptional regulation was once known, meaning that it was measured in an experiment and published in a paper but not curated, it is frequently not comparable to a patient. Studies in cell lines are not representative of human tissues' gene expression— cells' gene expression is inherently dysregulated by virtue of existing as a cell line— and regulation of the same pathway may differ between tissues. In our pathway models, we usually do not know which steps are transcriptionally regulated, which of subset those we should expect to see co-vary in our data, or what expression levels are typical or atypical. This would be valuable reference information either for computational tools or directly for physicians and scientists using the curated pathway model to make sense of a hypothesis generated from their data.

A note on non-transcriptional regulation

An obvious limitation of gene expression network approaches is that much of regulation occurs at the post-translational level, and it is not inconceivable that there are pathways that are *only* regulated at the post-translational level. It does seem likely

that some, if not many, pathways are primarily regulated at the post-translational level and that certain pathways may be primarily transcriptionally regulated only at a subset of steps.

Possibly due to the longstanding ability of biochemists to measure reaction kinetics as opposed to the newer ability to measure gene expression, there are many cases where we have a thorough understanding of how certain proteins are post-translationally regulated at certain steps in the pathway, particularly at rate-limiting steps, while transcriptional regulation is still murky. A great example of this is the glycolysis pathway, which breaks down glucose (a sugar) into pyruvate (a three-carbon molecule) to store energy in a chemically readily-available form through adenosine triphosphate (ATP) (Wu, Khan, and Lange, 2005; Düvel et al., 2010; Wiśniewski, Gizak, and Rakus, 2015; Choe et al., 2025). In terms of chemical kinetics, certain steps are rate-limiting, and we have a fair understanding of how they are inhibited or activated, but even recent work after decades of study has shown that the "unimportant" non-rate limiting steps without well-established post-translational regulation can theoretically play crucial roles (Choe et al., 2025). All this is to say that gene expression is worth studying, even when we think our regulatory understanding is complete, and even when we know that gene expression does not provide the complete picture.

1.5 The biological historical context

Microarrays and enrichment analysis

Some of the first methods for quantifying gene expression were built upon a technique called polymerase chain reaction (PCR) and were termed real time or quantitative PCR (qPCR) (Higuchi, Dollinger, et al., 1992; Higuchi, Fockler, et al., 1993). qPCR remains in use today for sensitively and accurately quantifying the relative abundances of mRNAs. However, they are not scalable and are used only for a few genes at a time. To understand why we have the computational methods we use today for analyzing gene expression and predicting pathway regulation, we need the historical perspective of what technological advancements were providing the data to answer certain questions for the first time. I would argue that this drives which topics are "hot" in computational biology and when they receive the attention, and it seems reasonable that if the computational techniques still work on data generated by the new technology, there is less reason for key players in the field to return or new students to enter if other technologies are driving new hot computational

problems. If we had an abundance of detailed, curated pathway models that included regulation and accurate, reproducible mRNA quantification techniques at the time, I would argue that the methods we use today would be quite different.

Microarrays are chips with DNA probes that can bind to their complementary mRNA targets to facilitate quantification. They were invented in 1995 (Schena et al., 1995) and over the following decade made quantitative measurements affordable and scalable. Their primary limitation was their restriction to only known parts of the genome — newly discovered or unsequenced genes could not be characterized as they would not have probes on the array — but they also suffered from challenges of background signal, low dynamic range, and difficulty in comparison across experiments, which required heuristic normalization methods (Bolstad et al., 2003; Wang, Gerstein, and Snyder, 2009). The limitation to only quantifying the genes with probes already on the array was a primary driver of the later development of RNA sequencing,[3] but by the early to mid 2000s, microarrays provided an efficient way to quantify relative mRNA abundance for large numbers of genes in samples from patient tumor tissue (West et al., 2001) to yeast cultures (Alter, Brown, and Botstein, 2000) and quantitatively compare genes (that were not too highly or too lowly expressed (Wang, Gerstein, and Snyder, 2009; Marioni et al., 2008)) between conditions within the same experiment.

As scalable assays for gene expression became available, there was an explosion of computational methods for analyzing the data, as it was intractable to manually interpret gene expression changes one by one for lists of hundreds of genes. Such interpretation would require unreasonable amounts of time, expert knowledge of many different fields of research, and vast demands on human memory capacity to synthesize an interpretation from such a large list. Critically, the genes were not chosen by the experimenter due to prior knowledge. Thus, enrichment analysis was introduced and made use of curated expert knowledge in the form of gene lists for biological processes, pathways, and other GO terms by quantifying their overrepresentation in the data (Huang, Brad T. Sherman, and Lempicki, 2009).[4]

[3] Personal communication from Barbara Wold, who kindly discussed the evolution of transcriptomics with me.

[4] Crucially, these methods (and often experiments that generate this type of data) are not hypothesis-driven, at least not in a simultaneously meaningful and precise sense at a physiological level. Obtaining massively parallel measurements of thousands of genes usually reflects a lack of a specific hypothesis, or rather a generic hypothesis that there must be some transcriptomic difference between experimental conditions. That is not to stereotype all such experiments, as such an assay would be a convenient way to assay a few tens of genes that are of interest. Even a mildly more specific hypothesis, such as "metabolism is probably different between cancer and non cancer cells"

Subsequent modifications leading to different generations of methods included more sophisticated statistical tests to avoid predefined differential expression cutoffs by analyzing ranked the lists of genes based on expression changes (Subramanian et al., 2005), and methods for analyzing the topology, or network structure, of the pathways (Glaab et al., 2012; Brad T Sherman et al., 2022). There are many variations on these; for a pair of good, chronologically spaced reviews, see (Huang, Brad T. Sherman, and Lempicki, 2009; Reimand et al., 2019). There were many papers in this field in subsequent years (68 tools as of a 2008 review (Huang, Brad T. Sherman, and Lempicki, 2009) and "over 100" as of 2023 (Zhao and Rhee, 2023), but the techniques that appear to be most widely used today and were recommended in (Reimand et al., 2019) in 2019[5] implement the simple early testing schemes of overrepresentation and ranked change analyses. It appears that research activity slowed down by the mid 2010s, which I speculate could be related to computational biologists' increasing interest in solving the array of novel statistical problems in the newly developed field of single-cell RNA-sequencing or even bulk RNA-sequencing. A common theme among enrichment tools is that their core method for biological hypothesis generation is to obtain a p-value — the chance of observing a feature of the data to that degree or more extreme — for the statistical null hypothesis that expression changes in all genes are independent, so more differentially regulated genes in a pathway than would be expected would indicate a regulation of the pathway's activity. At least the methods that survived and are commonly used today do not use external references for typical gene expression changes, and given the difficulty of comparing different microarray experiments and the low dynamic range of measurements (Wang, Gerstein, and Snyder, 2009; Marioni et al., 2008) and the lack of mention in reviews (Huang, Brad T. Sherman, and Lempicki, 2009; Reimand et al., 2019), it is not unreasonable to speculate that few if any such methods were made. I am sure they were considered, but the experimental technology and the curation were not ready when this field developed.

---

would have been a terrible experimental design prior to massively parallel analyses when obtaining each data point involved extensive labor. This is in contrast to a targeted hypothesis, that a specific physiological process is differentially active between conditions, in which case the appropriate experiment is a targeted biochemical assay or phenotypic assay of that process, potentially paired with quantitation of gene expression for the specific genes involved. This represents a meaningful shift in the way some research is conducted, as the meaningful hypotheses are generated *after* the experiment is analyzed. Thus, the "analysis" techniques could really be viewed as automated hypothesis generation techniques.

[5]Their recommendation does not necessarily reflect which methods are "most" statistically correct but rather which tools were reasonably correct, useful and easy to use, and have survived either standalone or been incorporated into widely used tool suites to have their knowledge databases maintained and updated.

RNA sequencing

RNA-sequencing (RNA-seq) was invented in 2008, with the first papers from Ali Mortazavi in Barbara Wold's laboratory at Caltech quantifying a mouse transcriptome (Mortazavi et al., 2008), the Grimmond laboratory at the University of Queensland studying stem cells (Cloonan et al., 2008), the Bähler laboratory at the Wellcome Sanger Institute studying yeast (Wilhelm et al., 2008), and the Snyder laboratory at Yale studying yeast as well (Lister et al., 2008). Generally, mRNA molecules are converted to complementary DNA molecules, which then get sequenced and mapped to a reference genome. This technology was not limited by prior knowledge of gene sequences as it did not use hybridizing probes, and over the years it became cheaper and widely used. Provided that long enough stretches (on the order of a few tens of bases) could be accurately sequenced, reads could be unambiguously be mapped to the genome for most genes. The key challenge was to normalize the data properly, the most critical challenges being accounting for the length of transcripts (longer mRNAs provided more sequencing reads) and variation in the amount of RNA present and the depth to which sequencing occurred, which could be corrected in a principled way with statistical models (Roberts et al., 2011). Though comparing RNA sequencing results produced with different technologies remains nontrivial, the TMM (Robinson and Oshlack, 2010) and DESeq (Anders and Huber, 2010) normalization methods published in 2010 enable cross-sample comparison, even with different tissues, and are widely used today. RNA-seq also provided the starting data for isoform quantification, leading to computational efforts to enable this as well (Trapnell et al., 2010). In summary, RNA-seq both improved the quantification and reproducibility limitations of microarrays and simultaneously provided a new set of well-defined problems[6] for computational biologists to work on.

Single-cell RNA-sequencing followed bulk RNA-sequencing, and although the first paper was published in 2009, only one year after, it was not widely available until later in the 2010s. Single-cell sequencing improved on existing technology by allowing for the relative quantification of mRNA abundances in single cells. However, the small amount of starting material for the sequencing process and the inherent biological stochasticity of gene expression (Gorin et al., 2022) at this level meant that all-new statistical methods needed to be developed to handle such data. Even today, there are still questions about how to best normalize, analyze, and integrate single-cell RNA-sequencing. Recent transcriptomic advancements

[6]This is in contrast to enrichment analysis, which is well-defined statistically but without a reliable biological ground truth.

technologically have been largely focused on obtaining spatial measurements in the data, either with variations in the RNA-seq technology (Lim et al., 2025) or with hybridization and imaging techniques (Lubeck et al., 2014; Trivedi et al., 2018). All of these again require development of principled computational normalization and analysis techniques.

Data from these bulk and single cell experiments have been accumulating in publicly available repositories such as the Sequence Read Archive (SRA), Gene Expression Omnibus (GEO) or other projects that build upon this in some form (e.g., (Lachmann et al., 2018; CZI Single-Cell Biology Program et al., 2023)), all the while curated knowledge bases have been growing (Ashburner et al., 2000; Thomas, Campbell, et al., 2003; Good et al., 2021; Milacic et al., 2024) and newer, better representations have been evolving (Thomas, Hill, et al., 2019) that could enable pathway analyses likely considered but never done in the microarray era. With the efforts that have been made to uniformly preprocess this data to eliminate computational batch effects (Lachmann et al., 2018), vast amounts of data produced with each sequencing method, and the possibility of modeling batch effects with machine learning to integrate this data at scale, millions of dollars' worth and a decade of data could suddenly become available to study pathway regulation across the swath of conditions and tissue or cell types in human health and disease. This dissertation was driven by that opportunity, and Chapter Four focuses on application of the method I develop in Chapter Three to start laying the foundation for such analyses. Gene expression networks differ across tissues and likely across cell types as well (Piro et al., 2011; Pierson et al., 2015; Saha et al., 2017; Gewirtz, 2022) but current efforts revolve around trying to discover new modules or construct new network connections, meanwhile our pathway models are without references for which genes in them covary (presumably because of regulation) and what typical expression levels and changes look like for each tissue. Once these typical changes and how expression is similar between certain tissues and different between others are well understood, we can begin correlating them with experimentally measured variables (either in new experiments or in the metadata of integrated publicly available experiments) to act as references in the prediction of transcriptional regulation of pathways.

I personally think this approach will help us disentangle ourselves from the p-value, which is pulling a double-duty as a likelihood of a gene expression event and the confidence metric for a pathway regulatory state. I argue that *most* of the time, we are not primarily concerned with the probability of observing a gene expression

change at least as extreme as the one in the experiment under a contrived null hypothesis of independent, uniformly selected differential expression of genes in the genome. Such a null hypothesis makes sense in the context of studying mutations and random drift in population genetics, a process that should generally be observed uniformly at random in the absence of selection. There, the null hypothesis actually describes a real process. If adjustments were made to account for any subtle inherent differences in mutation rates at different loci due to biochemical properties, we would obtain well-calibrated p-values that make sense. In the case of enrichment analysis, nobody *believes* the null hypothesis, and I would argue that we primarily do not *care* about having well well-calibrated p-values. If all genes in a hypothetical stress-response pathway were tightly controlled at the transcriptional level (and only at that level, so RNA-sequencing provides a perfect proxy for pathway activity) by a single transcription factor (so we violate the independence assumption as much as possible and the likelihood of observing a gene expression state is solely the likelihood of observing that level of the transcription factor) and we further suppose that this pathway is frequently activated by experimental conditions (being an organism[7] in a lab is a stressful existence), perhaps the well-calibrated p-value would be 0.5, as it is upregulated in most experimental conditions and inactive in most controls. But, we want to know whether it is active or not, and being active half of the time (as indicated by the p-value) does not make it any less active in any given experiment.

Separately, we may not care to follow up on this in downstream experiments because the activation is not unique, and relatively recent work has shown that many differentially expressed genes in studies are so-called "frequent-fliers" (Crow et al., 2019), but my key point here is that the question of whether to follow up with another experiment depends on the combination of uniqueness, relevance, and correctness of the enrichment result, and it is too much to ask a single metric to satisfy all three. After all, if we know only one gene, the rate-limiting one, in a relevant pathway is transcriptionally regulated and we see it is indeed upregulated in the experiment, we would want to follow up on this regardless of the p-value. It is better to separate the activity inference metric from the decision making metric. One path towards building such tools is to first studying the gene expression space for human tissue samples in the context of known pathways.

[7]Broadly defined as a living being— either the graduate student or the subject of their work.

The GTEx project

The Genotype and Tissue Expression (GTEx) project was specially funded by the U.S. National Human Genome Research Institute and began in 2010 with the goal of producing a database of human tissue expression variation using bulk RNA-sequencing and corresponding genotypes to understand how vairation in the genome affects variation in the transcriptome (Gibson, 2015; The GTEx Consortium et al., 2015; GTEx Consortium, 2017; T. G. Consortium, 2020), now totaling 19,788 samples from 946 donors across 54 tissues.[8] The v8 dataset analyzed in Chapter Four totaled 17,382 samples from 838 donors and 52 tissues. This work was conducted by a consortium of laboratories around the country and represented a joint effort to create a resource that would be accessible to the entire research community. Critically, sample collection and sequencing were standardized along with careful metadata collection to facilitate integrated analyses.

In bulk RNA-sequencing, cell type composition could be measured from tissue samples separately but is often computationally estimated using reference gene expression profiles for cell types (Aran, Hu, and Butte, 2017; Donovan et al., 2020; Kim-Hellmuth et al., 2020). These methods are benchmarked or validated in some tissues, and some studies suggest that different tools may be optimal for specific tissues (Sutton et al., 2022). In general, it is unclear how accurate these tools are and how well they generalize,[9] which is a major motivation for single-cell studies when cell type-specific hypotheses are investigated. Nonetheless, bulk RNA-sequencing holds an edge over single-cell in mRNA quantification accuracy, making the GTEx data an invaluable resource, and the GTEx consortium provides cell type enrichment estimates using a tool called Xcell (Kim-Hellmuth et al., 2020; Aran, Hu, and Butte, 2017).

Current efforts: single-cell, parallelized perturbations, network discovery

In contrast, single-cell sequencing analyses allow for much greater sample sizes, as tissues can be dissociated into single cells and sequenced to produce atlases on the order of one million data points (e.g., (Hickey et al., 2023; Lake et al., 2023; HuBMAP Consortium et al., 2019; Arutyunyan et al., 2023)), with the advantage of yielding cell type-specific information provided cell types can be accurately identified. In these studies, data is typically clustered by gene expression and then all cells in the cluster are assigned a cell type label based on reference gene

[8] According to the GTEx portal website on November 9th, 2025.

[9] Validation "in-silico" is only as good as the assumptions that went into the simulation.

expression (T. T. M. Consortium, 2018), or the reference gene expression is used either during the clustering process or to directly assign cell type labels to cells (Zhang et al., 2019; Booeshaghi, Galvez-Merchán, and Pachter, 2024). Data is then typically split by cell type to construct cell type profiles or other analyses.

Other new possibilities with single-cell sequencing are scalable assays of gene expression across a very large number number of experimental conditions using cell lines, or cultures of individual cells that survive outside the human body, like we would culture bacteria or yeast. One such example is Perturb-Seq, where CRISPR (a genome editing technology that can modify or inactivate genes) is used to perturb the genome and single-cell sequencing is used to assay the transcriptome with the goal of discerning the underlying gene network architecture (Dixit et al., 2016). Networks can be constructed with deep learning (in which case the network itself is unknown but transcriptional responses to perturbations can be predicted) (e.g., (Cui et al., 2024)), matrix factorization approaches (e.g., (Kunes et al., 2024)), graphical models (e.g., (Friedman, 2004; Karlebach and Shamir, 2008; Kotiang and Eslami, 2020)) or physics-inspired methods (e.g., (Jiang et al., 2023)).

A key challenge (aside from the serious issue of batch effects [10]) is the interpretability of the discovered modules or "programs" within the networks. Often, enrichment analysis itself is used to interpret the modules (Kunes et al., 2024). This challenge has been recognized and addressed by using gene sets in the network construction (Buettner et al., 2017; Elyanow et al., 2020; Kunes et al., 2024), though methods are imperfect, but a recent approach (Kunes et al., 2024) aims to account for cell type specifity or sharing of biological processes while incorporating prior knowledge in the network construction process. Cell types or states are specified before the network construction process, however, and the goal of this method is still to produce new network discovery for the entire transcriptome.

While these advances are important, the efforts to produce better gene networks for the whole genome have left behind our knowledge already codified into pathways. The second issue is that by virtue of being cell lines, the transcriptome is inherently dysregulated[11] or at the very least surviving in an unnatural environment even if genomic changes are minimized. A cancerously multiplying leukemia cell in a petri dish probably will not serve as a good reference for a muscle cell living in the human

[10] Unwanted variation introduced by technical experimental factors.

[11] Arbitrarily chosen human cells cannot survive and multiply in laboratory conditions, and cell lines frequently are cancer-derived and have rearranged and duplicated genomes or were forcibly differentiated in laboratories from stem cells outside their natural host tissue.

patient's heart. The gene expression networks would likely not be the same, and the "normal" gene expression levels for each would certainly be different. Thus, despite the unattractive complications of cell type composition variation and lower throughput with bulk RNA sequencing of human tissues,[12] I believe that a reference for human tissues is still necessary and clinically relevant. In Chapter Four, we take a complementary approach of characterizing gene expression covariation within the pathways specified in our pathway models, forcing interpretability by restricting analysis just to the genes in the pathway and providing utility by characterizing expression for human tissues in a way that can be added to the databases containing our pathway models.

1.6 The computational context

Chapter Three of this thesis introduces *k*-spaces, a probabilistic mixture model, that we believe will have wide impact beyond the pathway problem introduced above. The other applications we pursued in that study are introduced in that chapter. I give it its own introduction here and discuss the basic mathematical foundations.

Fitting models to data

Loosely put, a model is a representation of our understanding of data. A subset of models, probabilistic models, can be mechanistic and represent our understanding of how data points are created or simply represent our understanding of what values they tend to have. Both of these can be called generative models if they provide a procedure for making data, and both can be useful at various stages in the scientific process. A non-mechanistic model provides less information about how the data came to be but can reveal a signal in the data that can be studied to produce mechanistic understanding, while a mechanistic model can represent our current understanding and be matched up to data to see how good our representation of the system really is.

For example, consider a sensor outside that randomly measures whether it is light or dark outside twice-a-day, at 12:00 AM and 12:00 PM. A simple, non-mechanistic yet fully probabilistic or generative model fit to the binary light measurements alone could be a coin flip, or Bernoulli trial with a probability of 0.5 for light and 0.5 for

[12] Single-cell sequencing of human tissues would also be suitable, but we are further behind with accounting for batch effects in single-cell than bulk RNA-sequencing and scalability is severely limited for now due to the difficulty and cost of obtaining high quality single-cell RNA from patients given that the samples are usually removed during surgeries when pristine data collection is not the priority.

dark. This would agree well with the data and suggest that there is some physical process causing the sensor to detect light half of the time. This model could be updated to include time of day for the measurement, where the probability of light or dark is affected by the time. This is a bit more mechanistic, and we could make it unequivocally mechanistic by introducing the idea of Earth's rotation producing the day and night cycles. We can go further as well, by using the laws of gravity to explain the motion of the Sun and the Earth through space and the rotation of Earth as a conservation of angular momentum from the formation of the solar system, but adding these equations does not really help in the sense that we cannot directly identify any of these variables from our binary, twice a day light sensor system. All Earth models that produce light at 12:00 PM and darkness at 12:00 AM are equally good fits to this data.

Separately, we can also test a mechanistic theory with data to see if something is wrong. If we have a mechanistic model with the Earth's rotation producing day and night and this producing light and dark measurements alone, then we would say that the probability of light in the day is 1 and dark in the night is 1. But, every once in a while we will get a dark 12:00 PM measurement because of a solar eclipse — the moon is not in our equations. This tells us that the mechanistic model is either wrong or incomplete. We can change the probability for light measurement in the day to 0.9999 or some other value very close to 1 while searching for the missing component of our physical model, the moon. Once we fill this in, we could change the probability back to 1 during the day unless there is a forecasted total eclipse given Newton's laws of motion, in which case it is 0. Or we could be content to have learned the proper physics as a result of the mismatch and still keep the simplistic model with probabilities of sensing light at 0.9999 for day and 0 for night because the model still works very well on the data.

Biological models are in general not quantum and therefore not strictly random, or stochastic, but the systems are so complex that we blend mechanism and data-driven statistical distributions to represent our understanding. Our learning goes both directions from simplistic models that merely describe the data's location in feature space to detailed processes accounting for physics, chemistry, and molecular biology. The two approaches work together, with simplistic, even non-mechanistic, models serving as placeholders for yet unknown mechanistic processes (as is the approach in my work) until those processes are well-understood.

Principal Components Analysis

What is PCA?

Principal components analysis (PCA) as most biologists know it is a linear dimension reduction technique that retains the maximum variance in the data sample. Dimension reduction means to take data where individual data points, or observations, are described by D measurements, or features, and represent it with fewer than D dimensions. This can be done for computational reasons, where data may be reduced from very high dimension to moderate-to-high dimension or to 2-3 dimensions for practical reasons of visualization. Variance is one measure of spread in data points, the mean squared distance of data points from their mean value:

$$\mathrm{Var}(X) = \sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{1.1}$$

So, to most, PCA is the plot where data points' coordinates are determined by a linear equation $f : \mathrm{R}^D \rightarrow \mathrm{R}^d$ that takes in their high dimensional coordinates as input and outputs 2 or 3 dimensional coordinates such that the resulting plot spreads out the data as much as possible. Few think about how this equation is obtained beyond "the directions of maximum variance" or "the singular value decomposition (SVD) of the covariance matrix," which are *also* known as PCA. The geometric intuition for PCA is helpful for showing why PCA has so many interpretations and explanations. To use a programming term, the acronym PCA is overloaded.

PCA as a plot:

A PCA *plot*, implying that the A stands for analysis of the reduced-dimension data points, is the projection of data points onto an affine subspace that maximizes the variance of the projections. A projection of a data point onto a flat surface is the location where a line perpendicular to the surface intersects both the data point and the surface. If we shine a light directly facing the surface, it would be where a shadow cast by a data point *off* of the surface lands *on* the surface. Technically, an affine subspace is a region of a vector space satisfying certain criteria but lacking an origin, but as we are only discussing Euclidean spaces, we can say it is the generalization of lines, planes, prisms, etc. that do not pass through the origin. The dimension of each subspace, which I denote with d, is the number of perpendicular directions that one can move within the subspace. So, a line is 1-dimensional, as given a point on the line, one can only move forward or backward while staying on the line, a plane is 2-dimensional, and a cube is 3-dimensional. By this logic, a point

is *also* an affine subspace — a 0-dimensional one as there is no degree of freedom for a point to move while staying in the subspace. The dimension d of the subspace is independent of the dimension D of the ambient space it is contained in (except for the upper bound $D \geq d$): a line can exist on a 2-D plane or in the 3-D world and is still a 1-dimensional object either way. This is just what a 2-D PCA *plot* is, the projection of points onto an a plane oriented in the high dimensional space such that the projections of the data points onto the plane are spread out as much as possible.

PCA as an affine subspace:

The PCA *plane* is the one closest to the data or any plane parallel to it. It is the solution to this problem:

$$\arg\min_{\mathbf{S}} \sum_{j=1}^{N} \left\| \mathbf{x}_j - \text{proj}_s(\mathbf{x}_j) \right\|^2, \tag{1.2}$$

where $\mathbf{S}$ is the plane or subspace and $\text{proj}_s(\mathbf{x}_j)$ the projection of point $\mathbf{x}_j$ onto $\mathbf{S}$. From this perspective, PCA is either the process of finding the subspace or the subspace itself. Unlike the original feature space, the plane does not have preset axes, but a natural coordinate system uses the line that has the most variance in the projections of data points onto it as its first axis, principal component (PC) 1. To make this a plot axis, we can set the 0 on its number line to be the mean of the projections. The similarity of this axis construction process to the construction of PCA plane itself is important, and in fact, whether we take the projections onto the plane and then again project them onto the line or directly find the line that maximizes the variance of the projections our data from the high dimension, we get the same first axis for our coordinate system. Different dimensional PCAs are *nested.* The top two PCs are contained within the space spanned by the top three PCs, and if we define them to be the directions of maximum variance (i.e., constructing the plane or the cube from the principal components rather than the defining the principal components as basis vectors after finding the cube or plane) then PC 1 and PC 2 of a 2-D PCA plot should be exactly PC 1 and PC 2 of a 3-D PCA plot. If we remove PC 1 from the data by subtraction of all data points' projections onto PC 1, we can repeat this to obtain PC 2 and so on.

Viewing PCA from the perspective of the subspace first with subsequent definition of the axes as a coordinate system for the subspace also clarifies another relationship: maximizing the retained variance and minimizing the $L2$ norm. The $L2$ norm is

the sum of squares for all the entries in a vector and can be extended to a matrix and called the Frobenius norm, where it is just the sum of the $L2$ norms of all the constituent vectors. The earlier definition of the variance as the sum of squared residuals is just the $L2$ norm of the vector after subtracting its mean. In fact, of all the values that can be subtracted from the vector, including 0, the mean is the one that minimizes the $L2$ norm, which means it is the 0-D subspace that minimizes the sum of squared distances to the data points (here, just points on a number line).

If we measure closeness as the sum of squared residuals, or the $L2$ norm of the residuals, then the mean is the closest. We cannot retain any variance with a point, as it is a 0-dimensional subspace, but if we instead consider the process of trying to find the PC 1 line, and observe how the squared distances from points to their projections on the line and the squared distances from projections to their mean along the line change, we see that the projections are spaced out as much as possible along the line when the line is closest to the data points in the ambient space. We are either solving

$$\arg\max_{\mathbf{s}} Var(\text{proj}_s(X_1) - \text{proj}_s(\bar{X_1})) \; + \; ... \; + \; Var(\text{proj}_s(X_d) - \text{proj}_s(\bar{X_d})) \quad (1.3)$$

or

$$\arg\min_{\mathbf{s}} Var(X_{d+1} - \text{proj}_s(X_{d+1})) \; + \; ... \; + \; Var(X_D - \text{proj}_s(X_D)), \quad (1.4)$$

the latter of which is just eq 1.2 with the sum of squares replaced with the variance operator.[13] This is because of the decomposition of the total variance into variance within the subspace (along the line) and outside of the subspace (from the points to the line), and it shows how the common definitions "maximizes the retained variance of projections" and "minimizes the discarded variance of the data" are *almost* equivalent. The first describes many lines as candidates for a 1-D PCA—any translation of PC 1 will not change the projections' positions relative to each other and thus will not change the retained variance. The second only describes one line, the closest one, as translating the line will add distance from data points to their projections. If we view PCA as solely a dimension reduction, this is unimportant,

[13] Here, we are considering a coordinate system using the affine subspace's basis vectors for $1 \ldots d$ and any arbitrary coordinate system for $d+1 \ldots D$, so we have rotated the first d axes for the data into whichever subspace we are considering. The axes for $1 \ldots d$ do not have to be *the* principal components, but it is a natural choice. It is all added up in the end, and expansion of the terms leads us to a simplification with squared Euclidean distances between the points' projections and their mean for eq 1.3 or the squared Euclidean distances between the points and their projections in eq 1.4. In these new equations, the specific coordinate system is irrelevant.

but it matters if consider it a representation of the full data as a combination of the coordinates in the dimension reduction subspace and the position of the subspace within the ambient space.

PCA as an autoencoder or latent variable model:

From the geometric view, there are at least two equally reasonable directions to go, the parametric probabilistic interpretation and the machine learning autoencoder interpretation.[14] In the autoencoder interpretation of PCA, the PCA plot shows a view inside the latent space and is embedded within the ambient feature space. If the unobserved, true points of the data all lie on a d-dimensional linear manifold[15] but the observed data points are subject to measurement error, then the data projected onto the PCA latent space is a de-noised but still high dimensional version of the data, and the projections in the relative coordinate system defined on the subspace are a low dimensional representation of the data. Autoencoders are machine learning models that encode data in a low dimensional space and decode data to return it to the high dimensional space, which this combination enables.

Some information is lost in the process,[16] but if the dimension of the latent space is truly close to the dimension of the underlying manifold, then latent space paired with the projections serve as a representation of important variation in the data. PCA does exactly this. Centering, or subtracting the mean from the data, is often described as a preprocessing step *before* PCA, but from this perspective, calculating the mean is part of the process of determining the decoder, and after the mean is subtracted, the eigenvalue-eigenvector decomposition of the covariance matrix of the centered data gives the rest of the encoder and decoder schemes, where the encoder removes the noise in the data and the decoder maps the low dimensional representation back to the high dimensional space. From the autoencoder perspective, PCA is also a latent variable model. The PCA plot can be viewed as a latent space just as autoencoders can be used to learn useful latent representations of data (Goodfellow, Bengio, and Courville, 2016). In this subspace, the value of each coordinate can be treated as a latent variable.

PCA as a probabilistic model:

[14] Both are probabilistic, but parametric means an explicit probability distribution while most machine learning is not explicitly so but can be interpreted as implicitly representing probability.

[15] A manifold is a space that is locally Euclidean, or linear, and a linear manifold is a space that is globally euclidean — the affine subspaces described earlier.

[16] If the autoencoder is set up to perform data compression by storing data in fewer dimensions and uses a simple transformation to return the data to its original format, then there might be very little or no information loss.

We can further connect the above to a probabilistic interpretation of PCA. From the perspective of minimizing the $L2$ norm, the PCA subspace is a generalization of linear regression to orthogonal regression, where we not only minimize errors in the y variable as a function of x but also assume errors in both x and y — specifically when we assume errors are unbiased in magnitude between the two, yielding total least squares. PCA is often cited as being invented by Karl Pearson in 1901 (Pearson, 1901) and independently discovered many times and given different names depending on the field— the Wikipedia entry for "Principal Components Analysis" as of November 7th, 2025 lists: Hotelling transform after Harold Hotelling in 1933 (Harold Hotelling, 1933), the Karhunen–Loève transform,[17] quasiharmonic modes, spectral decomposition, proper orthogonal decomposition, and empirical modal analysis.[18] If we take the perspective of orthogonal regression, then the concept of PCA at least dates back to 1878 with J Robert Adcock proposing generalizing linear regression to finding the most likely line that generated data with errors in both x and y and equal variance (J. Robert Adcock, 1878). This orthogonal regression also allows the interpretation of a latent variable model, as we are in effect trying to correlate the x, y, z, and any further dimensions together and represent them with just one variable in PC 1.[19] Adcock implied a latent process combined with a noisy observation process and wanted to find that latent line along which the process occurred.

Whether he was aware or not, just as the process of minimizing squared residuals in y in linear regression has an interpretation as fitting a Gaussian, or normal, distribution to noise in the y variable after it is predicted as a function of x, this has an interpretation as fitting a line to minimize multivariate isotropic Gaussian noise. This noise is equal in magnitude all directions and is Gaussian in each dimension. Technically, this is distinct from the assumption that data points are perturbed orthogonally from latent points to observed ones, but the resulting line is the same. Because the noise distribution is rotationally symmetric, we can freely choose any coordinate system for the components of the noise, and if we choose our d principal components to be the first d of the D noise components, we get the within-subspace component of the noise for free as retained variance, since

[17] Of the alternate names, this is the only one I have personally encountered in a paper.

[18] It also lists the Eckart-Young Theorem (Eckart and Young, 1939) but this is the theorem that PCA is the optimal linear approximation, along with SVD and eigenvalue decomposition of the covariance matrix, which just describe a process that one can use to obtain PCA.

[19] Multiple linear regression is a distinct concept, where both x and y are used to predict z, so errors are still only considered to be in the response variable.

Adcock's and the regression perspectives on PCA only care about minimizing out-of-subspace noise. As stated by Collins et al., PCA has a probabilistic interpretation as the subspace that produces the maximum likelihood of obtaining data points perturbed by isotropic gaussian noise (Collins, Dasgupta, and Schapire, 2002). Any component of the perturbation that occurs within the subspace is ignored, as assuming the projected data point is the true location of the latent point just shifts the latent point within this model, so we obtain this equation:

$$\arg\max_{\mathbf{S}} \Pi_{j=1}^{N} f(\mathbf{x}_j - \text{proj}_s(\mathbf{x}_j); s), \tag{1.5}$$
$$f(\mathbf{x}) = N_{(D-d)}(0; \sigma^2_{\text{noise}} \cdot \mathbf{I}_{D-d}).$$

A reader familiar with probabilistic PCA may have expected a discussion of probabilistic PCA (pPCA) first (Roweis, 1998; Tipping and Bishop, 1999b), and it does critically differ from the above model as being truly fully probabilistic, but it is important to recognize that the principal components and the affine subspace of PCA can be obtained from a regression setting where orthogonal residuals are assumed. We can add to eq 1.5 to still have a fully generative probabilistic model, which the approach I take in Chapter Three.[20] We can assume a Gaussian distribution on each dimension of the subspace, which are orthogonal to the out-of-subspace dimensions with noise, and the reason this works is that PCA, as described earlier, can be obtained iteratively by finding PC 1 based on maximizing variance (in other words, fitting a 1-dimensional Gaussian for that axis) and then PC 2 in the same manner and so on, and when we stop at PC d, we can fit a single isotropic Gaussian to the remaining dimensions.

From here, we can also see that another interpretation is to fit a multivariate Gaussian to the data and drop the minor $D - d$ axes of the resulting ellipsoid and only keep the top d axes. As an aside, this definition should be extended to "principal components for PCA are the result of fitting any Gaussian with any arbitrary variances along its eigenvectors, provided they are in sorted order;" a proof of this is in the Chapter Three Appendix. This is where the eigenvalue decomposition of the covariance matrix comes in. Everything is tied together because we use a common mathematical foundation in research. Perhaps Gauss used his namesake distribution because it was analytically tractable given the tie to the $L2$ norm, or maybe he wanted to work out the probabilistic interpretation of minimizing the $L2$ norm or variance. I am not

[20] Not on purpose but rather because I derived it without knowledge of pPCA and from a k-means framework. The two are equally suitable for my applications.

familiar with the history. The idea of orthogonal perturbations are a bit awkward, but the variance of the estimated noise distribution can be subtracted from the variance along each subspace dimension if we want the maximum likelihood estimate for the variance of the latent variables represented by our subspace if they had the same noise as well.

This takes us to pPCA, another generative model which shares the components of PCA and the subspace of PCA, but it differs slightly on the $plot$ part of PCA because it does not assume orthogonal residuals and accounts for uncertainty in the latent points' locations unlike PCA. It has its own section.

PCA as the technique we use today has been rediscovered more than once across different fields because it is the natural solution to several frequently encountered optimization or analysis problems as a result of its relationships with the mean of the data, the covariance (a description of how the variations of different features are interrelated), and the analytical convenience of the $L2$ norm.

What *is* PCA?

The term is overloaded and the *concept* is too. To quote Henri Poincaré, or Lior Pachter, who relayed the English translation to me, "Mathematics is the art of giving the same name to different things."

Probabilistic PCA

pPCA is a particular generative model for data that was derived independently by Sam Roweis (Roweis, 1998) at Caltech and Michael Tipping and Christopher Bishop at Microsoft Research (Tipping and Bishop, 1999b) in 1998 and 1999, and these papers contributed a probabilistic model to the definition of PCA. Tipping and Bishop derive pPCA from factor analysis, a type of latent variable model, by specifying that the noise in the model must be isotropic. If we take the closest subspace perspective of Adcock in 1878 (J. Robert Adcock, 1878), Pearson in 1901 (Pearson, 1901), and others on what PCA is, then their contribution is a probabilistic interpretation for the location of the subspace. Adcock's paper begins with "Find the most probable position of the straight line determined by the measured coordinates, each measure being equally good..." and views the location of the subspace as the thing to determine as a function of observed data points (J. Robert Adcock, 1878). However, in the PCA *plot* as PCA is frequently used in biology, the interest is in the location of data points. From the perspective of a latent variable model, which is

quite similar to Adcock's idea that the line is in a fixed location but measurement error corrupts the *observed* data, there are observed data points and latent data points. This distinction has no impact on the location of the subspace, but given the awkwardness of orthogonal noise (isotropic noise only in the complementary space) pPCA contributes an idea of uncertainty in the position of the latent points due to isotropic noise in all measurements. So, a Gaussian distribution within the latent space first produces the latent point prior to an isotropic Gaussian that perturbs it *into* the complementary space and *along* the latent space. Instead of a fixed *inferred* position for the latent point given the observed data point, we have a probability distribution along the latent space. As the noise shrinks towards 0, the pPCA *plot* converges to the PCA plot, but the principal components are the same regardless. pPCA is particularly useful for handling missing data or denoising it, as it improves over the projection method for obtaining latent positions (Tipping and Bishop, 1999a). That being said, if one is *not* using PCA for this purpose, i.e., one only wants to find the subspace, the importance of pPCA is conceptual rather than technological.

Clustering

k-means clustering

k-means clustering represents data as a set of k means. It partitions observations into groups such that if only k points can be used to represent the data as a whole, these set of k means minimize the sum of squared distances from data to their cluster means, or the $L2$ norm. This *sounds* like finding an affine subspace closest to the set of points in the cluster, where the dimension of the subspace, d, is 0. We can view each cluster's mean as a 0-dimensional PCA for that data. Recalling the definition of variance in eq 1.1, it is clear that choosing clusters in order to minimize the squared distances to their centers is equivalent to minimizing the variance in the clusters. If k-means minimizes the intracluster variance, or the variance *discarded* by the representation of the data by k means, we again arrive at another of our interpretations of PCA. k-means could be viewed as a PCA mixture model.

To find the solution to the k-means optimization problem, an expectation-maximization (EM) approach (Dempster, Laird, and Rubin, 1977) such as Lloyd's algorithm is classically used (Lloyd, 1982). If the clusters were known, minimizing the sum of squared residuals would be trivial — the means of each cluster would be our k means. If the k means were known, we could assign each point to its nearest

mean to construct the cluster — for each point, one residual length will be added to the sum for the whole dataset and assigning a point to any mean other than the closest one will mean adding a longer residual and will increase the sum. Lloyd's EM algorithm for *k*-means clustering iterates between assigning points to clusters based on the chosen means and refining the means based on the chosen clusters. The EM algorithm is guaranteed not to decrease the objective function for *k*-means. This means at each step it either increases or does not improve, at which point we terminate the loop and say it converged. EM is not guaranteed to find the global optimum, only a local one given the clusters at initialization, so it is initialized many times to increase the chance of finding the best solution.

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) can be thought of as an extension of *k*-means in three ways: probabilistic generation of points from the Gaussian distributions representing each cluster, probabilistic assignment of points to clusters by the EM process, and nonspherical clusters. If a point is halfway between two clusters, they can share the responsibility for generating for that point; GMMs allow uncertainty in which cluster a point belongs to. They also are *generative* models unlike *k*-means, meaning that they represent an additional aspect of the data: they not only represent the data's location in feature space but also represent a process for generating the data. Their EM process explicitly aims to maximize the observed likelihood of the data, or the probability that the fitted model would generate the data.

1.7 Thesis structure and contributions

Chapter Two of this thesis revisits enrichment analysis, taking advantage of the newer curation model in GO-CAM pathways to test steps of pathways rather than genes. Unlike GO Biological Processes, which are annotated with unstructured sets of genes, biological pathways are composed of steps, which themselves are annotated in GO-CAMs with sets of genes that are individually sufficient to facilitate them. As our knowledge grows, so does the number of genes annotated to many of these pathways, deflating p-values, yet the number of steps is unchanged. In addition to identifying this issue and providing a solution, I work out a discrete probability distribution that could be used to obtain p-values while accounting for the increased number of genes. Rather than changing the generative process for gene selection, I change the test statistic to be the number of steps with differentially regulated genes. If 5 of 9 single-gene steps in a pathway have an upregulated gene yet the 10th step

has 86 possible genes, it is still a good hypothesis that this pathway is upregulated and therefore should be identified by enrichment analysis.

There is far more mechanistic and topological information contained in GO-CAMs, however, and as part of a long term strategy to harness this, it is useful to characterize the correlation structure of genes and their observed expression levels in human tissues in physiological conditions. Subsequent work suggested that mixing (sometimes unlabeled) sample types was decreasing the positive correlations between many genes and introducing negative correlations through Simpson's paradox (Simpson, 1951), motivating me to develop a method for simultaneously learning subpopulations in data and the linear subspaces they occupy.

Chapter Three of this thesis is the development of the *k*-spaces algorithm, a probabilistic generalization of *k*-means to affine subspaces of mixed dimension, which I discovered was also a mixture model of PCAs through the "closest subspace" definition of PCA. The marginal distribution for data from this model that I independently derived is the same as Probabilistic PCA — with extensions to mixed dimension and optional shared noise parameters — and in fact Tipping and Bishop had explored mixtures of pPCAs (Tipping and Bishop, 1999a) for piecewise approximation of nonlinear manifolds and some clustering, but this approach has not been widely used or deployed for that purpose to the best of my knowledge and certainly not within biology. So, this work introduces the mixture model to biology and reintroduces this to the broader scientific community.

Though developed for clustering gene expression data onto lines, the linear algebra generalizes to any linear subspace, this work makes several important contributions to biological data analysis outside of bulk RNA-sequencing. The first is in the identification of the spatial usage of gene expression states (subspaces) and gradients for their usage (latent variables), which I demonstrate in spatial gene expression data of a developing zebrafish embryo. The second is to cluster on the transposed data matrix with subspaces of different dimension to find the solution to a mathematical problem posed in bulk epigenomics and to provide the probabilistic model for an existing deconvolution approach that employs heuristics. I further provide theoretical analyses of this problem to assess suitability of *k*-spaces, which are enabled by the analytical tractability of the model. The third is to provide a new method for dimensionality reduction in exploratory data analysis by projecting single-cell data across an optimal set of linear dimensionality reductions found by *k*-spaces. The current practice is to rely on nonlinear dimension reductions where axes have no

meaning and apparent structures in data are unreliable. I do this with *C. elegans* and demonstrate that we can visualize cell trajectories in a meaningful latent space while extracting the genes contributing to the structures' formation.

Having completed the first round of development of *k*-spaces — extensions to other probability distributions, especially for single-cell count data are planned — I return to the pathway problem it was developed for in Chapter Four. Turning to the glycolysis pathway and bulk RNA-sequencing in the GTEx dataset of human tissues, I discover that gene expression changes for this and two other metabolic pathways I studied can in fact be well represented with low dimensional latent variables but also that there is more diversity in the stoichiometry of gene expression between tissues than I expected. Nonetheless, *k*-spaces recovers coherent groupings of labeled tissue subtypes into shared subspaces and provides better representations of the data than Pearson correlation coefficients or using one subspace to model the entire dataset alone. Along the way, I identified new challenges for scaling this approach to the rest of our pathways in an automated manner and outline concrete strategies to continue on the path towards combining pathway models with gene expression covariation.

Overall, this dissertation conceptually advances the way we view enrichment analysis in pathways, concretely contributes to the manner in which we can study their gene expression, and ultimately provides a technique for learning linear latent spaces in biology that yields interpretable solutions to the problems it is employed on.

References

Alter, Orly, Patrick O. Brown, and David Botstein (Aug. 2000). "Singular value decomposition for genome-wide expression data processing and modeling". en. In: *Proceedings of the National Academy of Sciences* 97.18, pp. 10101–10106. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.97.18.10101.

Anders, Simon and Wolfgang Huber (2010). "Differential expression analysis for sequence count data". en. In.

Aran, Dvir, Zicheng Hu, and Atul J. Butte (Dec. 2017). "xCell: digitally portraying the tissue cellular heterogeneity landscape". en. In: *Genome Biology* 18.1, p. 220. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1349-1.

Arutyunyan, Anna et al. (Apr. 2023). "Spatial multiomics map of trophoblast development in early pregnancy". en. In: *Nature* 616.7955, pp. 143–151. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-05869-0.

Ashburner, Michael et al. (May 2000). "Gene Ontology: tool for the unification of biology". en. In: *Nature Genetics* 25.1, pp. 25–29. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/75556.

Avery, Oswald T., Colin M. MacLeod, and Maclyn McCarty (Feb. 1944). "Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated From Pneumococcus Type III". In: *Journal of Experimental Medicine* 79.2, pp. 137–158.

Bolstad, B.M. et al. (Jan. 2003). "A comparison of normalization methods for high densityoligonucleotide array data based on variance and bias". en. In: *Bioinformatics* 19.2, pp. 185–193. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/19.2.185.

Booeshaghi, A. Sina, Ángel Galvez-Merchán, and Lior Pachter (Mar. 2024). *Algorithms for a Commons Cell Atlas*. en. DOI: 10.1101/2024.03.23.586413.

Buettner, Florian et al. (Dec. 2017). "f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq". en. In: *Genome Biology* 18.1, p. 212. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1334-8.

Chapman, Alec R. et al. (Dec. 2022). "Correlated gene modules uncovered by high-precision single-cell transcriptomics". en. In: *Proceedings of the National Academy of Sciences* 119.51, e2206938119. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2206938119.

Choe, Mangyu et al. (May 2025). "Glycolysis model shows that allostery maintains high ATP and limits accumulation of intermediates". en. In: *Biophysical Journal* 124.10, pp. 1562–1586. ISSN: 00063495. DOI: 10.1016/j.bpj.2025.03.037.

Cloonan, Nicole et al. (July 2008). "Stem cell transcriptome profiling via massive-scale mRNA sequencing". en. In: *Nature Methods* 5.7, pp. 613–619. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1223.

Collins, Michael, Sanjoy Dasgupta, and Robert E. Schapire (Nov. 2002). "A Generalization of Principal Component Analysis to the Exponential Family". en. In: *Advances in Neural Information Processing Systems 14*. Ed. by Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani. The MIT Press, pp. 617–624. ISBN: 978-0-262-27173-8. DOI: 10.7551/mitpress/1120.003.0084.

Consortium, The GTEx (2020). "The GTEx Consortium atlas of genetic regulatory effects across human tissues". en. In: *Science* 369.6509, pp. 1318–1330. DOI: DOI:10.1126/science.aaz1776.

Consortium, The Tabula Muris (Oct. 2018). "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris". In: *Nature* 562, pp. 367–372. DOI: https://doi.org/10.1038/s41586-018-0590-4.

Crick, Francis (Aug. 1970). "Central Dogma of Molecular Biology". en. In: *Nature* 227.

Crow, Megan et al. (Mar. 2019). "Predictability of human differential gene expression". en. In: *Proceedings of the National Academy of Sciences* 116.13, pp. 6491–6500. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1802973116.

Cui, Haotian et al. (Aug. 2024). "scGPT: toward building a foundation model for single-cell multi-omics using generative AI". en. In: *Nature Methods* 21.8, pp. 1470–1480. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-024-02201-0.

CZI Single-Cell Biology Program et al. (Nov. 2023). *CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data*. en. DOI: 10.1101/2023.10.30.563174.

Dempster, A. P., N. M. Laird, and D. B. Rubin (Sept. 1977). "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39.1, pp. 1–22. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

Dixit, Atray et al. (Dec. 2016). "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". en. In: *Cell* 167.7, 1853–1866.e17. ISSN: 00928674. DOI: 10.1016/j.cell.2016.11.038.

Donovan, Margaret K. R. et al. (Feb. 2020). "Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants". en. In: *Nature Communications* 11.1, p. 955. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14561-0.

Düvel, Katrin et al. (July 2010). "Activation of a Metabolic Gene Regulatory Network Downstream of mTOR Complex 1". en. In: *Molecular Cell* 39.2, pp. 171–183. ISSN: 10972765. DOI: 10.1016/j.molcel.2010.06.022.

Eckart, Carl and Gale Young (Sept. 1939). "The approximation of one matrix by another of lower rank". In: *Psychometrika* 1, pp. 211–218.

Elyanow, Rebecca et al. (Feb. 2020). "netNMF-sc: leveraging gene–gene interactions for imputation and dimensionality reduction in single-cell expression analysis". en. In: *Genome Research* 30.2, pp. 195–204. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.251603.119.

Fire, Andrew et al. (Feb. 1998). "Potent and specific genetic interference by double-stranded RNA in". en. In: *Nature* 391.

Francis H. Crick et al. (Dec. 1961). "General Nature of the Genetic Code for Proteins". In: *Nature* 192, pp. 1227–1232.

Friedman, Nir (Feb. 2004). "Inferring Cellular Networks Using Probabilistic Graphical Models". en. In: *Science* 303.5659, pp. 799–805. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1094068.

Gewirtz, Ariel D.H. (2022). "Exploring Context-Specific Expression Variation and Distal Gene Regulation Via Latent Variable Models". PhD thesis. Princeton, NJ: Princeton University.

Gibson, Greg (May 2015). "GTEx detects genetic effects". en. In: *Science* 348.6235, pp. 640–641. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aab3002.

Glaab, Enrico et al. (Sept. 2012). "EnrichNet: network-based gene set enrichment analysis". en. In: *Bioinformatics* 28.18, pp. i451–i457. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/bts389.

Good, Benjamin M et al. (Oct. 2021). "Reactome and the Gene Ontology: digital convergence of data resources". en. In: *Bioinformatics* 37.19. Ed. by Zhiyong Lu, pp. 3343–3348. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btab325.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. MIT Pres.

Gorin, Gennady et al. (Dec. 2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". en. In: *Nature Communications* 13.1, p. 7620. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34857-7.

GTEx Consortium (Oct. 2017). "Genetic effects on gene expression across human tissues". en. In: *Nature* 550.7675, pp. 204–213. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature24277.

Guo, Jimmy K. and Mitchell Guttman (Oct. 2022). "Regulatory non-coding RNAs: everything is possible, but what is important?" en. In: *Nature Methods* 19.10, pp. 1156–1159. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-022-01629-6.

Hall, Benjamin D. and S. Spiegelman (Feb. 1961). "SEQUENCE COMPLEMENTARITY OF T2-DNA AND T2-SPECIFIC RNA". en. In: *Proceedings of the National Academy of Sciences* 47.2, pp. 137–146. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.47.2.137.

Han, Min et al. (May 1993). "C. elegans lin-45 raf gene participates in let-60 ras-stimulated vulval differentiation". In: *Nature* 363, pp. 133–140.

Harold Hotelling (Oct. 1933). "Analysis of a complex of statistical variables into principal components". In: *Journal of educational psychology* 24.6, pp. 417–441.

Hershey, Alfred D. and Martha Chase (Sept. 1952). "Independent functions of viral protein and nucleic acid in growth of bacteriophage". In: *Journal of General Physiology* 36.1, pp. 39–56.

Hickey, John W. et al. (July 2023). "Organization of the human intestine at single-cell resolution". en. In: *Nature* 619.7970, pp. 572–584. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-05915-x.

Higuchi, Russell, Gavin Dollinger, et al. (Apr. 1992). "Simultaneous Amplification and Detection of Specific DNA Sequences". en. In: *Bio/Technology* 10.4, pp. 413–417. DOI: 10.1038/nbt0492-413.

Higuchi, Russell, Carita Fockler, et al. (Sept. 1993). "Kinetic PCR analysis: real-time monitoring of DNA amplification reactions". In: *Bio/Technology* 11.9, pp. 1026–1030. DOI: doi:10.1038/nbt0993-1026.

Horace Freeland Judson (1996). "II. RNA. The Functions of the Stucture: The breaking of the genetic code, the discovery of the messenger". In: *The Eighth Day of Creation*. 25th Aniversary Edition. New York, USA: Cold Spring Harbor Laboratory Press, p. 333. ISBN: 978-0-87969-478-4.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki (Jan. 2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". en. In: *Nucleic Acids Research* 37.1, pp. 1–13. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkn923.

HuBMAP Consortium et al. (Oct. 2019). "The human body at cellular resolution: the NIH Human Biomolecular Atlas Program". en. In: 574.7777, pp. 187–192. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-019-1629-x.

J. Robert Adcock (Mar. 1878). "A Problem in Least Squares". en. In: *The Analyst* 5.2, pp. 53–54.

Jiang, Jialong et al. (Apr. 2023). *D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing principles of cellular perturbation response*. en. DOI: 10.1101/2023.04.19.537364.

Kanehisa, Minoru and Susumu Goto (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes". en. In: *Nucleic Acids Research* 28.1.

Karlebach, Guy and Ron Shamir (Oct. 2008). "Modelling and analysis of gene regulatory networks". en. In: *Nature Reviews Molecular Cell Biology* 9.10, pp. 770–780. ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm2503.

Kim-Hellmuth, Sarah et al. (Sept. 2020). "Cell type–specific genetic regulation of gene expression across human tissues". en. In: *Science* 369.6509, eaaz8528. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaz8528.

Kotiang, Stephen and Ali Eslami (May 2020). "A probabilistic graphical model for system-wide analysis of gene regulatory networks". en. In: *Bioinformatics* 36.10. Ed. by Alfonso Valencia, pp. 3192–3199. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btaa122.

Kotliar, Dylan et al. (July 2019). "Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq". en. In: *eLife* 8, e43803. ISSN: 2050-084X. DOI: 10.7554/eLife.43803.

Kunes, Russell Z. et al. (July 2024). "Supervised discovery of interpretable gene programs from single-cell data". en. In: *Nature Biotechnology* 42.7, pp. 1084–1095. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-023-01940-3.

Lachmann, Alexander et al. (Apr. 2018). "Massive mining of publicly available RNA-seq data from human and mouse". en. In: *Nature Communications* 9.1, p. 1366. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03751-6.

Lake, Blue B. et al. (July 2023). "An atlas of healthy and injured cell states and niches in the human kidney". en. In: *Nature* 619.7970, pp. 585–594. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-05769-3.

Langfelder, Peter and Steve Horvath (Dec. 2008). "WGCNA: an R package for weighted correlation network analysis". en. In: *BMC Bioinformatics* 9.1, p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559.

Lim, Hyun Ju et al. (Jan. 2025). "A practical guide for choosing an optimal spatial transcriptomics technology from seven major commercially available options". en. In: *BMC Genomics* 26.1, p. 47. ISSN: 1471-2164. DOI: 10.1186/s12864-025-11235-3.

Lister, Ryan et al. (May 2008). "Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis". en. In: *Cell* 133.3, pp. 523–536. ISSN: 00928674. DOI: 10.1016/j.cell.2008.03.029.

Lloyd, Stuart P. (Mar. 1982). "Least Squares Quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137.

Lu, Jonathan et al. (Jan. 2021). "Causal network inference from gene transcriptional time-series response to glucocorticoids". en. In: *PLOS Computational Biology* 17.1. Ed. by Christina S. Leslie, e1008223. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008223.

Lubeck, Eric et al. (Apr. 2014). "Single-cell in situ RNA profiling by sequential hybridization". en. In: *Nature Methods* 11.4, pp. 360–361. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2892.

Lyon, Mary F (Apr. 1961). "Gene Action in the X-chromosome of the Mouse (Mus musculus L.)" en. In: *Nature* 190.4773, pp. 372–373. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/190373a0.

Marioni, John C. et al. (Sept. 2008). "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays". en. In: *Genome Research* 18.9, pp. 1509–1517. ISSN: 1088-9051. DOI: 10.1101/gr.079558.108.

Milacic, Marija et al. (Jan. 2024). "The Reactome Pathway Knowledgebase 2024". en. In: *Nucleic Acids Research* 52.D1, pp. D672–D678. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkad1025.

Mortazavi, Ali et al. (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". en. In: *Nature Methods* 5.7, pp. 621–628. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.1226.

Nirenberg, Marshall W. and J. Heinreich Matthaei (Oct. 1961). "The Dependence of Cell-free Protein Synthesis in E. coli Upon Naturally Occuring Or Synthetic Polyribonucleotides". In: *Proceedings of the National Academy of Sciences* 47.10, pp. 1588–1602.

Pearson, Karl (Nov. 1901). "LIII. *On lines and planes of closest fit to systems of points in space*". en. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11, pp. 559–572. ISSN: 1941-5982, 1941-5990. DOI: 10.1080/14786440109462720.

Pierson, Emma et al. (May 2015). “Sharing and Specificity of Co-expression Networks across 35 Human Tissues”. en. In: *PLOS Computational Biology* 11.5. Ed. by Isidore Rigoutsos, e1004220. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004220.

Piro, Rosario Michael et al. (Nov. 2011). “An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction”. en. In: *European Journal of Human Genetics* 19.11, pp. 1173–1180. ISSN: 1018-4813, 1476-5438. DOI: 10.1038/ejhg.2011.96.

Quinodoz, Sofia A. and Mitchell Guttman (May 2022). “Essential Roles for RNA in Shaping Nuclear Organization”. en. In: *Cold Spring Harbor Perspectives in Biology* 14.5, a039719. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a039719.

Reimand, Jüri et al. (Feb. 2019). “Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap”. en. In: *Nature Protocols* 14.2, pp. 482–517. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/s41596-018-0103-9.

Roberts, Adam et al. (Mar. 2011). “Improving RNA-Seq expression estimates by correcting for fragment bias”. en. In: *Genome Biology* 12.3, R22. ISSN: 1474-760X. DOI: 10.1186/gb-2011-12-3-r22.

Robinson, Mark D and Alicia Oshlack (Mar. 2010). “A scaling normalization method for differential expression analysis of RNA-seq data”. en. In: *Genome Biology* 11.3, R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25.

Roweis, Sam (July 1998). “EM Algorithms for PCA and SPCA”. en. In: Denver, Colorado, USA, pp. 626–632.

Saha, Ashis et al. (Nov. 2017). “Co-expression networks reveal the tissue-specific regulation of transcription and splicing”. en. In: *Genome Research* 27.11, pp. 1843–1858. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.216721.116.

Schena, Mark et al. (Oct. 1995). “Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray”. en. In: *Science* 270.5235, pp. 467–470. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.270.5235.467.

Sherman, Brad T et al. (July 2022). “DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update)”. en. In: *Nucleic Acids Research* 50.W1, W216–W221. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkac194.

Simpson, E. H. (July 1951). “The Interpretation of Interaction in Contingency Tables”. en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 13.2, pp. 238–241. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/j.2517-6161.1951.tb00088.x.

Subramanian, Aravind et al. (Oct. 2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0506580102.

Sutton, Gavin J. et al. (Mar. 2022). "Comprehensive evaluation of deconvolution methods for human brain gene expression". en. In: *Nature Communications* 13.1, p. 1358. ISSN: 2041-1723. DOI: 10.1038/s41467-022-28655-4.

Temin, Howard M. and Satoshi Mizutani (June 1970). "RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus". In: *Nature* 226, pp. 1211–1213.

The ENCODE Project Consortium (Sept. 2012). "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414, pp. 57–74. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11247.

The Gene Ontology Consortium et al. (May 2023). "The Gene Ontology knowledgebase in 2023". en. In: *GENETICS* 224.1. Ed. by A Baryshnikova, iyad031. ISSN: 1943-2631. DOI: 10.1093/genetics/iyad031.

The GTEx Consortium et al. (May 2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". en. In: *Science* 348.6235, pp. 648–660. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1262110.

Thomas, Paul D., Michael J. Campbell, et al. (Sept. 2003). "PANTHER: A Library of Protein Families and Subfamilies Indexed by Function". en. In: *Genome Research* 13.9, pp. 2129–2141. ISSN: 1088-9051. DOI: 10.1101/gr.772403.

Thomas, Paul D., David P. Hill, et al. (Oct. 2019). "Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems". en. In: *Nature Genetics* 51.10, pp. 1429–1433. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0500-1.

Tipping, Michael E. and Christopher M. Bishop (Feb. 1999a). "Mixtures of Probabilistic Principal Component Analyzers". en. In: *Neural Computation* 11.2, pp. 443–482. DOI: 10.1162/089976699300016728.

– (Sept. 1999b). "Probabilistic Principal Component Analysis". en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3, pp. 611–622. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/1467-9868.00196.

Trapnell, Cole et al. (May 2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". en. In: *Nature Biotechnology* 28.5, pp. 511–515. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.1621.

Trivedi, Vikas et al. (Jan. 2018). "Multidimensional quantitative analysis of mRNA expression within intact vertebrate embryos". en. In: *Development* 145.1, dev156869. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.156869.

Volkin, Elliot, L. Astrachan, and Joan L. Countryman (Oct. 1958). “Metabolism of RNA phosphorus in Escherichia coli infected with bacteriophage T7”. en. In: *Virology* 6.2, pp. 545–555. ISSN: 00426822. DOI: 10.1016/0042-6822(58)90101-6.

Wang, Zhong, Mark Gerstein, and Michael Snyder (Jan. 2009). “RNA-Seq: a revolutionary tool for transcriptomics”. en. In: *Nature Reviews Genetics* 10.1, pp. 57–63. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2484.

Watson, James D. and Francis H. Crick (Apr. 1953). “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171, pp. 737–738.

West, Mike et al. (Sept. 2001). “Predicting the clinical status of human breast cancer by using gene expression profiles”. en. In: *Proceedings of the National Academy of Sciences* 98.20, pp. 11462–11467. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.201162998.

Wilhelm, Brian T. et al. (June 2008). “Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution”. en. In: *Nature* 453.7199, pp. 1239–1243. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature07002.

Wiśniewski, Jacek R., Agnieszka Gizak, and Dariusz Rakus (Aug. 2015). “Integrating Proteomics and Enzyme Kinetics Reveals Tissue-Specific Types of the Glycolytic and Gluconeogenic Pathways”. en. In: *Journal of Proteome Research* 14.8, pp. 3263–3273. ISSN: 1535-3893, 1535-3907. DOI: 10.1021/acs.jproteome.5b00276.

Wu, Chaodong, Salmaan A. Khan, and Alex J. Lange (Nov. 2005). “Regulation of glycolysis—role of insulin”. en. In: *Experimental Gerontology* 40.11, pp. 894–899. ISSN: 05315565. DOI: 10.1016/j.exger.2005.08.002.

Zhang, Allen W. et al. (Oct. 2019). “Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling”. en. In: *Nature Methods* 16.10, pp. 1007–1015. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0529-1.

Zhao, Kangmei and Seung Yon Rhee (Apr. 2023). “Interpreting omics data with pathway enrichment analysis”. en. In: *Trends in Genetics* 39.4, pp. 308–319. ISSN: 01689525. DOI: 10.1016/j.tig.2023.01.003.

Chapter 2

ENRICHMENT ON STEPS, NOT GENES, IMPROVES INFERENCE OF DIFFERENTIALLY EXPRESSED PATHWAYS

Markarian, Nicholas, Kimberly M. Van Auken, Dustin Ebert, and Paul W. Sternberg (Mar. 2024). "Enrichment on steps, not genes, improves inference of differentially expressed pathways". en. In: *PLOS Computational Biology* 20.3. Ed. by Marc Robinson-Rechavi, e1011968. ISSN: 1553-7358. DOI: https://10.1371/journal.pcbi.1011968. URL: https://dx.plos.org/10.1371/journal.pcbi.1011968.

Abstract

Enrichment analysis is frequently used in combination with differential expression data to investigate potential commonalities amongst lists of genes and generate hypotheses for further experiments. However, current enrichment analysis approaches on pathways ignore the functional relationships between genes in a pathway, particularly OR logic that occurs when a set of proteins can each individually perform the same step in a pathway. As a result, these approaches miss pathways with large or multiple sets because of an inflation of pathway size (when measured as the total gene count) relative to the number of steps. We address this problem by enriching on step-enabling entities in pathways. We treat sets of protein-coding genes as single entities, and we also weight sets to account for the number of genes in them using the multivariate Fisher's noncentral hypergeometric distribution. We then show three examples of pathways that are recovered with this method and find that the results have significant proportions of pathways not found in gene list enrichment analysis.

Author Summary

Genome-scale experiments typically identify sets of genes which are primarily analyzed by enrichment analysis to identify relevant pathways that may be perturbed. Curated pathway models have rich structure that we believe can be exploited to get better results. Some pathway steps are enabled by sets of interchangeable genes which inflate the gene count of their respective pathways relative to the number of steps. We improve sensitivity towards these pathways in enrichment analysis by performing enrichment on steps. We then use this approach to identify pathways that would otherwise be missed in medically relevant datasets to gain new insights.

Introduction

High-throughput experiments regularly output large lists of genes that vary in expression across conditions and cell types or are perturbed in disease states (e.g. 1,2). While these experiments can establish transcriptional signatures for cells or diseases, the interpretation of these lists of genes in the context of physiology or phenotype can prove difficult, even for domain experts, as the collective body of knowledge in the literature grows at an ever increasing rate (3). Enrichment tools aim to aid that analysis by comparing annotations, such as disease, process, and pathway associations, associated with the outputted list of differentially expressed genes to those of sets of genes in databases that previous studies have linked to specific diseases, biological processes, and pathways (4–7). While the enrichment analysis field has made many advances, it has treated pathway enrichment the same as enrichment on categorical terms without consideration for pathways' inherent structure. In its simplest form, enrichment analysis searches for overrepresented annotations within lists of genes. It performs pairwise evaluations of the overlap between the list from a particular experiment and reference lists in knowledgebases to determine if any of those overlaps are greater than would be expected by chance. Early work searched for overrepresentation in categories such as diseases or particular Gene Ontology (GO) terms for cellular compartments, molecular functions, or biological processes (6,8,9). As pathway databases such as Reactome and KEGG were developed and expanded (10,11), pathways were incorporated into enrichment analyses by applying the same algorithms and treating pathway membership as an annotation to form a list, although this eliminated causal relationships and pathway structure. Later work introduced more sophisticated statistical procedures to address open problems in enrichment such as utilizing fold changes in expression (12), leveraging relationships between annotation terms (7), or incorporating protein-protein interaction networks (13), but these methods still treated pathways the same as other reference gene sets, considering pathway membership as an annotation. Unlike the previously mentioned classifications for genes, such as the category of genes with products active in a specific organelle, biological pathways have structure (i.e., have causal, directional relationships between participating genes), and thus are not simply categories or lists, but this issue has not been addressed.

We utilized an aspect of pathway structure, sets, in our enrichment on pathways modeled in Gene Ontology Causal Activity Models (GO-CAMs) (14), and this enabled us to recover pertinent biological pathways that could otherwise be missed. GO-CAMs are a type of pathway model centered around GO molecular functions

and use other ontology terms to provide relevant biological context. GO-CAMs are typically curated manually by members of the GO consortium (15), but an additional source of human GO-CAMs is computationally generated from conversion of pathways in Reactome (16), a popular pathway database also used to define gene lists for pathways in widely used enrichment analysis tools such as PANTHER and DAVID (4,7,10). In examining causal flow in GO-CAMs, we realized that another relationship between genes annotated to pathways has been neglected in the conversion to lists but is becoming recognized in other analyses (17): OR logic via interchangeability of gene products at certain steps in pathways. This interchangeability is represented explicitly in Reactome (and by extension, GO CAMs) as "sets," defined as groups of proteins or protein complexes that are individually sufficient to perform the same step in a given pathway, and implicitly in KEGG, where they can be inferred by annotation of multiple Enzyme Commission numbers to a reaction, making our work broadly applicable. For example, the Reactome set "Glucokinase and Hexokinases" is comprised of glucokinase and hexokinases 1, 2, and 3. Any one of these proteins is sufficient to phosphorylate glucose, the first step in the glycolysis pathway. Furthermore, glucokinase is only expressed in the liver and pancreas and is not available to be up or downregulated by other cell types. Thus, sets can either be a consequence of annotation decisions, such as using one pathway diagram that may differ from cell type to cell type, or they can be a direct representation of biology, where multiple gene products may substitute for one another, albeit with potentially distinct reaction kinetics.

We can contrast sets with complexes in terms of logic. For example, microtubules are formed from tubulin $\alpha \cdot \beta$ dimers. Microtubule formation requires both tubulin α AND tubulin β. In contrast, phosphorylating glucose requires either glucokinase OR hexokinase 1 OR hexokinase 2 OR hexokinase 3. (In fact, there are actually 8 genes that encode α tubulins and 9 genes that encode β tubulins, many of which have cell or tissue specific expression, so sets can be found in the context of complexes as well (18)). Sets enable curators to avoid creating multiple, otherwise redundant instances of pathways when different gene products may perform the same step in different cells or within the same cell; a single instance of a pathway model is created, and the set indicates the variability at that step. Ideally, enrichment analysis would acknowledge this variability and have some degree of robustness to the decision to annotate additional genes that can enable a pathway.

However, widely used enrichment tools such as those at PANTHER and Reactome

do not account for these sets, nor does any other tool of which we are aware. This can be problematic, because sets inflate the count of all genes annotated to a pathway when they are expanded to create a gene list for enrichment, but they do not increase the number of steps. For example, the BMP signaling pathway has 7 receptors, each individually sufficient to facilitate signaling, and they are expressed in many tissues at varying levels (19). There are many other steps in this pathway, but for argument's sake, suppose there were only two other steps, one enabled by a complex of two gene products and the other enabled by one gene product, for a total of 10 genes annotated to the pathway. If a cell upregulated expression of one member of the set of receptors and the single gene product for the last step, this scenario will be treated as 2 of 10 genes in the pathway, even though 2 of 3 steps are affected. Furthermore, complexes are treated the same as sets even though the logical relationship between their members differs. Increased expression of just one member of the proteasome complex likely does not mean increased proteasome activity, but increased expression of a member of a set of enzyme activators, receptors, or enzymes may be impactful. Due to the inflation of n, the gene count of the pathway, the pathway may not be captured by the enrichment analysis. Researchers using enrichment tools usually seek to uncover which pathways are more active in different conditions, a question that is more directly dependent on the proportion of steps in a pathway that are up and down regulated than on the proportion of genes annotated to a pathway, given that some of those genes can act in each other's stead.

This study implements enrichment on Gene Ontology Causal Activity Models (GO-CAMs) (14) and explores the impact of "sets," a feature in pathway models neglected in current enrichment tools, seeking to integrate it into analysis. We discover that some very large gene sets greatly inflate the gene count of the pathways in which they are members if sets are treated the same as complexes, impeding the pathways from being captured by enrichment analysis tools. We propose accounting for this by performing enrichment analysis on the pathway steps rather than directly on the genes themselves. Using a one-tailed hypergeometric test while treating sets as single entities, we showcase three examples of enriched pathways and then evaluate results on datasets from six studies (20–25). We show that while the enrichment results largely overlap with those yielded by enriching directly on the list of genes, a significant proportion of results are unique. Lastly, we consider how the assumptions of the null hypothesis change when treating sets as single entities and introduce enrichment analysis on pathway steps via the multivariate Fisher's

noncentral hypergeometric distribution to weight sets according to the number of genes, in line with the traditional assumption that each gene is chosen with uniform probability.

Results

Formulating a step-focused null hypothesis for enrichment by including sets

Traditionally, enrichment analysis with the hypergeometric test asks the question "What is the probability that k or more out of n genes associated with the pathway are in a list of length N, where those N genes are sampled from a background of size M?" We want to propose a question focused on the steps that form a pathway instead. Defining entities as: 1) the proteins, 2) protein subunits of complexes, 3) and sets of proteins that perform steps in a pathway, we ask "Given the steps in the pathway and the entities that enable them, what is the probability that the k or more out of n entities required to enable those steps are in a list of length N, where those N entities are sampled from a background of size M?" Both null hypotheses assume that each gene or entity is sampled independently and with equal probability. We don't formally state the question as "What is the probability that k or more out of n steps in a pathway are sampled," because complexes are split into their protein subunits, but that is the underlying idea. In a pathway with no complexes, the questions are equivalent, and we want to represent a scenario where cells regulate pathways by selecting steps to regulate without replacement. We illustrate the comparison in the gene lists used by the two methods in Fig. 2.1 and detail the procedures for enrichment in the next section.

Pathway enrichment procedure

Enriching on steps, as shown above, requires mapping the input list of genes from an experiment to the list of step-enabling entities that those genes belong to. This list consists of 1) any sets that have at least one member in the input and 2) any genes in the input that are the sole genes to enable a step in a pathway (not part of a set). Enrichment can be performed using the one-tailed hypergeometric test with this modified input list and the step-enabling list of entities for pathway i, L_i. This is also known as a one-tailed Fisher's exact test (26). The key result is that n_i, the length of list L_i, is the minimum number of genes required to enable all steps of the pathway. Traditionally, ni is the total number of genes associated with a pathway, which could greatly exceed the number of steps in the pathway due to large sets. A comparison of the algorithms is shown in Fig. 2.2.

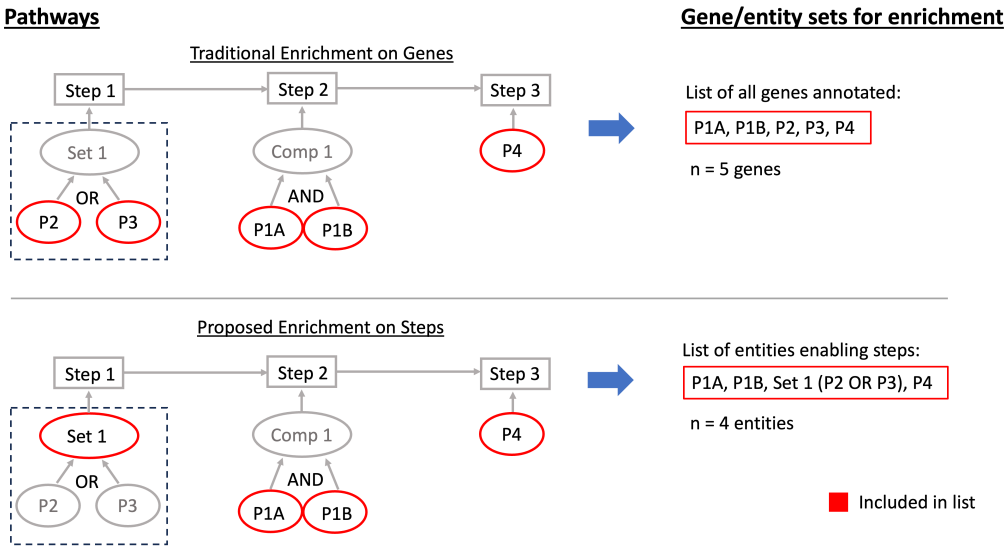


Figure 2.1: **Enriching directly on genes inflates the number of pathway elements relative to the number of steps.** The pathway model (in blue) consists of 3 steps, enabled by a complex, a set, and a protein respectively. Traditional enrichment extracts the list of all genes associated with a pathway, treating complexes and sets equivalently, even though the logical interpretation of a complex is a joining of its members through an AND relation while members of a set are linked by OR. Enrichment on steps accounts for sets while creating the lists, where Set 1 acts as a placeholder for "Protein 1 OR Protein 2", and Complex 1 is treated as "Protein 3A AND Protein 3B". The pathway is enabled by "Set 1 AND Complex 1 AND Protein 4," which is equivalent to "(P1 OR P2) AND (P3A AND P3B) AND P4." Hence, the list we enrich on is "Set 1, P3A, P3B, P4," where Set 1 is (P2 OR P3). Importantly, the size of the list used in enrichment is equal to the minimum number of genes required to enable the pathway in our step-centric enrichment but not in traditional enrichment on gene lists, which uses the total gene count.

Complexes pose a design challenge with enriching on steps, because it is unclear what it would mean to alter expression of one of the members of a protein complex but not the others. This depends on whether a particular complex is assembled upon translation or later through protein interactions, as well as knowledge of which proteins are the limiting factor due to stoichiometry and/or assembly kinetics. In addition, some annotated complexes in pathway databases are transiently formed during signal transduction, such as IL7R-JAK-STAT, a complex in Reactome (10). We decided to treat complexes the same way as they have been previously: each complex is mapped to its protein members, and those proteins are considered part of the list for the pathway, just as each member of a complex is traditionally added to the list of genes for the pathway (e.g. 4). If a protein complex is necessary to

Enrichment on Gene Lists:	Enrichment on Steps:
1. For each pathway *i*, construct a list of *genes*, L_i	1. For each pathway *i*, construct a list of *entities* (genes and sets) that enable steps, L_i
1.1 Extract all entities that enable steps in pathway.	1.1 Extract all entities that enable steps in pathway.
1.2 Add entities that are protein-coding genes to L_i.	1.2 Add entities that are protein-coding genes to L_i.
1.3 Map complexes to subunits, add them to L_i.	1.3 Map complexes to subunits, add them to L_i.
1.4 Map sets to members and treat those members according to (1.2) and (1.3).	1.4 Add each set directly to L_i as its own entity. a) If a set contains complexes, treat the intersection of all members' subunits as a complex according to (1.3). Form a set with the remainder and treat according to (1.4).
1.5 After all lists are created, define M, the background size, as the total number of unique genes across all lists.	1.5 After all lists are created, define M, the background size, as the total number of unique entities across all lists
** n_i, the length of L_i, is the number of genes annotated to pathway *i* **	** n_i, the length of L_i, is the minimum number of protein-coding genes required to enable all steps in pathway *i* **
2. Compute p-value for each pathway *i*	2. Compute p-value for each pathway *i*
2.1 Remove any genes in input list not found in any L_i. Define N as the length of this list.	2.1 Map input list to any sets which have at least one member in the input. Remove any genes in input list not found in any L_i not as a set. Concatenate these two lists and define N as the length of this list.
2.2 Map genes on the filtered list to each L_i.	2.2 Map the entities on the new input list to each L_i.
2.3 For each L_i, define k_i as the size of the overlap (# genes from the input mapped to L_i).	2.3 For each L_i, define k_i as the size of the overlap (# genes from the input mapped to L_i).
2.4 Use the hypergeometric distribution to calculate the probability of an overlap ≥ k_i given n_i, N, and M.	2.4 Use the hypergeometric distribution to calculate the probability of an overlap ≥ k_i given n_i, N, and M.

Figure 2.2: **Step by step comparison of enrichment algorithms.** Multiple testing correction is not shown here but is done with the Benjamini-Hochberg procedure.

perform a step in a pathway, we consider their protein members to be necessary as well, acknowledging the limitation that this allows for partial contribution to enrichment when in some cases, it should biologically be all-or-nothing.

Sets of complexes usually have one or more common subunits and differ only in one subunit, so we create a new complex out of the common subunits (the subunit intersection across all the complexes), and then a new set out of the remainder (Box 2.2). That new complex is then mapped to its subunit members, and each is treated as an entity. Except in the cases where the set of complexes is a heterogenous group or has multiple specific subunits, this faithfully represents sets of complexes

in a manner consistent with our representation of complexes and of sets of proteins. For example, Prolyl 4-hydroxylase is a complex with 2 P4HB1 beta subunits and 2 identical alpha subunits from P4HA1, P4HA2, or P4HA3. This is represented as a set of complexes (2 P4HA1: 2 P4HB1 OR 2 P4HA2: 2 P4HB1 OR 2 P4HA3: 2 P4HB1), but we represent it with P4HB1 AND (P4HA1 OR P4HA2 OR P4HA3). We recursively apply the above logic to reduce these to proteins and sets.

Comparing parameter changes in gene list and step-centric enrichment

These changes affect the hypergeometric test primarily by reducing n, the size of the pathway against which the overlap, k, is compared to. M, the size of the background list is also reduced, because 2300 genes only appear in pathways as part of sets and thus are not unique entities. All else equal, the reduction of n lowers the p-value, while the reduction of M increases it. We constructed lists of genes for each pathway via the standard gene list method and compared these to the entity lists produced by our method for each pathway. While the change in n is pathway-specific, the median reduction per pathway from the gene-list method to ours is 6%, with a 75th percentile reduction of 40%, indicating that most models are unaffected by the change, but a minority are significantly impacted (Table 2.1). The change in M is a reduction from 5386 genes annotated across all pathways to 3983 entities (genes and sets of genes) annotated across all pathways. (Of the 5386 genes, only 3086 are the sole entities that enable at least one step, meaning they appear outside of sets, while 1400 are only annotated as part of sets. There are a total of 915 sets). N can change as well, but the magnitude and direction of change are dependent on the input list. Lastly, k can be reduced because we only allow for each set to count once towards the overlap, even if more than one gene in the set is on the input list. This reduction is a consequence of enriching on entities and treating each set as one entity. We discuss this issue further in the section on limitations.

Pathway size change (gene list length vs entity list length)	Absolute Reduction $n_{\text{gene list}} - n_{\text{entity list}}$	Relative Reduction $\frac{(n_{\text{gene list}} - n_{\text{entity list}})}{n_{\text{gene list}}}$
50^{th} percentile	1	3%
75^{th} percentile	4	40%
95^{th} percentile	19	76%
Max	255	98%

Table 2.1: **Absolute and relative reductions in pathway sizes (n) from gene enrichment to step enrichment.**

Case studies of pathways missed by gene list enrichment analysis

We sought specific use cases of published RNAseq or quantitative proteomics datasets to compare the gene-list method and our step-centric method using two criteria: 1) single cell sequencing or bulk sequencing on only one cell type and 2) that the datasets were generated from different organ systems and different pathologies. We wanted to compare one cell type or state versus another, as our study is motivated by accounting for the sufficiency of members of sets for enabling steps in pathways, and pooling data from multiple cell types or conditions could undermine this and lead to spurious results. For example, a list of mutated genes implicated in gastric cancers do not characterize a single cell state, as these mutations do not all coexist in one cell but are aggregated from many patients' tumors. We also wanted diversity in our datasets as the GO-CAM pathway database does not yet represent all processes in the cell, and annotations (and therefore enrichment) may be relatively scant for some organ systems or processes. We evaluated enrichment on six datasets, one proteomic, representing cell line response to p97 inhibitors (21), and five transcriptomic: changes in platelets during SARS-CoV-2 infection (20), dilated and arrhythmogenic cardiomyopathy (22), nonalcoholic steatohepatitis (24), astrocytes in various regions of aging mouse brains (23), and activation of macrophages in vitro (25). These datasets were from the first six papers we found that met our criteria and had readily available differential expression data, as we wanted to provide an unbiased assessment in our intended use case. First, we present case studies from three of these datasets (platelets in SARS-CoV-2 infection, p97 inhibitors, and dilated and arrhythmogenic cardiomyopathy) and then we aggregate results from all six to characterize the divergence between enrichment methods.

Platelets in SARS-CoV-2 infection may produce collagen, and VLCFA synthesis is reduced

COVID-19 patients suffer from increased rates of thrombotic events, and Manne et al., 2020 sought to understand whether platelets in these patients are more prone to aggregation and, if so, how (20). They collected platelets from the blood of patients hospitalized with COVID-19, performed bulk RNAseq, and compared them to platelet transcriptomes of healthy donors. While platelets themselves do not have nuclei and cannot synthesize mRNA, they retain mRNA from their megakaryocyte precursors and can synthesize proteins. One striking result we uncovered is the collagen biosynthetic pathway, which has 10 entities of 12 upregulated in SARS-CoV-2 infection (Fig. 2.3), but there are 65 members in the gene list for this pathway

(unadjusted p-value of 0.087 for 8/65 genes in our implementation; not reported as a result by PANTHER for 9/67 genes; unadjusted p-value of 0.22 for 9/76 entities in the Reactome Analysis tool due to slight annotation differences (4,10)). GSEA analysis utilizing the log2 fold changes yielded an FDR adjusted p-value of 0.52 (12,27). In this case GO CAM model size is clearly inflated by the set of '4-Hydroxyproline collagen propeptides', which represents 44 procollagen genes. See Discussion for further discussion of the potential significance of this and support in literature, particularly from Kyselova, 2021, who provided the first biochemical evidence that platelets can be a source of collagen in another disease, Polycystic Ovarian Syndrome. While collagen I levels in COVID-19 patients would need to be assayed to confirm this, and if true, its significance or lack thereof to increased platelet activity in COVID-19 would need to be studied, collagen I is a known activator of platelets, and accounting for sets allowed us to generate a hypothesis that would have been missed by the standard gene list method.

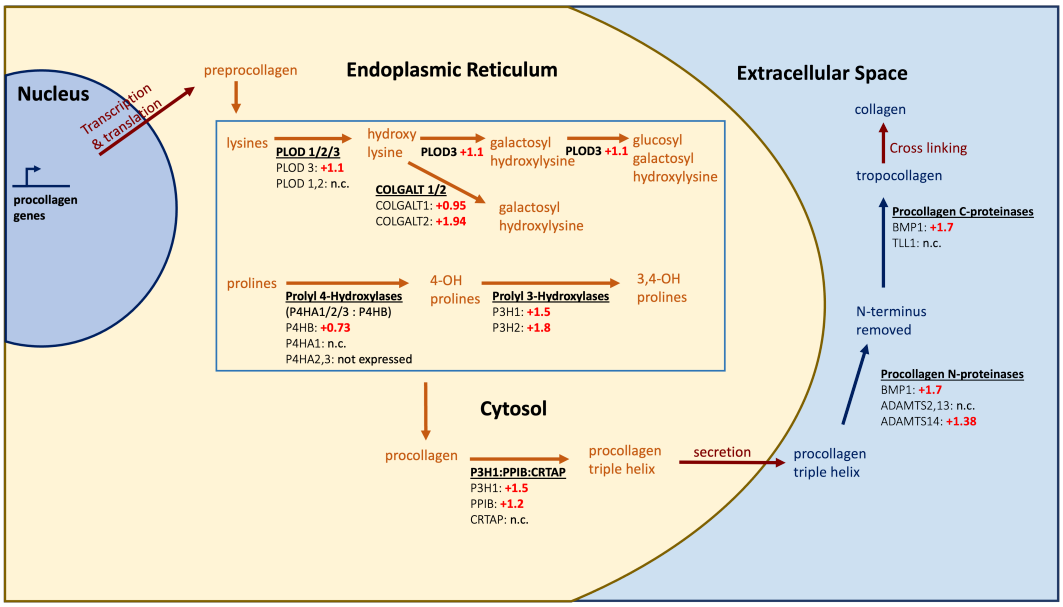


Figure 2.3: **Nearly all steps in the collagen synthetic pathway have at least one gene that is upregulated two-fold in platelets during SARS-CoV-2 infection (20).** Diagram derived from GO-CAM Collagen Biosynthesis and Modifying and edited for clarity. Log fold changes shown (20). Log2 fold change of ADAMTS14 of +1.38 is not statistically significant. N.c. stands for "no change." Sets are denoted with text titles (such as "Procollagen N-proteinases") or forward slashes (such as "P4HA123"). Complexes are denoted with colons.

In addition, we found that the pathway "Synthesis of Very Long Chain Fatty Acyl Co-A" (`http://model.geneontology.org/R-HSA-75876`) is downregulated

(6/12 step enabling entities) while this is missed by gene-list enrichment because of pathway size inflation (unadjusted p-value of 0.067 for 5/24 genes in our implementation; not reported as significant for 5/24 genes in PANTHER; with 5/51 entities in the Reactome Analysis tool due to annotation differences) (4,10). GSEA analysis utilizing the log2 fold changes yielded an FDR adjusted p-value of 0.93 (12,27). This putative downregulation is confirmed in another study of lipids in platelets during SARS-CoV-2 infection, where a specific, class-wide decrease of very long chain fatty acids was observed (see Fig 3B of (28)).

P97 inhibitors may alter PI3K signaling

Wang et al. (21) performed quantitative proteomics on cultured cell lines exposed to p97 inhibitors, proteome inhibitors, and DMSO. We analyzed their list of proteins with altered abundance after shRNA knockdown of p97 and found that the gene list method missed PI5P, PP2A and IER3 Regulate PI3K/AKT Signaling (`http://model.geneontology.org/R-HSA-6811558`), which in our step centric method had a p-value of 0.000002 with $k = 8$ and $n = 13$. Via the gene list method, the pathway has a p-value of 0.218 with $k = 8$ and $n = 115$. Comparing this to the Reactome analysis tool, the result is similar (p-value of 0.67 with $k = 7$ and $n = 118$); the change in k is an artifact of annotation. We also analyzed the same data with DAVID (7), which had an adjusted p-value of 0.9, and PANTHER (4), which did not list it among results with FDR ≤ 0.05. Analysis with GSEA did not report this specific pathway among results but had lower FDR adjusted p-values of 0.27 and 0.55 for Constitutive Signaling by Aberrant PI3K in Cancer and Negative Regulation of the PI3K AKT network. Overlay of the fold changes of these proteins on the pathway is shown in Fig. 2.4. A similar result is obtained from protein differential expression data with two p97 inhibitors also tested in the same study. While it is not apparent whether AKT signaling activity is up- or downregulated as a result, as PI3K seems to be downregulated but PIP2 concentration may be increased, the aim of enrichment analysis is hypothesis generation, and it is plausible that this signaling pathway may be meaningfully affected. While we found no literature specifically regarding PI3K/AKT signaling and p97 inhibitors, the involvement in PI3K/AKT signaling in cancers through enhancing cell survival and proliferation is well documented (29).

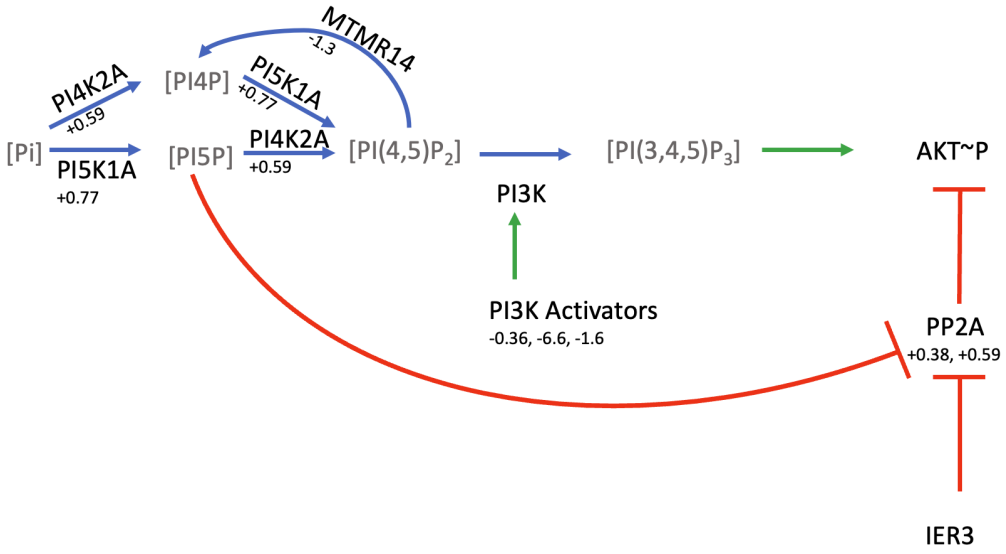


Figure 2.4: **Regulation of PI3K signaling has multiple steps with perturbed protein expression.** Pathway diagram derived from PI5P, PP2A and IER3 Regulate PI3KAKT Signaling (R-I-6811558) with positive regulatory steps shown in green, negative regulatory steps shown in red, and phosphate additionremoval shown in blue. Log fold changes in proteins as measured by label-free proteomics in p97 shRNA knockdown versus control shown in small print (21). Multiple fold changes may be listed due to multiple changes within sets or complexes.

Dilated and arrhythmogenic cardiomyopathy may involve dysregulated complement

Lastly, the Regulation of Complement Cascade pathway (`http://model.geneontology.org/R-HSA-977606`) was significantly enriched in several of the datasets we examined from Reichart et al. (2022) who performed single nucleus RNAseq on hearts from 61 patients with dilated or arrhythmogenic cardiomyopathy. We examined left ventricle arterial smooth muscle (SMC), fibroblasts (FB), and cardiomyoctes (CM) datasets from their supplemental data, and we observed that Regulation of Complement Cascade was enriched for several of the genotypes for both FBs and CMs. The genes and their roles in the complement cascade are shown below in Table 2.5.

Surprisingly, the transcriptomic changes seem to suggest downregulation of several components this pathway, which is known to be upregulated in several cardiovascular diseases. See Discussion for further comments on this result in the context of published literature. While the downregulation of complement pathway genes is surprising, its involvement in dilated cardiomyopathy is another example of a

		Fibroblasts logFC					Arterial Smooth Muscle logFC					Cardiomyocyte LogFC				
Genotype		LMNA	PKP2	RBM20	TTn	PVneg	LMNA	PKP2	RBM20	TTn	PVneg	LMNA	PKP2	RBM20	TTn	PVneg
Number of patients		12	6	8	12	8	12	6	8	12	8	12	6	8	12	8
Gene	Role															
MASP1	Activates MASP2 in lectin pathway, cleaves C3	+1.8	+2.4	+1.1	+1.7											
CFD	Rate-limiting step of alt pathway	-1.8				-1.8	-2.8	-1.3	-1.3	-2.3	-1.7	-2.8			-2.4	
C1R	Part of classical pathway	-1.3	-2.1			-1.2	-1.1		-1.5	-1.4	-1.0	-2.3			-2.4	
C3	Initiates alt pathway and is part of classical + lectin pathway	-1.5		-2.0		-1.4	-1.8	-1.5	-1.5	-2.1	-1.6	-1.9		-2.5	-2.1	
ELANE	Activates C5	-2.4	-2.9	-1.8		-2.5										
CFH	Inhibits alt pathway						+1.0		+1.2		+1.5					
C1-INH	Inhibits classical pathway	-1.1			-1.1	-1.1	-1.0				-.9	-2.2				-2.3

Figure 2.5: **Log fold changes of genes involved in regulation of the complement pathway in the left ventricle (22) correspond to increased expression of pro-complement activation genes and decreased expression of inhibitors.** Green shades indicates that the gene product promotes complement activation, while magenta corresponds to inhibition. Log2 fold changes are not shown if adjusted p-value > 0.05.

plausible hypothesis recovered by our enrichment method. The GO-CAM model for complement activation has 11 step-enabling entities but the gene list has 117 members again due to set inflation (unadjusted p-value of 0.27 for 7117 genes in fibroblasts in the LMNA genotype in our implementation; not reported as a significant result by PANTHER for 8120 genes; unadjusted p-value of 0.70 for 9156 entities in the Reactome Analysis tool) (4,10).

Gene and step enrichment analyses yield non-overlapping results

Next, we sought to understand the degree of divergence between the results across the datasets from those six papers in our analyses. Enrichment analysis does not

produce probabilities that pathways are differentially regulated, so researchers need to examine pathway results by hand to decide to investigate further. Having already shown specific examples of plausibly differentially regulated pathways, we wanted to know if the unique results from our method could simply be produced by running standard enrichment analysis with a higher false discovery rate (FDR). If the missed pathways could be recaptured with a reasonable FDR, this approach could work, but if the FDR required is too great, then the number of false positives would be too high and there would be too many results for researchers to assess. Therefore, in addition to comparing the percentages of unique results at FDR = .10, we iteratively lowered the FDR threshold for the gene list method while keeping the FDR fixed at 0.1 for our method and vice versa (Fig. 2.6). Our method generally produced higher percentages of unique results, but the standard method produced many more results in some cases, such as the platelet and p97 inhibitor datasets. At least two phenomena could contribute to this: the gene list method counts all genes while ours does not count multiple genes that belong to the same set, and the number of entities across all pathways in our database is 4000 while there are 5300 genes. This background size decrease of 25% can affect p-values by an order of magnitude or more, which can also interact with the Benjamini-Hochberg procedure and allow many results to be considered significant.

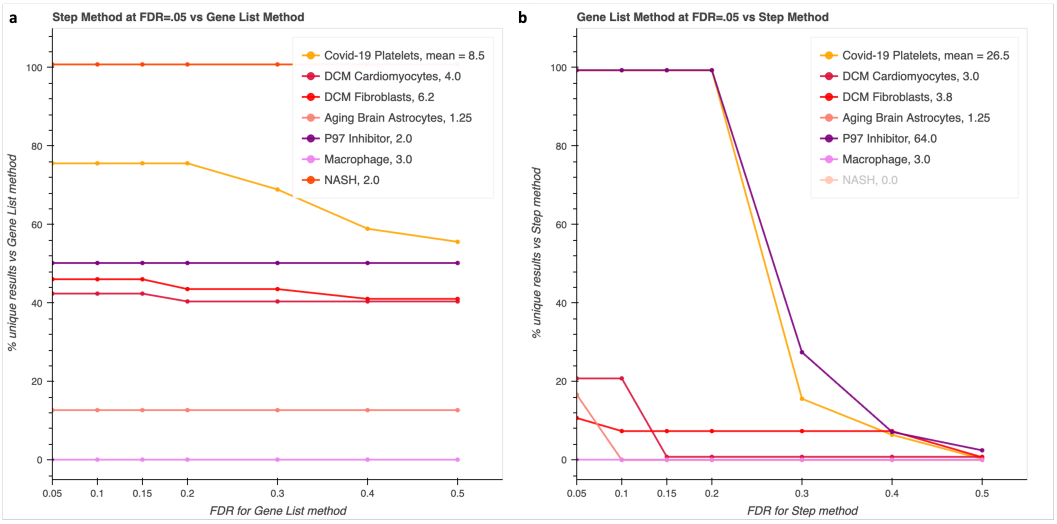


Figure 2.6: **Comparing the percentage of unique (non-overlapping) results for test datasets. a,** Step-centric enrichment shows increased sensitivity to pathway that are not recovered even with high false discovery rates **b,** while traditional, gene list enrichment shows overall increased sensitivity, likely systematic due to a larger background size that is recovered by increasing the FDR for the step-centric method.

We aggregated the counts across the datasets and saw similar proportions of unique

results versus total results at FDR = 0.1, but at FDR = 0.4, a much higher proportion of results are unique in our step-centric method (Table 2.7). Looking at the curves for the p97 inhibitor dataset and the platelet dataset, we see that most of these unique results were contributed by just those two datasets at 0.1, and there is a steep drop when the FDR is raised from 0.2 to 0.3, indicating that there was a systematic loss of sensitivity (such as due to the background size) rather than a targeted one for specific pathways (as is the case when very large sets inflate pathway sizes for gene list enrichment).

	Gene List Method at FDR =					
	0.10	0.15	0.20	0.30	0.40	Total Number of Results
Step, FDR = 0.10	53	53	50	48	44	94
	Step Method at FDR =					
	0.10	0.15	0.20	0.30	0.40	Total Number of Results
Gene List, FDR = 0.10	136	134	132	46	23	177

Figure 2.7: **Comparing the number of unique (non-overlapping) results summed across test datasets.** The traditional gene list method produces more results at FDR = 0.1 (177 vs 94), but they can be recovered by raising the FDR on the step-centric method. Raising the FDR for the gene list method does not recover unique results from the step-centric method.

The multivariate Fisher's noncentral hypergeometric distribution

In gene list enrichment, the null hypothesis specifies that each gene has an equal probability of being sampled, while the method we proposed assumes that each entity has an equal probability of being sampled, which is equivalent to assuming each step in the pathway has an equal probability of perturbation if there are no complexes. To account for the number of genes in sets, we weighted the probability of sampling sets according to the number of genes within them using the multivariate Fisher's noncentral hypergeometric distribution and the BiasedUrn package (30,31). For example, a set of five genes is assigned a weight of 5, while an entity of just one gene has a weight of 1, corresponding to the relative odds of selecting these two entities with a single draw of a gene from the genome. The hypergeometric distribution describes sampling balls from an urn (or genes from a genome) without replacement, where each ball has an equal probability of being drawn and the process continues until N balls have been drawn from the urn. The multivariate Fisher's noncentral

hypergeometric distribution is a generalization that allows different types of balls to be sampled with different probabilities. See Methods for further description.

To obtain a probability mass function for the distribution of our test statistic, k, we took advantage of its discrete nature and enumerated all possible outcomes where the number of entities sampled equals exactly k. We then computed the probability of each outcome where exactly k entities are sampled with the BiasedUrn package and sum to obtain the total probability of sampling exactly k entities. We repeated this for every valid value greater than k, and we sum to obtain a p-value. This procedure is detailed in Fig 2.8. It should be noted that sets may share members, which is discussed further in Limitations.

Alternately, we can simulate the traditional null hypothesis of independently selecting genes uniformly at random while using our proposed test statistic to avoid this problem of dependence while still including the notion of steps. See Methods for further discussion of drawbacks of simulation and our justification for using the multivariate Fisher's noncentral hypergeometric distribution.

Comparing results with Fisher's noncentral hypergeometric distribution

We next performed pairwise comparison on our test datasets between weighted enrichment done with Fisher's noncentral hypergeometric distribution and the gene list method, as well as the step-centric method (Fig. 2.9, Table 2.10). The weighted step enrichment and standard gene list methods had varying proportions of unique results, but they had generally high agreement, while the weighted step method also captured many of the enrichment results unique to the unweighted step method (compare Fig. 2.9a with Fig. 2.9d). Fig. 2.9b also shows that our weighted enrichment method recaptured many of the results that the standard gene method produced but the unweighted step method missed. 4 of 6 test datasets had less than 10% of results at FDR=0.1 that are unique to the standard method. Note that one of the remaining two was Aging Brain Astrocytes, which had a denominator of 1.25 enriched pathways. We included it for transparency, but it has little value in interpreting.

One might expect our unweighted step method's results to be supersets of our weighted method's results because all else equal, increasing the weight of sets in the pathway would increase p-values. However, p-values from the unweighted method are increased by the decreased background probability mass compared to the gene list method as the number of genes in the background exceeds the number

Parameters specific to the pathway: $\mathbf{m}, \mathbf{w}, n$
Variables determined by the experiment: $\mathbf{x}, N$

Let $\mathbf{x}$ be a selection vector representing an outcome and $\mathbf{w}$ be a weight vector. $x_i = 1$ or 0 to indicate if entity i is selected, and w_i = weight of entity i in $\mathbf{x}$ (# of genes in set). n is the number of entities in the pathway. x_{n+1} and w_{n+1} represent the background. $x_{n+1} = M - n$, and w_{n+1} = the mean weight of entities in the background.

For efficiency, $\mathbf{w}$ and $\mathbf{x}$ can be compressed to obtain $\mathbf{w}'$, $\mathbf{x}'$, and $\mathbf{m}'$: let $\mathbf{w}'$ be the sorted unique weights in $\mathbf{w}$ and each m'_i in $\mathbf{m}'$ denote the number of entities in the pathway with weight w'_i. Similarly, each x'_i in $\mathbf{x}$ denotes the number of selected entities in the pathway with weight w'_i. The $n+1^{th}$ position is excluded from this procedure, as selecting a gene from the pathway versus a gene from the background are not exchangeable outcomes.

For example, if $\mathbf{w}$ = [2,1,2,2], $\mathbf{x_a}$ = [1,1,0,0], $\mathbf{x_b}$ = [0,1,1,0] and $\mathbf{x_c}$ = [0,1,0,1] ($n + 1^{th}$ position for background not shown) have the same probability and could be instead represented with $\mathbf{w}'$ = [1,2], $\mathbf{m}'$ = [1,3], and $\mathbf{x}'$ = [1,1]. This eliminates redundant calculations and can provide over a 50-fold reduction in the number of calculations per dataset, primarily by acting on large pathways.

$\mathbf{w}'$: weight vector.
$\mathbf{m}'$: entity vector. m'_i indicates how many entities with weight w'_i are in the pathway.
$\mathbf{x}'$: selection vector. x'_i indicates how many entities with weight w'_i are selected.
S'_k: the set of all outcomes $\mathbf{x}'$ such that $\sum_{i=1}^{n'} x'_i = k$ and $0 \leq x'_i \leq m'_i \ \forall i$.

$$P(K = k) = \sum_{\mathbf{x}' \in S'_k} \text{fisher_nchgd_pmf}(\mathbf{x}', \mathbf{w}', \mathbf{m}', N)$$

$$P(K \geq k) = \sum_{j=k}^{n} \sum_{\mathbf{x}' \in S'_j} \text{fisher_nchgd_pmf}(\mathbf{x}', \mathbf{w}', \mathbf{m}', N)$$

Figure 2.8: **Computing a p-value.** We use the multivariate Fisher's noncentral hypergeometric distribution to obtain probabilities for outcomes and sum to obtain our probability mass function for the event $P(K = k)$. We then sum over $K \geq k$ to obtain a p-value. fisher_nchgd_pmf is calculated using BiasedUrn (30). Sorting and compressing the pathway vectors can greatly reduce the size of S'_k. Across our test datasets, the median reduction of the number of calls per dataset to BiasedUrn by sorting and compressing the pathway vectors is 65 fold (the median of $\frac{\sum_{\text{pathways}} \sum_{j=k}^{n} |S_j|}{\sum_{\text{pathways}} \sum_{j=k}^{n} |S'_j|}$).

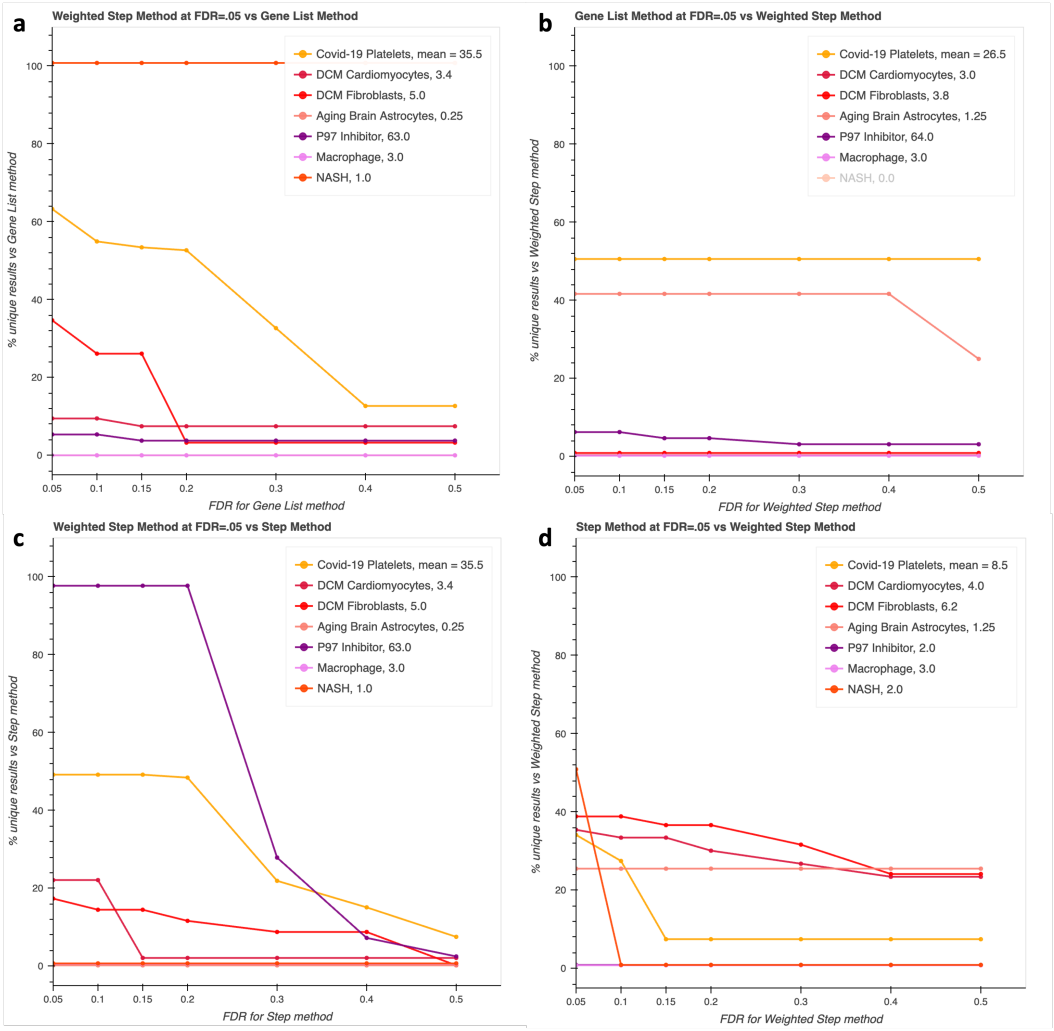


Figure 2.9: **Pairwise comparisons with the multivariate Fisher's noncentral hypergeometric distribution.** Comparison of the results of weighted enrichment and the standard gene list method shows that the weighted method provides unique results **a,** while having generally good agreement with the gene list method **b,** compared to the unweighted step method (Fig. 2.9b). **c,** Comparison between the weighted and unweighted methods shows that the weighted method produces more results than the unweighted step method **d,** while still having relatively good capture rates of results from the unweighted method.

of entities. This systematically decreases sensitivity, which is ameliorated in the weighted step method because entities in the background are weighted as well. Weighted step-centric enrichment and unweighted step-centric enrichment showed more divergence in results at typical FDRs but with gradual recapture as the FDR drops for the weighted method (Fig. 2.9; Table 2.10). Lastly, we note that of the case studies shown earlier, the collagen synthesis pathway and PI3K regulation were

	Gene List Method at FDR =					
	0.10	0.15	0.20	0.30	0.40	Total Number of Results
Weighted Step, FDR = 0.10	65	30	20	16	13	228
	Weighted Step Method at FDR =					
	0.10	0.15	0.20	0.30	0.40	Total Number of Results
Gene List, FDR = 0.10	14	12	11	10	10	177
	Step Method at FDR =					
	0.10	0.15	0.20	0.30	0.40	Total Number of Results
Weighted Step, FDR = 0.10	176	174	171	87	65	228
	Weighted Step Method at FDR =					
	0.10	0.15	0.20	0.30	0.40	Total Number of Results
Step, FDR = 0.10	42	35	34	31	27	94

Figure 2.10: **Comparing the number of unique (non-overlapping) results summed across test datasets between the gene list, step, and weighted step methods.** The gene list and weighted step method have less overall divergence than the unweighted step method and the gene list method.

still enriched with our weighted method, and the complement pathway was enriched in some genotypes with the weighted method but was past the cutoff for others.

Discussion

We introduce enrichment analysis on steps in pathways and integrate the notion of "sets," which we have not seen elsewhere in analysis tools but is commonly used in pathway databases. Our motivation is to perform an enrichment analysis that better reflects the biology of pathways and distinguishes between protein complexes and sets. We show that this can lead to significantly different results from enrichment with greater sensitivity towards pathways suffering from gene count inflation due to sets, which is critical for researchers looking to use enrichment analysis to plan follow up experiments. We do this by reanalyzing data from several papers- both characterizing the number of differential results in aggregate and discussing specific examples of pathways that are plausibly differentially active. Some of these examples have previously been shown in the literature, such as decreased synthesis of very long chain fatty acids by platelets during SARS-CoV-2 infection, and some are novel, such as collagen production by platelets during SARS-CoV-2 infection. We then utilize Fisher's noncentral hypergeometric distribution to weight sets according

to the number of genes within them to partially preserve an assumption in the null hypothesis of gene list enrichment analysis.

We compare the results obtained with the hypergeometric distribution between our step-centric method that uses a list of step-enabling entities and the standard method using a gene list, and we discover that up to half or more of results from each method with FDR = .1 can be unique. Furthermore, increasing the FDR for the opposing method only partially recaptures those non-overlapping results. Note that we did not evaluate enrichment results only produced by the standard gene-list method, because it is already widely used and accepted. We believe that results unique to standard enrichment analysis are not necessarily irrelevant, as there may be cases where many members of a set are upregulated in a biologically meaningful manner, which our method may miss. Also, the larger background size of the traditional gene-list method systematically decreases p-values, which could lead to reporting many more results as significant, as for the P97 dataset when comparing our unweighted step enrichment procedure with the traditional gene-list method in Fig. 2.6b. Likewise, we found that the results of our case study pathways were not significantly enriched in the popular GSEA algorithm, but we did not evaluate the credibility of results unique to GSEA as it is also well accepted (12).

One-way, pairwise comparisons with weighted enrichment analysis on steps using the multivariate Fisher's noncentral hypergeometric distribution showed it had the least divergence from the standard method but still provided unique results. Unique results from this method could be evaluated alongside results from standard gene-list enrichment. However, it unclear if weighting is always necessary. We use enrichment to generate viable hypotheses that can be evaluated based on literature and investigated further as opposed to providing accurate p-values with which to reject a null hypothesis that we already know is incorrect, as cells do not select genes independently with uniform probability due to intricacies of gene regulation that are unrelated to regulation of pathway activities. More importantly, sets sometimes indicate tissue-specific expression, in which case neither weighting nor the standard gene list method is biologically meaningful, because cells of that type do not actually have access to sampling all members of the set. Even when sets do not indicate gene-expression switches between tissues, weighing sets by their gene content might not be appropriate; the collagen pathway was primarily inflated by a set corresponding to 44 procollagen genes, yet these encode different types of collagen proteins and only two need to be expressed, let alone differentially expressed, to produce collagen.

What is necessary, however, is deciding whether this method is appropriate for a given dataset. Our core idea is that an individual cell (or population of cells employing the same regulatory strategy to respond to a condition) is unlikely to need to significantly perturb expression of every known member of a set in order to regulate a step in a pathway. Some members may not be expressed and therefore cannot be downregulated, are lowly expressed for other functions but are not relied upon by that cell type for that pathway, or are not "accessible" to upregulate because that cell does not express those genes under any normal conditions. Given this idea, we believe it is appropriate to use our method on single cell data and bulk data corresponding to a single cell type, while using our method on a heterogeneous source (i.e. mutation data pooled from many patients' tumors) could lead to spurious results. It may be possible to use our method on bulk-RNA seq from a whole tissue, but care should be taken to interpret each result based on whether most cells in the sample likely employed the same regulatory strategy for the pathways of interest. If a dataset meets the criteria above, we believe using our method alongside standard enrichment analysis is beneficial, but if not, we recommend the standard method.

Assuming a given dataset is an appropriate choice for our enrichment methods to do a comparison, p-values can differ by orders of magnitude between the various algorithms, but even within the same method, p-values are sensitive to relatively small changes in background size and changes to cutoffs for generating differentially expressed gene lists from the experiments themselves. As mentioned above, the p-values themselves are not the focus of enrichment analysis tools. Enrichment analysis aims to provide researchers with hypotheses to follow up on based on the results of their experiments. As a hypothesis generating tool, ideal enrichment analysis would have a high sensitivity and enable us to uncover novel biology. So long as there are not so many results that it becomes impractical for a researcher to consider each one critically when deciding to investigate further, it is perhaps better to err on the side of false positives than false negatives. This bias motivates us to formulate an enrichment question more representative of the biology at hand.

Discussion of pathways in case studies

Collagen I production by platelets

Collagen is known to contribute to platelet aggregation and clot formation, but platelets have not historically been considered sources of collagen. However, there is recent support in the literature for this possibility. Collagen modifying enzymes have been shown to be expressed in platelets, although their activity against mature

collagen triple helices in the extracellular space was not observed, suggesting that platelets do not modify extracellular collagen triple helices produced by other cells (32). Recently, collagen I synthesis in activated platelets has been observed (33) in platelets derived from patients with polycystic ovarian syndrome (PCOS), which is known to be associated with increased platelet aggregation. These platelets were shown to contain higher levels of hydroxyproline than platelets from healthy donors, COL1A1 mRNA was detected, and collagen was detected at higher levels in platelets from PCOS patients. To demonstrate that the detected collagen originated from within the platelets themselves as opposed to contamination, the authors showed that collagen levels increased when platelets were activated with thrombin and coated the surfaces of the platelets, while in the absence of activation, collagen was detected only after membrane permeabilization with Triton X-100. They note that procollagen mRNA was not significantly different between the platelets of healthy donors and the PCOS patients, yet collagen levels were different, speculating that this may be due to increased activity of enzymes in the collagen synthesis and modification pathway. Manne et al. (20) did not assay collagen levels, and mRNA levels were not increased for collagen I, yet mRNA levels for enzymes at every step of the pathway were increased, consistent with the findings of Kyselova et al. (33) that collagen levels and markers of collagen production were increased despite no increase in collagen mRNA.

Dilated and Arrhythmogenic Cardiomyopathy

The role of complement in dilated cardiomyopathy is unclear, but complement activation has been implicated in various cardiovascular and cardiac diseases, and complement activation has been shown specifically to play a role in congestive heart failure (34,35). Endogenous production of complement components following myocardial infarction has been shown (36), and C3 and CFD have been shown to play a role in cardiac remodeling in right heart failure in mice, where C3 and CFD deletion ameliorated heart failure (37). It should be noted, however, that the remodeling in right heart failure due to pulmonary hypertension typically results in hypertrophy, not dilation. Curiously, the changes in gene expression in Reichart et al. (22) all suggest downregulation of the complement cascade at the initial steps of the alternative (CFD, CFH, C3), classical (C1R), and lectin (MASP1) pathways as well as downstream (C3, ELANE), except for increased MASP1 in fibroblasts and consistently decreased C1 Inhibitor in the LMNA and Pathogenic Variant negative (PVneg) genotype for all three cell types. While survivorship

bias is a possibility (i.e. cells that did not downregulate complement activation were destroyed), complement has roles beyond cell lysis; intracellular complement activation and local complement secretion by cells is beginning to be understood as having important roles for regulating cell activities, including in endothelial cells and fibroblasts (38). Ito et al. (37), who studied RV heart failure, also observed that complement system genes are increased in healthy controls in the right versus left ventricle, and deletion of C3 did not ameliorate LV failure due to aortic constriction in mice (37), raising the possibility that complement pathway plays different roles in the left and right ventricles. Its downregulation may play a protective role in dilated cardiomyopathy, explaining why C3 deletion had no effect on LV in Ito et al., or it might be involved in the pathogenesis of dilated cardiomyopathy through dysregulation of currently unknown cellular functions. Proteomics of plasma have shown that decreased C3 levels are associated with poorer survival in congestive heart failure (39). Alternatively, it is possible that complement genes are downregulated due to medications; however, primary treatment for DCM includes angiotensin-converting enzyme inhibitors and beta blockers, neither of which have been shown to decrease complement.

Limitations

Sets count once towards overlap In the hypergeometric test (or Fisher's exact test), the actual fold change of a gene is not accounted for. Genes are added to the list if they pass some cutoff criteria. In keeping with that idea, we treat selection of two members of a set the same as selection of one member. If their activities contribute additively towards the total activity of the step, then this is analogous. However, there are examples of interactive behavior between sets, even a decrease in activity when multiple members of a set are upregulated. In BMP signaling, there are many ligands that can interact with many receptors, and in some cases, increased concentration of multiple ligands can actually decrease BMP signaling due to binding differences in affinities and reaction kinetics (19). Ideally, the interactions between members of sets would be characterized and accounted for. Furthermore, sets sometimes denote genes with tissue-specific expression, as in glycolysis. Not all members of a set are available for given cell types to up or downregulate, so upregulation of just one or two members of the set might actually constitute upregulation of all members of the set that are actually expressed in that cell. Additionally, in the case of downregulation in response to a treatment, only one member of a set may be expressed prior to treatment, and again, not all members of the set are available to

be selected for regulation. We currently do not distinguish between these cases and err on the side of leniency.

Positive dependence in the weighted distribution We note that this model suffers from the problem of overlaps between sets. This creates a relatively weak dependence between sets and a strong correlation between genes as solo entities and any sets that contain them. This violates the assumption of independence in the distribution. Typically, genes will not appear in more than one step within a single pathway; they are more likely to appear in multiple entities across pathways due to artifacts of annotation or genuine biology, but they are in the background relative to the pathway being tested. In some cases, some enzymes can catalyze multiple steps in pathways, and it is possible to "double count" genes in this manner if an enzyme can uniquely catalyze one step but is one of several in another step, as seen in the collagen synthesis example.

Methods

The multivariate Fisher's noncentral hypergeometric distribution

Fisher's distribution is more appropriate than Wallenius's for enrichment

Fisher's noncentral hypergeometric distribution and Wallenius's noncentral hypergeometric distribution both describe sampling processes where the balls are not drawn with equal probability; they are weighted (30,31). However, Fisher modeled a process in which all balls were sampled independently, and the total number of balls drawn is itself a random variable (i.e., all at once), while Wallenius modeled a process in which balls were sampled one by one until the desired number of balls were drawn. In that scenario, the order of drawing the balls matters- drawing a ball with very high weight first and removing it from the urn increases the chance of drawing the remaining balls, whereas drawing a ball with low weight first does not substantially affect the probability of drawing the heavy ball next. To maintain independence, Fisher's distribution treats the number of balls drawn in the experiment as a random variable itself, and then once the experiment has been performed and N is known, the multivariate distribution of balls drawn of each type can be conditioned on N. This aligns with the process of differential expression experiments, where we do not know the number of differentially expressed genes prior to the experiment, and there is no reason why a cell would only perturb a fixed number of genes. Therefore, we use Fisher's noncentral hypergeometric distribution to model outcomes of the experiment (specific arrangements of gene perturbations) and sum the probabilities of those outcomes to obtain p-values for our events (the number of

perturbed genes in the pathway).

Fisher's distribution over simulating the traditional null hypothesis with a new test statistic

Let k be the number of entities with at least one gene member selected. A cell randomly selects genes independently with uniform probability. N genes are selected without replacement from M total. Assuming no entities share overlapping genes within a pathway, this shares the same generative process as the traditional, gene-list method, but it has a different test statistic. Repeated sampling of an entity does not count towards k. Assuming a gene-first generative model, simulating uniform probability, independent gene selection will give accurate p-values with respect to that null hypothesis without problems of dependence. However, this suffers from two drawbacks. First, we believe it is more biologically relevant to have a generative model where a cell selects steps to regulate without replacement and the genes merely reflect that. Second, simulation is slow, as multiple testing correction necessitates that we have accurate estimates of p-values several orders of magnitude smaller than 0.05. While this can be pre-computed for many values of N and then interpolated, it would need to be recomputed for custom backgrounds that a user may specify for their experiment, and simulation on the fly would be intractable.

Representing the background to select from

In the Results, we discuss k, the overlap and detail our procedure for calculating p-values with weighted entities in Fig. 2.8. Here, we discuss the parameters M and N for the hypergeometric distribution and how they factor into our use of the multivariate Fisher's noncentral hypergeometric distribution. There is no M for this distribution, although it exists implicitly. The multivariate Fisher's noncentral hypergeometric distribution is parameterized by m, w, x, and N. Let there be c colors of balls in the urn. m and w are vectors of length c, respectively, where m_i is the number of balls of color i and w_i denotes the weight for balls of color i. x corresponds to an outcome in the sample space. It is a vector of observed counts for the sample, and x_i is the number of balls sampled of color i. N is the total number of balls sampled. In our problem, let there be n entities in the pathway. We can group them according to their weights to produce a vector m for our pathway, where m_i indicates the number of entities with weight w_i, which is either 1 for proteins or is the cardinality of the set represented by element i. $\mathbf{x}$ takes the same form, and $\sum_{i=1}^{len(\mathbf{x})} x_i = k$ is the number of entities selected from the pathway in the outcome

represented by $\mathbf{x}$. Each element x_i is subject to the constraint $x_i \leq m_i$ (i.e. if there are only 3 sets in the pathway with 5 genes in them, we cannot observe an event $\mathbf{x}$ where we have 4 entities of weight 5). However, these vectors only describe the pathway and lack a description for the background against which entities are selected, previously represented by M. We added an $(n+1)^{th}$ element to each vector to represent the entities outside of the pathway, where$w_(n+1)$ is the mean weight for the remaining entities in the background , and $m_{(n+1)} = M - n$. An exact representation would add more than 1 element to represent the pathway; after constructing $\mathbf{m}_{\text{pathway}}$ and $\mathbf{w}_{\text{pathway}}$, the same procedure could be used to construct $\mathbf{m}_{\text{background}}$ and $\mathbf{w}_{\text{background}}$ vectors for the background (representing each entity not in the pathway and their individual weights), and $\mathbf{x}_{\text{background}}$ would be constrained by summing to $N - k$. We did not do this because representing every entity in the background produces a much larger sample space and would be not computationally feasible. Instead, we approximated the background with entities of the mean weight (rounding to 2 decimal places) to preserve the odds of entities within the pathway vs not in the pathway. Approximating the background at a less coarse resolution with two groups of entities by sorting the background entities by weight, splitting at the median into upper and lower halves, and approximating each group with its mean weight incurred a two order of magnitude increase in runtime while making a negligible difference in p-values. For the nonalcoholic steatohepatitis dataset (24), the two significantly enriched pathways were the same, and their p-values decreased very slightly from 2.6810^{-7} to 2.7110^{-7} and 1.2910^{-4} to 1.3010^{-4}; for the platelets dataset (20), the 67 enriched pathways were the same and their p-values were reduced by a median of 11% with 2.5^{th} and 97.5^{th} percentile reductions of 1% and 24%). Therefore, we approximated the entire background as one group with its mean weight. Implementation was done with python v3.10.9, R v4.2.2, rpy-2 v3.5.9, SciPy v1.9.3, and BiasedUrn v2.0.10. A full description of the Conda environment is available at our GitHub repository in environment.yml.

Pathway data collection GO-CAM models used in this study were from the 2022-07-01 release (`doi:10.5281/zenodo.6799722`). All GO-CAM models in production state were filtered to select those with at least three molecular functions (steps) linked by two causal edges (142 pathway models). We did this because these GO-CAMs are more likely to represent a process or pathway corresponding to a GO Biological Process than other, partial models, which may reflect curation practice for entering annotations into the GO. Then, GO-CAM models imported from Reactome, which are not yet officially released in GO-CAM form, were selected

as well (1087 pathway models). These were generated from Reactome using the methods described in Good et al. (16). As a note, GO curators do not currently curate sets, but they are allowed in the GO-CAMs that are created by direct conversion from Reactome. Entities enabling the molecular functions in these models were extracted using the "enabled by" relation (RO:0002333). All pathways were annotated as human or mouse; genes in mouse pathways were mapped to human orthologs using SimpleMine (v5.0.0) at `http://mangolassi.caltech.edu/~azurebrd/cgi-bin/forms/agr_simplemine.cgi`. Reactome entities were mapped to UniProt accession IDs using the Reacto ontology (v1.0) at `http://www.geneontology.org/ontology/extensions/reacto.owl` and all protein members of complexes and sets were retained while ions and non-protein chemical entities were discarded. Some complexes and sets contain subsets and/or subcomplexes within them. Generally, these can be recursively simplified by taking complexes to mean an AND joining of their members and sets to entail an OR relationship (Box 2.2). 100 random examples of Reactome entities were checked manually after recursive conversion to entities with UniProt accession numbers. 2 were represented logically inconsistently and failed in the following way: let Set 1 be a set of 3 complexes C1, C2, and C3. C1 = P1 & Pa; C2 = P1 & Pb; C3 = P1 & Pc & Pd. Then S1 = P1 & Pa or Pb or Pc or Pd. An input of [P1, Pc] would count as the entirety of S1 being perturbed.

Test data sources

Gene lists were filtered for adjusted p-value < 0.05 and log fold change > 1 or < -1 unless done otherwise in their respective papers. Gene lists were split by up and downregulated genes for Manne et al. (20) and Boisvert et al. (23) and by genotype in Reichart et al. (22) and brain region in Boisvert et al. (23).

Enrichment with PANTHER, REACTOME, GSEA, and DAVID Differentially expressed genes or protein were analyzed with PANTHER's Gene List Analysis tool by selecting "statistical overrepresentation test," "Reactome pathways," "Homo sapiens genes," "Fisher's Exact," and "Calculate False Discovery Rate." The same list of proteins were queried at DAVID by selecting "Functional Annotation Tool," and "REACTOME_PATHWAY" and at Reactome by selecting "Analysis Tools" on the home page and the default "Project to Human" option on step 2. GSEAPreranked was performed with GenePattern using the reported log fold changes, "c2.cp.reactome.v2023.2.Hs.symbols.gmt" for the gene sets database, the default 1000 permutations, and "No_Collapse" (12,27). GSEA was not performed on the

Dataset	Source	Paper
Colon cancer cell line response to p97 shRNA	Supp 1: shRNA DE	Wang et al., 2022
Platelets in Covid-19	Supp 3	Manne et al., 2020
Left Ventricle Cardiomyocytes	Supp 7, 8: all genotypes and PVneg	Reichart et al., 2022
Left Ventricle Fibroblasts	Supp 16, 17: all genotypes and PVneg	Reichart et al., 2022
Arterial Smooth Muscle	Supp 23, 24: all genotypes and PV neg	Reichart et al., 2022
Astrocytes in Aging Mouse Brain	Supp Table 5, CB and HTH	Boisvert et al., 2018
Macrophage activation in vitro	Table S1	Orecchioni et al., 2021
Nonalcoholic steatohepatitis	Table S2	Govaere et al., 2023

Table 2.2: **Test datasets.** Gene lists were filtered for adjusted p-value < 0.05 and log fold change > 1 or < -1 unless done otherwise in their respective papers. Gene lists were split by up and downregulated genes for Manne et al. (20) and Boisvert et al. (23) and by genotype in Reichart et al. (22) and brain region in Boisvert et al. (23).

dilated and arrythmogenic cardiomyopathy dataset as Reichart et al., reported log fold changes only for significantly differentially expressed genes (22).

Acknowledgments

The authors thank Mark Zhang for writing feedback as well as Chris Mungall and his group in the Gene Ontology consortium for their support. This work was supported by the National Human Genome Research Institute (U24HG012212).

Data availability statement

All data and code used is available on the GitHub repository `https://github.com/nmarkari/gocam_enrichment`. We have also used Zenodo to assign a DOI to the repository: 10.5281/zenodo.8310236 (`https://zenodo.org/records/8310236`) . Primary sources for datasets used in testing are listed in Table 2.2, and their corresponding csv files after filtering are in our GitHub at `https://github.com/nmarkari/gocam_enrichment/tree/main/test_data/processed`.

References

1. BRAIN Initiative Cell Census Network (BICCN), BRAIN Initiative Cell Census Network (BICCN) Corresponding authors, Callaway EM, Dong HW, Ecker JR, Hawrylycz MJ, et al. A multimodal cell census and atlas of the mammalian primary motor cortex. Nature. 2021 Oct 7;598(7879):86–102.

2. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. eLife. 2017 Dec 5;6:e27041.

3. Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun. 2018 Apr 10;9(1):1366.

4. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019 Mar;14(3):703–21.

5. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. Protein Sci. 2022 Jan;31(1):8–22.

6. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan;37(1):1–13.

7. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022 Jul 5;50(W1):W216–21.

8. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007 Jul;35(suppl_2):W193–200.

9. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder — open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004 Dec 12;20(18):3710–5.

10. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022 Jan 7;50(D1):D687–92.

11. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. 1999 Sep;

12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct 25;102(43):15545–50.

13. Glaab E, Baudot A, Krasnogor N, Schneider R, Valencia A. EnrichNet: network-based gene set enrichment analysis. Bioinformatics. 2012 Sep 15;28(18):i451–7.

14. Thomas PD, Hill DP, Mi H, Osumi-Sutherland D, Van Auken K, Carbon S, et al. Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. Nat Genet. 2019 Oct;51(10):1429–33.

15. Ebert D, Feuermann M, Gaudet P, Harris NL, Hill DP, Lee R, et al. The Gene Ontology knowledgebase in 2023.

16. Good BM, Van Auken K, Hill DP, Mi H, Carbon S, Balhoff JP, et al. Reactome and the Gene Ontology: digital convergence of data resources. Lu Z, editor. Bioinformatics. 2021 Oct 11;37(19):3343–8.

17. Nanda S, Jacques M, Wang W, Myers CL, Yilmaz LS, Walhout AJ. Systems-level transcriptional regulation of Caenorhabditis elegans metabolism. Mol Syst Biol. 2023 May 9;19(5):e11443.

18. Lopes D, Maiato H. The Tubulin Code in Mitosis and Cancer. Cells. 2020 Oct 26;9(11):2356.

19. Antebi YE, Linton JM, Klumpe H, Bintu B, Gong M, Su C, et al. Combinatorial Signal Perception in the BMP Pathway. Cell. 2017 Sep;170(6):1184-1196.e24.

20. Manne BK, Denorme F, Middleton EA, Portier I, Rowley JW, Stubben C, et al. Platelet gene expression and function in patients with COVID-19. Blood. 2020 Sep 10;136(11):1317–29.

21. Wang F, Li S, Houerbi N, Chou TF. Temporal proteomics reveal specific cell cycle oncoprotein downregulation by p97/VCP inhibition. Cell Chem Biol. 2022 Mar;29(3):517-529.e5.

22. Reichart D, Lindberg EL, Maatz H, Miranda AMA, Viveiros A, Shvetsov N, et al. Pathogenic variants damage cell composition and single cell transcription in

cardiomyopathies. Science. 2022 Aug 5;377(6606):eabo1984.

23. Boisvert MM, Erikson GA, Shokhirev MN, Allen NJ. The Aging Astrocyte Transcriptome from Multiple Regions of the Mouse Brain. Cell Rep. 2018 Jan;22(1):269–85.

24. Govaere O, Hasoon M, Alexander L, Cockell S, Tiniakos D, Ekstedt M, et al. A proteo-transcriptomic map of non-alcoholic fatty liver disease signatures. Nat Metab. 2023 Apr 10;5(4):572–8.

25. Orecchioni M, Ghosheh Y, Pramod AB, Ley K. Macrophage Polarization: Different Gene Signatures in M1(LPS+) vs. Classically and M2(LPS–) vs. Alternatively Activated Macrophages. Front Immunol. 2019 May 24;10:1084.

26. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics. 2007 Feb 15;23(4):401–7.

27. Kuehn H, Liberzon A, Reich M, Mesirov JP. Using GenePattern for Gene Expression Analysis. Curr Protoc Bioinforma [Internet]. 2008 Jun [cited 2023 Nov 13];22(1). Available from: https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/0471250953.bi0712s22

28. Schuurman AR, Léopold V, Pereverzeva L, Chouchane O, Reijnders TDY, Brabander JD, et al. The Platelet Lipidome Is Altered in Patients with COVID-19 and Correlates with Platelet Reactivity. Thromb Haemost. 2022 Oct;122(10):1683–92.

29. Hemmings BA, Restuccia DF. PI3K-PKB/Akt Pathway. Cold Spring Harb Perspect Biol. 2012 Sep 1;4(9):a011189–a011189. 30. Fog A. Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. Commun Stat - Simul Comput. 2008 Feb;37(2):241–57.

31. Harkness WL. Properties of the Extended Hypergeometric Distribution. Ann Math Stat. 1965 Jun;36(3):938–45.

32. Anttinen H, Tuderman L, Oikarinen A, Kivirikko K. Intracellular enzymes of collagen biosynthesis in human platelets. Blood. 1977 Jul 1;50(1):29–37.

33. Kyselova A, Zukunft S, Puppe D, Wittig I, Mann WA, Dornauf I, et al. Human platelets are a source of collagen I. Haematologica. 2021 Mar;106(3):899–902.

34. Lappegård KT, Garred P, Jonasson L, Espevik T, Aukrust P, Yndestad A, et al. A vital role for complement in heart disease. Mol Immunol. 2014 Oct;61(2):126–34.

35. Aukrust P, Gullestad L, Lappegård KT, Ueland T, Aass H, Wikeby L, et al. Complement Activation in Patients With Congestive Heart Failure. 2001;

36. Yasojima K, Schwab C, McGeer EG, McGeer PL. Human Heart Generates Complement Proteins That Are Upregulated and Activated After Myocardial Infarction. Circ Res. 1998 Oct 19;83(8):860–9.

37. Ito S, Hashimoto H, Yamakawa H, Kusumoto D, Akiba Y, Nakamura T, et al. The complement C3-complement factor D-C3a receptor signalling axis regulates cardiac remodelling in right ventricular failure. Nat Commun. 2022 Sep 15;13(1):5409.

38. West EE, Kemper C. Complosome — the intracellular complement system. Nat Rev Nephrol. 2023 Jul;19(7):426–39.

39. Cuvelliez M, Vandewalle V, Brunin M, Beseme O, Hulot A, De Groote P, et al. Circulating proteomic signature of early death in heart failure patients with reduced ejection fraction. Sci Rep. 2019 Dec 16;9(1):19202.

Chapter 3

K-SPACES: MIXTURES OF GAUSSIAN LATENT VARIABLE MODELS

Abstract

Principal component analysis (PCA) and *k*-means clustering are two seemingly different methods for dimension reduction and clustering, respectively, but can be understood as special cases of inference in a Gaussian latent variable model framework. We leverage this insight to develop a probabilistic framework and methods for simultaneous dimension reduction, clustering, and latent space learning that are efficient and interpretable, and that can replace current ad hoc combinations of PCA and clustering. The algorithm, *k*-spaces, has broad applicability, which we demonstrate in several distinct genomic settings. In particular, we show how *k*-spaces can be used to model gene expression in quantitative hybridization chain reaction (qHCR) images, for inference in epigenomics, and for dimension reduction of single-cell RNA-sequencing data.

Introduction

Dimension reduction is an essential step in the analysis of large tabular data for exploration and visualization of patterns in the data. Two classical and useful dimension reduction techniques for matrices are principal component analysis (PCA) and *k*-means clustering, which reduce the dimensions of features and partition observations, respectively. Although seemingly unrelated methods, both PCA and *k*-means can be viewed as special cases of an algorithm that iteratively clusters samples while reducing the dimension of the feature space by fitting a set of affine subspaces (i.e., points, lines, planes, etc.) to the features in order to capture distinct linear relationships within subpopulations in the data.

PCA may be derived as a limiting case, where the variance approaches zero, of a probabilistic homoskedastic latent Gaussian factor model called probabilistic PCA (pPCA) (Michael E. Tipping and Christopher M. Bishop, 1999b; Roweis, 1998). This formulation leads to an interpretation of PCA as the optimal inference procedure (in the limit) to find a latent space for points perturbed by isotropic Gaussian noise. From this point of view, zero-dimensional pPCA fits a single isotropic Gaussian to the data. In a Gaussian mixture model framework, it is therefore natural to consider *k*

mixtures of Gaussian latent variable models, each possibly of a different dimension, with isotropic noise in their complementary spaces. We derive an algorithm to perform inference for such models that contains Gaussian mixture models and pPCA as special cases, and we call this framework k-spaces. These specific models were considered in earlier work (Michael E. Tipping and Christopher M. Bishop, 1999b; Roweis, 1998; Geoffrey E Hinton, Revow, and Dayan, 1995; Ghahramani and Geoffrey E. Hinton, 1996; Su and Dy, 2004); in particular, Tipping and Bishop explored the extension of their pPCA to mixture models for piecewise local linear dimension reduction of nonlinear manifolds (Michael E Tipping and Christopher M Bishop, 1999a), but their model has not been widely explored or deployed for clustering to the best of our knowledge. We develop a framework for fitting subspaces of mixed dimension and shared noise parameters as well as performing model selection, which makes k-spaces suitable for a variety of biological settings where it is useful to simultaneously cluster and reduce dimension.

Our work was motivated by several problems in biology that require fitting multiple linear latent spaces to data. One such problem is detecting distinct gene coexpression relationships within different cell types or states. We use k-spaces to analyze a quantitative hybridization chain reaction (qHCR) image, where our results and k-spaces's behavior can be interpreted in an anatomical and developmental context. qHCR imaging overcomes multiple shortcomings of traditional enzyme-based imaging methods to enable multiplex, quantitative, high-resolution RNA imaging, yielding subcellular voxel intensities that scale linearly with target abundance (Trivedi et al., 2018; Choi et al., 2018; Schulte, Fornace, et al., 2024). In prior work (Trivedi et al., 2018), a qHCR image staining the expression levels of four mRNAs was used to identify the spatial organization of distinct circuit states active during zebrafish somitogenesis. In that study, circuit states for regions of interest (ROIs) were identified by manual clustering of their gene expression vectors across several pairwise scatter plots and then analyzed in spatial context by shading voxels of the image within ROIs by the identified circuit states. Here, we use k-spaces to process the entire image and automatically fit k lines to the entire dataset in the full 4-D voxel intensity space. We show that this automated approach to the analysis technique introduced by that study both recapitulates and extends earlier findings (Trivedi et al., 2018). By analyzing all cells in the image at once, we identify spatial usage of distinct gene circuit states and their spatial gradients of expression scaling.

Our k-spaces method has immediate application to diverse problem types, as we

demonstrate with additional examples from epigenomics and RNA sequencing. The epigenomics application, namely detection of differentially methylated sites between underlying cell types in bulk data, challenges k-spaces through the dimension of the clusters. Clustering on features rather than observations of the methylation matrix, we project uncorrelated and correlated CpG sites (i.e., the sites of methylation) onto separate 0- and 4-dimensional linear manifolds. In particular, sub-populations in the data are not linearly separable but are separable by the dimensionality of their subspaces, and our model lends itself to tractable theoretical analysis. Similarly, we show that k-spaces is a powerful tool for the analysis of high-dimensional single-cell RNA-sequencing (scRNA-seq) data by virtue of unifying the dimension reduction and clustering steps, allowing us to visualize the distinct yet partially synchronized developmental trajectories of *C. elegans* hypodermis and seam cells across a series of linear dimension reductions.

Results

k-spaces overview

The k-spaces algorithm uses Gaussian latent variables and an expectation maximization (EM) procedure to learn a set of k latent spaces with dimension $d_1, d_2, ..., d_k$. Observed data points are assumed to be generated by a process with isotropic Gaussian noise in the complementary space, with either a distinct variance parameter for noise for each latent space or one that is shared for the entire model. Observations are iteratively assigned to subspaces (grouping rows of the data matrix) and an optimal linear dimension reduction is computed to represent each submatrix (reducing columns of the matrix; Fig. 3.1). k-spaces produces affine subspaces that cluster the observations along with projections of the clusters onto their subspaces. Depending on the application, different aspects of the k-spaces method may be of interest. For example, we show that the clustering of epigenomics data produces a classification of sites. In the gene coexpression analysis, the affine subspaces provide insights into gene expression scaling in qHCR. In single-cell RNA-sequencing data analyses, the latent spaces reveal complex structure in qHCR and single-cell relationships.

A self-contained description of our model and derivations of the inference procedures are provided in the Methods and Supplement.

k-spaces clustering applied to qHCR data identifies distinct gene circuit states and their oscillating expression levels in the developing zebrafish embryo

We first evaluated the behavior of k-spaces on systems data reflecting heterogeneous cell types and gene circuit states using spatial embryogenesis data produced with qHCR. The read-out/read-in analysis framework we build upon enables bi-directional quantitative discovery (Trivedi et al., 2018): read-out from anatomical space to expression space to discover co-expression relationships; read-in from expression space to anatomical space to discover the anatomical regions in which selected gene co-expression relationships occur. Trivedi and coworkers did the following: in a *read-out* step, to identify circuit states for regions of interest (ROI) within the image, voxel intensities were displayed as six 2-D scatter plots (one plot for each pairwise combination of the four mRNAs assayed in their 4-plex qHCR image) and gene expression clusters of interest were identified manually. Second, in a *read-in* step, to identify the anatomical location of each circuit state within the embryo, the voxels within the image were shaded by their expression cluster index. Here, we revisit their analysis of four mRNAs active during zebrafish somitogenesis.

We started with the 4-plex qHCR image (Fig. 3.2a). Following normalization of the signal for each channel, we used k-spaces to read-out all 17,340 subcellular (2×2×6 μm) voxels to 4-D expression scatter plots and fit 9 lines to the voxel intensities directly within the 4-D space (Fig. 3.2b); each line is an automatically generated expression cluster corresponding to a circuit state with different mRNA stoichiometry. We then used k-spaces to read these expression clusters back into the anatomical context of the image, coloring voxels according to cluster index to automatically segment the anatomy based on circuit state (Fig. 3.2c).

Reconstruction of anatomical features based on circuit state provides a powerful method to study the spatial usage of gene expression programs. Reconstruction also enables assessment of under-/over-clustering for different values of k based on visual inspection of expression space (Fig. 3.6) and anatomical space (Fig. 3.7). We varied k from 1 to 12 and determined that $k = 9$ was most appropriate for our data (Fig. 3.6a). Two of the nine clusters (blue and orange in Fig. 3.2c) correspond to voxels outside the embryo or voxels with a low basal expression level of the four mRNAs, leaving seven expression clusters to represent somitogenesis circuit states. Anatomical features emerge as k increases (Fig. 3.7a). The first feature identified was non-sample versus sample ($k = 2$). As k increases, the notochord, somites (S7 - S10), presomitic mesoderm (PSM), and adaxial cells are identifiable in the

segmented image; eventually these anatomical structures become subdivided (Fig. 3.7a). Lines fitting multiple clusters (i.e., bimodality) in the scatter plots indicated underfitting, and overfitting was apparent in the image when voxels were partitioned on subcellular spatial scales (Figs. 3.6a, 3.7a).

Beyond segmenting the image into discrete circuit states, we use voxel saturation to depict the relative position of a voxel's projection onto its assigned subspace (Fig. 3.2d), enabling visualization of co-expression changes produced by varying developmental programs in the embryo. Notably, this representation reveals that the red program (the mRNA stoichiometry in the expression cluster for which voxels are shaded red in Figs. 3.2b-d) is turning on with increasing strength as somites mature (increasing maturity S9→S8→S7; Fig. 3.2d).

Automated cluster identification and image segmentation of the whole image with k-spaces replicated and earlier extended findings (Trivedi et al., 2018) (Fig. 3.2c; see Fig. 3.8 for a side-by-side annotated comparison of segmented images). Replicated findings include: 1) usage of distinct gene expression programs in the maturing somites (S9, S8, S7; red program) versus presomitic mesoderm (PSM; pink program); 2) increasing expression of the red program as somites mature (Fig. 3.2d); 3) anterior-posterior regionalization of the nascent somite (S10), with the posterior region featuring the pink program found in the adjacent PSM and the anterior region featuring the red program found in adjacent maturing somites (S9, S8, S7); 4) corresponding anterior/posterior regionalization of the adaxial cells adjacent to the nascent somite (green program anterior, blue program posterior). New findings include: 5) additional segmentation of the adaxial cells between the nascent somite (S10) and the PSM (alternating between green and blue programs); 6) additional expression clusters corresponding to possible detection of distinct cell types within the somites (red versus brown) — discussed below; 7) a structure between the PSM and the left nascent somite (S10) that deviates from bilateral symmetry and features the same pink program featured in the PSM and the posterior region of S10.

Interestingly, k-spaces subdivides each somite, identifying distinct programs in the interior voxels versus an anterior and lateral border. Early in somite maturation, non-adaxial cells form transcriptionally-distinct anterior, medial, and posterior compartments, with the adaxial and posterior cells committing early to formation of slow muscle types (Devoto et al., 1996; Keenan and Currie, 2019; Hollway et al., 2007) and medial fast fibers (Stellabotte et al., 2007; Keenan and Currie, 2019), respectively, before migration within the somites as the somite undergoes rotation

(Hollway et al., 2007). Within the anterior compartment, there are anterior border cells (Stellabotte et al., 2007), also called row one cells (Hollway et al., 2007), which are located on the anterior and lateral edge of the somite (Fig. 1E of (Stellabotte et al., 2007); Figs. 3 and 4 of (Hollway et al., 2007)) and give rise late to lateral fast fibers (Stellabotte et al., 2007; Keenan and Currie, 2019). In the maturing somites that we analyzed (S9, S8, S7), *tpm3* and *myod1* are expressed proportionally for both the red program (inner) and brown program (anterior and lateral), but with a distinctly lower expression level for *tpm3* in the brown program. Human *TPM3.12* is the muscle-specific isoform of *TPM3* found in sarcomeres, and studies have shown it is further specific to slow muscle in humans, mice, and zebrafish (Yuen et al., 2015; Geeves, Hitchcock-DeGregori, and Gunning, 2015; Pieples and Wieczorek, 2000; Bonnet et al., 2017). Consistent with their slow muscle fates (Nguyen-Chi et al., 2012), the green adaxial cells express a substantially higher level of *tpm3*. The width of the brown region is approximately the expected ≈10 μm width of a cell in the embryo (approximated as the cube root of cell volume in Fig. S2 of (Kametani et al., 2024)). Based on location and low *tpm3* expression, the brown region could specifically be anterior somitic compartment cells, but it may additionally reflect non myogenic lineages in the somite (Nguyen et al., 2014).

In summary, with *k*-spaces we use an interpretable framework to i) identify seven gene circuit states, ii) automatically segment the image into anatomical regions featuring those circuit states, iii) visualize oscillations of the developmental gene expression program, iv) observe predominantly symmetrical but also asymmetrical circuit states in the developing embryo, and v) potentially identify distinct cell types within maturing somites. This analysis suggests that *k*-spaces is able to recapitulate patterns found using other techniques but also capture novel patterns reflecting known biology about a system.

k-spaces clustering can model DNA methylation probes in bulk epigenomics data driven by cell composition variation

Epigenome-wide association studies (EWAS) can be confounded by varying percentages of blood cell types in whole blood (Houseman et al., 2012; Listgarten et al., 2012; Rahmani et al., 2016). This issue does not arise in classical genome-wide association studies, but cell types have unique DNA methylation profiles, and disease is often associated with increases or decreases in certain cell types (Price et al., 2010). Reference-based methods use known methylation profiles of cell types to estimate cell-type proportions to control false positives (Houseman et al., 2012).

On the other hand, reference-free methods may use heuristics to identify a small number of sites in differentially methylated regions, and use PCA on those sites (hereafter referred to as differentially methylated regions, or "DMRs," for notational consistency with prior work) as a proxy for cell-type composition (Rahmani et al., 2016).

A popular reference-free method, ReFACTor (Rahmani et al., 2016), uses heuristics to identify a preset number (500 by default) of differentially methylated sites in the GALA II dataset (Esteban Burchard, 2016) and in simulations with a reference dataset (Reinius et al., 2012) to capture information on the proportions of the underlying cell types. The study posed the following model: for each DMR, global cell-type means are sampled from a normal distribution, and samples' underlying cell-type methylation profiles are sampled from another normal distribution with those cell-type means and a site-specific variance. These sample-specific underlying cell-type profiles combine with sample-specific underlying cell-type proportions to produce observed data. The study used extensive preprocessing of the GALA II data (see "GALA II analysis" in Methods, Fig. 3.12) prior to running the ReFACTor algorithm to identify the preset number of DMRs informative of cell-type composition. We set out to classify sites as DMRs or non-DMRs through probabilistic modeling and without preprocessing or heuristics.

We first demonstrate that k-spaces can solve the mathematical problem of classifying DMRs (Rahmani et al., 2016) using theory (see Supplement Section 2) and simulations. In our analyses, we centered the data so correlated deviations from global means may be modeled while retaining information about the variance. Simulated data was generated using cell-type proportions (Esteban Burchard, 2016) and cell-type methylation profiles (Reinius et al., 2012) from prior work. We ran k-spaces with a 4-D subspace (dimension one less than the number of cell types) and a 0-D subspace, respectively clustering and fitting sites with variation driven by cell-type composition (DMRs) or pure noise (non-DMRs). As subspace-specific noise parameters led to an unwanted partitioning of sites by high versus low site-specific variance (Fig. 3.13), and theory showed that a shared variance parameter between the spaces (see Supplement Section 3.4) may compensate for variation in the scale of noise in each methylation site, we exploited the constraint of sharing the variance parameter across the k-spaces.

When we sampled the site-specific standard deviations for noise from a half-normal distribution and bootstrapped the cell proportions from the GALA II data ($N = 95$),

sensitivity was 90% and specificity was 94%. When we bootstrapped the variances from the reference data (Reinius et al., 2012) used by the ReFACTor study ($N = 6$ individuals, $k = 5$ cell types, $M = 485,577$ sites) sensitivity and specificity were 79% and 97%, respectively. In ideal conditions with high variance in cell-type proportions, near perfect performance was achieved with sensitivity > 99.5% and specificity > 98% (Fig. 3.3b). Using our probabilistic model, we attained not only a high positive predictive value, like ReFACTor did on a small preset number of sites, but we also produced a classifier with high sensitivity and specificity for DMRs in theory and simulation in the absence of batch effects.

Following prior work (Rahmani et al., 2016), we turned to the GALA II dataset to evaluate its suitability out-of-the-box for correcting confounding by cell composition variation in EWAS; here, the dataset does not contain purified cell type profiles from which to determine a ground truth for DMRs versus non-DMRs. In prior work (Rahmani et al., 2016), PCA of selected DMRs was correlated with cell-type composition measurements to support the claim that informative DMRs were identified by ReFACTor. However, in that study, both samples and sites were arbitrarily removed prior to testing the software, and we found that using the full dataset led to a failure of ReFACTor's heuristic (Fig. 3.3c; Fig. 3.15a, b) and performance no better than PCA (Figs. 3.12, 3.15b).

We chose not to preprocess the data for k-spaces. An initial attempt at clustering with 4-D and 0-D spaces and a shared noise parameter predicted cell-type composition poorly in a downstream DMR-PCA versus cell composition correlation analysis (Fig. 3.15c). We hypothesized this may be because many of the called DMRs could be misclassified non-DMRs, or they could be *uninformative* DMRs because, while subtle effects of cell type composition variation may be detectable in a large sample, for any one individual, DMRs with low cell-type information may be dominated by technical noise, batch effects, or artifacts introduced by batch correction. Using the ground truth data from our prior simulation, we confirmed that classification of sites as DMRs or non-DMRs and selection of informative DMRs are not equivalent problems— PCA of all DMRs produced performance no better than using all sites for PCA (Fig. 3.16).

Because of the non-equivalency between the DMR classification problem and the cell proportion variation problem, we followed prior work (Rahmani et al., 2016) in selecting the 500 sites closest to the subspace for our analysis of the GALA II dataset. As we did not remove low-variance sites prior to clustering, we needed to

remove sites close to the subspace solely by virtue of having low variance overall to avoid using DMRs that were more likely to be dominated by noise. To do this, we filtered out called DMRs using the estimated variance parameter of the 0-D space modeling sites with pure isotropic noise (Fig. 3.17). In addition, we incidentally found that not using the shared variance constraint led to better results on the full GALA II data (Fig. 3.15). We suspect that, unlike in the simulation, there were unaccounted latent factors (either biological or batch effects) in the GALA II data resulting in a subset of sites with "structured noise" rather than either isotropic noise or a signal reflecting cell-type composition variation. These latent factors were visually evident, showing that at least a portion of noise was likely non-isotropic through persistent streaks in the visual representation of the data (Fig. 3.17). The free variance parameter may have enabled the 0-D space to capture the "structured noise" sites with additional variance from latent factors along with the expected non-DMRs with isotropic noise.

To account for these latent effects, we ran *k*-spaces with an additional 4-D space to capture structured noise. Surprisingly, PCA of the sites captured by the 4-D space representing structured noise revealed distinct subpopulations in the data (Fig. 3.18) that did not correlate with sex or ancestry labels (Fig. 3.19). With the additional space to model structured noise in the full GALA II dataset and the post-hoc selection of informative DMRs, *k*-spaces achieved the high performance in predicting cell type composition variation seen with ReFACTor applied to the preprocessed data (Figs. 3.3c, 3.15). Comparisons with ReFACTor and simple PCA of all sites on the preprocessed data are shown in Fig. 3.15. While the unsupervised application of *k*-spaces to this type of data in practice is complicated by unexpected and unknown sources of variation, we found that *k*-spaces was suitable for the mathematical problem of manifold separation and DMR classification, and we showed it had the capability to model batch effects or unwanted variation.

k-spaces offers subspaces for distance-preserving dimension reduction for exploratory data analysis in single-cell sequencing

Next, we used the dual functions of clustering and dimension reduction in *k*-spaces to resolve gene expression similarities and differences in the diverging developmental trajectories of two classes of epidermal cells—seam cells and hypodermis cells—in the *C. elegans* embryo. The embryogenesis atlas includes 86,024 single cells collected from bleach-synchronized embryos at three time points (estimated time post first cell cleavage was provided as metadata) of which 2,748 cells were seam

cells and 7,093 were hypodermis cells (Packer et al., 2019).

Traditional linear dimension reduction methods such as PCA have been replaced with embedding methods such as t-SNE and UMAP for single-cell RNA sequencing data, based on claims that the complexity of structures in the data cannot be represented with linear maps (Laurens van der Maaten and Geoffrey Hinton, 2008; McInnes, Healy, and Melville, 2020; Kobak and Berens, 2019). In turn, these nonlinear maps to two dimensions are known to produce artifactual structures and to heavily distort global distances (Chari and Pachter, 2023). Here, we fit 20 3-D spaces to the 86,024 cells in the *C. elegans* embryo dataset to gain insight into relationships between "structures" in the UMAP visualization. We do this by coloring cells according to their *k*-spaces clustering (Fig. 3.4b) while also studying *k*-spaces's informative and interpretable set of linear dimension reductions on these complex data (Fig. 3.4d, e).

After clustering, we examined the near-loop (indicated by lower black arrowhead) of the UMAP as the colors of several spaces were split across its arms, suggesting proximity between different sides of the loop, and additionally, the split orange cluster suggested proximity between the loop and the upper structure in the UMAP. PCA of those cells (the orange, red, pink, blue, and green clusters) revealed a broken loop as well (Fig. 3.4c), which was similar to that seen in the global PCA in Fig. 3.4a; however, the arrangement of colors along the structure revealed a different topology than was implied by UMAP. Therefore, we examined the clusters in each of their *k*-spaces subspaces, as each space is optimized for best representing the variance in its assigned cells in 3-D.

Unlike t-SNE and UMAP, which are not only nonlinear but also do not produce functions defined on the entire input space, the 3-D projections produced with *k*-spaces can be used to project any cell onto any space such as those from other clusters or from held out data. This allows us to observe the entire trajectory by projecting the cells across the set of linear subspaces (Fig. 3.4d) to reveal a bifurcation between hypodermis and seam cell trajectories, as suggested by the UMAP. In particular, the basis vectors of two of the subspaces - spaces 12 and 18 - represent directions in gene expression space along which seam cell and hypodermis cell developmental trajectories diverge, making these subspaces useful as dimension reductions to visualize and analyze the bifurcation event.

In addition, this application of *k*-spaces also revealed both lineage-specific changes (space 9 and space 12) and parallel gene expression changes shared by the two

lineages (space 9) that cannot be inferred from UMAP as UMAP axes do not correspond to gene expression. For the shared changes as well as the axis of bifurcation (space 18 basis 1), we further observed that more mature cells returned to the same state as early progenitors with respect to these basis vectors. We also observed the divergence of the seam cells and hypodermis cells from developing neurons (space 13), properly connecting the ectoderm lineages. These results together suggest that *k*-spaces yields more interpretable linear projections and clustering of complex, high-dimensional data than the 2-D nonlinear alternatives.

Although *k*-spaces was only applied to gene expression data, cluster membership aligned with intervals of embryo-time reported in the atlas (Fig. 3.4e). Thus, each subspace was optimized for coordinated gene expression changes during a different developmental period. A second benefit of plot axes corresponding to meaningful directions in gene expression space is the ability to extract the genes contributing most to each basis vector. We compared the genes best capturing each subspace to ground truth knowledge of embryonic events (Fig. 3.22, Tables 3.24, 3.23). For example, we interpreted two basis vectors in one space (1 and 3 in space 9) as potential shared gene expression programs for cuticle synthesis, which both cell types participate in. Examining bases of other spaces revealed several signaling genes, such as warthog and groundhog-like gene family members, genes involved in molting, and dumpy family members (named for mutants producing a shortened worm), among others. Collectively, these results suggest that *k*-spaces structures recapitulate known biological signals more accurately from unsupervised data analyses than their nonlinear counterparts.

Discussion

In this study, we presented a probabilistic model that we call *k*-spaces for joint clustering and dimension reduction of data generated by mixtures of linear subspaces defined by latent Gaussian variables. We described how the *k*-spaces model relates to other methods in this space, and showed its applications in biological data analyses. One of the primary motivations for this study was to develop a framework to study the latent space of gene expression and how correlated modules of genes scale their expression across time or space in response to external stimuli or innate physiological demand. Our *k*-spaces framework may generally be used to study systems that may be approximated as linear latent variables modulated by categorical latent variables such as cell type, tissue type, batch, sex, disease status, or genetic background. Though inspired and initially derived from a *k*-means framework (Bradley and

Mangasarian, 2000; Agarwal and Mustafa, 2004; Teng Zhang, Arthur Szlam, and Gilad Lerman, 2009; René Vidal, 2010; Gitlin et al., 2018; Lipor et al., 2021), we have connected k-spaces to PCA, GMMs, and probabilistic PCA, reiterating closely related results and interpretations in prior work (Michael E Tipping and Christopher M Bishop, 1999a; Roweis, 1998), while also generalizing the model to subspaces of mixed dimension and shared noise parameters.

In retrospect, it is unsurprising that this more general model is connected with these specific models, as the L_2 norm and the Gaussian distribution are fundamentally linked (John Aldrich, 1998; Robin L. Plackett, 1949). Methods that find the mean of data (such as total least squares (Robert J. Adcock, 1978; W. Edwards Deming, 1943) or k-means clustering), which corresponds to minimizing the L_2 norm or sum of squared residuals, implicitly fit Gaussian distributions, and PCA has Gaussian interpretations both from the perspective of minimizing the L_2 norm and maximizing the likelihood of Gaussian data (see Supplement Section 1). Thus, we can approach this k-spaces framework both as an interpretable clustering and dimension reduction method or as a linear latent space approximation where a theoretical analysis of failure modes is tractable. We term our method k-spaces to emphasize the focus on probabilistic learning of subspaces rather than the specific distribution, and future work can include generalizations to other data-relevant distributions (Archambeau, Delannay, and Verleysen, 2008; Chiquet, Mariadassou, and Robin, 2018).

Our application to qHCR imaging of zebrafish somitogenesis shows the potential for k-spaces to be used to scale and automate current methods of analyzing biological data that are done by hand, especially where the learned subspaces are themselves valuable for study and can be assessed with human expertise but manual clustering is intractable as direct visualization of the complete data is not possible.

Given the exploratory power of quantitative analyses moving back and forth between anatomical space and expression space, it is of interest to increase the scale of read-out/read-in analyses (more genes, more ROIs, more expression clusters). However, as the problem scale increases, a manual approach becomes unwieldy, both because the number of 2-D scatter plots increases quadratically with the number of multiplexed targets (e.g., 10 targets yields 45 2-D scatter plots), and because the number of regions of interest (in anatomical space) and expression clusters of interest (in expression space) can grow out of hand as the spatial extent of the study increases, as in whole-embryo analyses. qHCR spectral imaging now enables robust 10-plex quantitative imaging in highly autofluorescent samples including whole-mounts and

delicate specimens that are not amenable to repeated imaging (Schulte, Fornace, et al., 2024), opening the door for 10-dimensional automated read-out/read-in using *k*-spaces on whole-mount embryos to examine more and larger circuits. qHCR imaging has also been extended to enable compatible quantitative imaging of RNA (Trivedi et al., 2018; Choi et al., 2018), protein (Schwarzkopf et al., 2021), and protein:protein complex targets (Schulte, Shin, et al., 2024) in highly autofluorescent samples, creating the opportunity for automated read-out/read-in using *k*-spaces to examine circuits involving selected targets from the transcriptome, proteome, and interactome.

Application of *k*-spaces to qHCR data revealed another benefit in terms of its inductive bias. Tipping and Bishop demonstrated the usefulness of pPCA over full rank Gaussian models in the setting of high dimensional, sparse data, where fewer free parameters makes pPCA more robust to overfitting (Michael E. Tipping and Christopher M. Bishop, 1999b). Here, the inductive bias of *k*-spaces—that gene expression is scaled along lines—allows it to identify clusters that would be merged by full rank GMMs, and indeed, cluster labeling of voxels appears to be more anatomically correct with *k*-spaces than with GMMs or *k*-means (Figs. 3.7 and 3.11). Furthermore, full rank GMMs lack the notion of a latent space, while *k*-means has a 0-D latent space and no continuous latent variable, but only discrete clusters and noise (though we should properly refer to this as a spherical GMM with equal covariance matrices in a hard assignment setting). In contrast, *k*-spaces has an explicit notion of a latent space and allows us to visualize oscillations of gene expression programs in complex systems such as a developing embryo. *k*-spaces could also be applied to qHCR images of other systems such as bacterial biofilms (Livingston et al., 2022) or to spatial RNA sequencing data, though application to unnormalized sequencing data would likely require an extension of the *k*-spaces software using a probabilistic model for count data (Chiquet, Mariadassou, and Robin, 2018).

k-spaces also strikes a balance between being sophisticated enough to model complex data while being simple enough to theoretically analyze. In the context of bulk epigenomics, we analytically showed that the *k*-spaces model is appropriate for this data (even when the variation in noise across methylation sites produces a non-Gaussian distribution of noise) by (mis)using a shared noise parameter between the spaces. Furthermore, application to epigenomics allowed us to identify a sudden failure of the model at α between 40 and 50 (Fig. 3.3c). Our prior theoretical

analysis suggested a possible cause, and we support this with simulation evidence by removing the top 0.03% noisiest sites to restore robustness to low cell proportion variation (Fig. 3.30; see Supplement Section 3.4).

The inspiration for applying *k*-spaces to bulk epigenomics data, ReFACTor, was tested on a subset of the GALA II data, both in terms of samples and sites, and used heuristics (Rahmani et al., 2016). While we showed that the DMR classification problem as set up by the simulation in (Rahmani et al., 2016) could be solved by *k*-spaces, we learned that selection of informative DMRs, whether in real or simulated data, was a distinct problem (Fig. 3.16). Theoretical justification or empirical validation of our heuristic selection of "informative" DMRs after clustering would be needed for wider use. Furthermore, while visualization of the data (Fig. 3.17) suggested which of the two 4-D spaces represented the desired signal and which was structured noise, a robust selection strategy of the correct dimension for the second space is unclear; we merely chose 4 as it seemed plausible that a subset of sites were behaving systematically differently yet also affected by underlying cell proportion variation. In the future, the out-of-the-box *k*-spaces framework could be tailored to this epigenomics problem. However, with the wider availability of single-cell epigenomics assays (Preissl, Gaulton, and Ren, 2023), development of a robust reference-free computational method to identify DMRs may be less of a pressing need and is outside of the scope this study, though we used this example to challenge *k*-spaces with a problem requiring manifold separation through dimension.

We also explored other applications in transcriptomics where *k*-spaces may be used. In addition to visualizations for single-cell data (Fig. 3.4), we also examined species mixture "barnyard" plots and gene correlation in single-cell sequencing. In droplet-based RNA-sequencing, there is some low level of background contamination in the sample. *k*-spaces offers a principled way to select the threshold for when nonzero UMI counts are too high and the droplet likely captured more than one cell (See Supplement Section 6, Figs. 3.26a, 3.27).

We next assessed whether distinct relationships between pairs or groups of genes can be identified and characterized with *k*-spaces clustering in single-cell data, analogous to our qHCR analysis (see Figs. 3.26, 3.28, 3.29). This is important because experiments that identify co-regulated modules of genes using techniques, such as weighted gene co-expression network analysis (Langfelder and Horvath, 2008), often begin by constructing a gene-gene correlation matrix, but this assumes the linear correlation between a pair of genes is representative of their relationship.

While *k*-spaces was able to detect distinct relationships, it overclustered the data when used in an unsupervised fashion with model selection via the Bayesian information criterion (BIC) or the integrated completed likelihood (ICL), which adds an entropy penalty to BIC to approximate the joint likelihood of the data and the clustering solution (Biernacki, Celeux, and Govaert, 2000; Bertoletti, Friel, and Rastelli, 2015). One particularly challenging artifact contributing to this overclustering is a linear structure in pairwise scatter plots – especially in genes with low counts – introduced by depth normalization because angular position is determined by pairs of discrete counts, producing streaks in the data with quantized slopes but pseudo-continuous scaling along the lines. In the extreme case, *k*-spaces with Gaussian distributions will fit a separate noiseless line to each streak, and even outside of this extreme setting, it is difficult to determine whether the "distinct" spaces are artifacts rather than biology. To address this and other challenges, we envision an extension of our *k*-spaces framework with a generalization of the negative binomial distribution to mixture models, where the multidimensional gamma priors in the gamma-Poisson formulation capture the latent spaces to be inferred. We anticipate this will be particularly useful in handling the artifactual linear structures as well as overdispersion. In genes with low counts, it is not necessarily true that *no* shared latent variable controls gene expression and that all such gene abundances are independent; rather, we lack the resolution to make any inference with *certainty*. Modeling this within a Bayesian framework should make this evident with wide, uncertain posteriors.

Finally, we believe one of the most important contributions of *k*-spaces is the expansion of the biologist's exploratory data analysis (EDA) toolkit to provide insights in to global relationships and better visualize finer structures with *k*-spaces. Most applications discussed in this study involve directly modeling latent biological processes, but we show the ability to tease apart these finer structures with the single-cell RNA-sequencing data for *C. elegans* embryogenesis. Importantly, this neither requires proper selection of k nor finding the "correct" clustering solution, as underclustering to capture multiple structures in one subspace allows their comparison, and visualization with different *k*-spaces results can yield different insights. Tipping and Bishop used a mixture of pPCAs to partition a dataset and view the relative maximum a posteriori positions of data points within their own assigned pPCAs (Michael E Tipping and Christopher M Bishop, 1999a), and they used a hierarchical technique that performed mutual projection to only show ambiguously assigned points (Christopher M. Bishop and Michael E. Tipping, 1998). We extend

this idea by projecting points onto each other's subspaces to view entire structures in the context of gene expression space. The hierarchical technique for pPCA mixtures proposed by Su and Dy (Su and Dy, 2004) could also be useful in guiding EDA, though the nonhierarchical clustering we use may be more helpful for detecting global similarities in distantly related cells.

One limitation of our approach is that *k*-spaces visualizations can plot points in proximity to each other that are well-separated in another axis of variation. This limitation echoes the same behavior in PCA in obscuring fine structures when the data are underclustered. We addressed the former by decreasing the sizes of points to reflect increasing distance from the 3-D space of projection. Thus, larger points that are plotted in proximity are likely to be close in the ambient gene expression space, points of drastically different sizes are guaranteed to be separated in unseen dimensions, and small points are understood to be far away but may diverge from the subspace in different directions. The Eckart-Young theorem (Eckart and Young, 1939) states that PCA is the optimal linear dimension reduction for visualizing variance in data. By definition of being nonlinear, other methods will distort distances or, worse, provide embeddings with features that do not correspond to continuous axes – linear or nonlinear – in gene expression space. Visualizing the data with piecewise linear dimensionality reductions trades off the global optimality of PCA for local optimality to offer more fine grained visualizations of structure while still providing interpretable axes. Critically, the shortcomings of *k*-spaces are tractable to analyze for a specific application. In this particular case each space can be viewed as a 3-D PCA for a subset of the data. By using this framework to visualize single-cell data, we can not only see high-resolution structures in data but also quantify the reliability of our inferences.

Data and code availability

The *k*-spaces software package is available at `https://github.com/pachterlab/k-spaces`.

The purified human blood cell reference epigenomics data was accessed on September 11, 2024 at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35069`. The GALA II epigenomics data was accessed on September 17, 2024 at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77716`. Code to preprocess data is included in our notebooks in our repository but sample and probe IDs for the full and the filtered datasets are additionally available at

https://github.com/pachterlab/MEPSP_2024/data_paper.

HCR data was downloaded as part of the Read-out/Read-in Software Package on November 15, 2024 from https://www.moleculartechnologies.org/info/software. Both the raw data and preprocessed matrices are available in our repository at https://github.com/pachterlab/MEPSP_2024/data_paper.

The *C. elegans* embryogenesis data was accessed on August 25, 2025 at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE126954.

The human-mouse mixture, 10K PBMC, and 5K PMBC datasets were downloaded from the 10X Genomics website at https://www.10xgenomics.com/datasets/10k-hgmm-3p-gemx, https://www.10xgenomics.com/datasets/10k-human-pbmcs-3-v3-1-chromium-x-without-introns-3-1-high, and https://www.10xgenomics.com/datasets/5k_Human_Donor4_PBMC_3p_gem-x (filtered matrix), respectively.

We used anndata v0.9.2, MATLAB v24.1.0.2653294 (R2024a) Update 5, Numpy v1.23.5, Pandas v2.1.4, Python v3.10.9, ReadoutReadin v1.0, ReFACTor v1.0, scanpy v1.10.2, scipy v1.11.4, and sklearn v1.3.0. The full environment is available at https://github.com/pachterlab/MEPSP_2024. Code for reproducing the results and figures in the manuscript is available at https://github.com/pachterlab/MEPSP_2024 under Notebooks.

Author contributions

The *k*-spaces clustering algorithm was conceived by BE and LP, and developed and implemented by NM. NM, BE, NP, PS, and LP conceived the applications, and NM generated the results. NP provided qHCR data and directed its analysis. NM, BE, NP, PS and LP analyzed results. The manuscript was drafted by NM, and was edited and approved by all authors. PS and LP supervised the project.

Acknowledgments

We thank the Banff International Research Station for supporting and hosting the workshop on challenges and synergies in the analysis of large-scale population based biomedical data in Oaxaca, Mexico in 2017. The workshop led to discussions between BE and LP that led to this project. We thank Caleb Ghione, Ulrich Herget, and Peter Currie for their kind assistance in interpreting HCR data and discussions about zebrafish embryology and anatomy. This work was supported by NIH F30 1F30GM156092-01 to NM and by NIH-NHGRI grant U24HG010859 to PWS, a

Bren Professor of Biology, and by the Beckman Institute at Caltech (PMTC). BEE was funded in part by grants from the Parker Institute for Cancer Immunology (PICI), the Chan-Zuckerberg Institute (CZI), the Biswas Family Foundation, NIH NHGRI R01 HG012967, and NIH NHGRI R01 HG013736. BEE is a CIFAR Fellow in the Multiscale Human Program.

Methods

Overview

Our model consists of a mixture of latent Gaussian variables defining distributions on k affine subspaces $s_1 \dots s_k$ of dimensions $d_1 \dots d_k$ with the latent points perturbed by isotropic Gaussian noise in the complementary space to generate observed data. We consider the cases where the variance for noise can vary by latent space and where it is the same for all subspaces. For conciseness, we refer them to the individual latent spaces as "spaces," and we refer to our mixture model as k-spaces.

We model each observation $\mathbf{x} \in \mathbb{R}^D$ as the sum of a structured component in a latent subspace and an independent noise component in the complementary subspace such that:

$$\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp},$$

where $\mathbf{x}^{\parallel} \in \mathbb{R}^d$ lies in a latent affine subspace associated with a mixture component (signal), and $\mathbf{x}^{\perp} \in \mathbb{R}^{D-d}$ lies in the orthogonal complementary space (noise). Under this model, the density of $\mathbf{x}$ conditioned on mixture component space s factorizes as

$$f_X(\mathbf{x} \mid s) = f_{\parallel}(\mathbf{x}^{\parallel} \mid s) \cdot f_{\perp}(\mathbf{x}^{\perp} | s), \tag{3.1}$$

where $f_{\parallel}(\mathbf{x}^{\parallel} \mid s)$ is the multivariate Gaussian density of the latent component under the mixture model, and $f_{\perp}(\mathbf{x}^{\perp}|s)$ is the isotropic Gaussian density for the independent noise component of $\mathbf{x}$ that lies in the complementary space.

We denote by $f_X(\mathbf{x}_j; s_i)$ the probability density function of observation $\mathbf{x}_j$ under space s_i. Let $n_i = D - d_i$ for convenience. The probability density function for each space is therefore

$$f_X(\mathbf{x}_j; s_i) = \frac{1}{(2\pi)^{d_i/2} \prod_{a=1}^{d_i} \sigma_{ia}} e^{-\frac{1}{2} \sum_{a=1}^{d_i} \frac{\text{disp}_{s_i}(\mathbf{x}_j)_a^2}{\sigma_{ia}^2}} \cdot \frac{1}{(2\pi)^{n_i/2} \sigma_i^{n_i}} e^{-\frac{1}{2} \frac{r_{ij}^2}{\sigma_i^2}}, \tag{3.2}$$

where $\text{disp}_{s_i}(\mathbf{x}_j)_a$ and σ_{ia}^2 correspond to the displacement of the projection along principal axis a of subspace i and that axis's respective variance, r_{ij}^2 refers to the length of the orthogonal residual from $\mathbf{x}_j$ to s_i, and σ_i^2 refers to the variance of

the noise in each complementary dimension for space i. Note that the resulting distribution is Gaussian and the density function can be written concisely with a single covariance matrix. However, this form, which emphasizes the spectral decomposition, allows for separation into the first term corresponding to the latent space and the second corresponding to the complementary space. It is of interest in biological applications where the position relative to other latent points or to the mean within the latent subspace rather than position in the ambient space (or conversely the distance from the subspace) is relevant.

Each component has a mixing proportion π_i, where $\sum_{i=1}^{k} \pi_i = 1$. The marginal likelihood of $\mathbf{x}_j$ under the mixture model is then given by

$$f_X(\mathbf{x}_j) = \sum_{i=1}^{k} \pi_i f_X(\mathbf{x}_j | s_i).$$

Unless noted otherwise, throughout the Methods and Supplement our derivations relate to this probabilistic model. We estimate the parameters by maximizing the model likelihood with an expectation-maximization (EM) approach, in which points are soft-assigned to spaces based on their probability density functions and mixing proportions, and then spaces are fit in parallel based on eigenvalue decomposition of their points' weighted covariance matrices. The noise is determined as a function of the residuals (see Supplement Section 2). In the case of hard-assignment, where each point is fully assigned to its most likely space of origin, the M-step is performed with a singular value decomposition (SVD) solver. Lastly, also implemented in our software package is a direct generalization for k-means, where points are assigned to the closest space and spaces are fit to their assigned points by SVD. This is distinct from our probabilistic model as it assumes no distribution over the latent space and is equivalent to assuming the noise is equal for all spaces ($\sigma_i^2 \ \forall i$), i.e., this approach directly minimizes reconstruction error and is not a generative model.

Derivation

We can generalize the objective function for k-means clustering,

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \left\| \mathbf{x}_j - \boldsymbol{\mu}_i \right\|^2, \tag{3.3}$$

to an objective for clustering onto a set of k affine subspaces $s_1 \ldots s_k$ of dimensions $d_1 \ldots d_k$:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in \mathcal{S}_i} \left\| \mathbf{x}_j - \text{proj}_{s_i}(\mathbf{x}_j) \right\|^2 . \tag{3.4}$$

In k-means, C_i is the set of points closer to the mean of cluster i than any other cluster. In the most immediate generalization, points can be assigned to their nearest space; this was done in the image compression and computer vision communities (Agarwal and Mustafa, 2004; Bradley and Mangasarian, 2000; René Vidal, 2010). However, we assume a latent distribution and therefore for hard assignment we use $\mathbf{x}_j \in \mathcal{S}_i$ if $i = \arg\max f_X(x_j; s_i)\pi_i$. Relaxing the hard-assignment requirement, we obtain a maximization of the expected complete log likelihood

$$\arg\max_{\mathbf{S}} \mathcal{L}(\mathbf{X}; \mathbf{S}), \tag{3.5}$$

$$= \arg\max_{\mathbf{S}} \sum_{j=1}^{N} \sum_{i=1}^{k} \pi_{ij} \log(f_X(\mathbf{x}_j; s_i)\pi_i), \tag{3.6}$$

which we optimize by EM. Setting $d_i = 0$ and $\sigma_i = \sigma \; \forall i$ yields k-means clustering in the hard-assignment setting or its probabilistic analog: spherical Gaussians of equal variance. Setting $k = 1$ yields PCA. In the M step of the EM procedure for this algorithm, PCA can be performed to fit spaces to their assigned points, as the optimal decomposition into $\mathbf{x}^{\|}$ and $\mathbf{x}^{\perp}$ minimizes the distances from points, $\mathbf{x}$, to their projections, $\mathbf{x}^{\|}$, as in equation (4), making this a PCA mixture model (see Supplement Section 1.1). In many contexts, it is desirable to perform a probabilistic assignment of points to spaces as well as to assume a probability distribution over the latent space, which is critical to proper clustering when the dimensions of the subspaces varies, as will be discussed shortly. Without a latent distribution, higher dimensional spaces are favored over lower dimensional ones for clustering. When we assume latent Gaussian distributions with isotropic Gaussian noise in the complementary space, we obtain a constrained Gaussian Mixture Model (GMM), which is approximately equivalent to a mixture of pPCAs with varying dimension.

Model selection reveals the need for a latent distribution

An approach based on likelihoods of residuals alone fails when comparing spaces of different dimension, as subspaces of higher dimension observe out-of-subspace variance in fewer dimensions of the data. Consider a sufficiently large set of points

sampled from a bivariate standard normal $\mathcal{N}_2(\mathbf{0}, \mathbf{I})$. Our objective in equation (4) minimizes out-of-subspace variance, the per-point mean squared distance to the subspace. Fitting a single point ($k = 1, d = 0$) will yield a space with basis = $\emptyset$ and translation = $\mathbf{0}$. Because the bivariate standard Normal is rotationally symmetric, any line that passes through $\mathbf{0}$ is an equally good fit, so without loss of generality, assume fitting a single line ($k = 1, d = 1$) yields the basis [1,0] and translation $\mathbf{0}$. Thus, it is apparent that the mean squared residual length $\sigma^2_{line} = 1$, as the residuals' lengths correspond to the y-values of the points in the sample. However, for the point, $\sigma^2_{point} \neq 1$.The mean squared residual is 2 — the distances from points to the origin are $\sqrt{x^2 + y^2}$, and each dimension has a variance of 1. Note that the distances do not follow a Half-normal distribution; they follow a Rayleigh distribution, but regardless, the likelihood of residuals when fitting a point must account for variance in both the x and y dimensions, while the likelihood of residuals when fitting a line only needs to account for variance in the y dimension. Although the data was generated from a 0-dimensional subspace with isotropic noise, if we were choosing between models with $d = 1$ and $d = 0$, the likelihood of the residuals under the $d = 1$ model would be better, and if we had a model with both spaces and were assigning points in the expectation step, the $d = 1$ space would always be favored except exactly at the origin. For these reason, we assume a Gaussian distribution over the latent space, akin to Tipping and Bishop's pPCA.

qHCR Image Analysis

Preprocessing

Data for the analysis in Fig. 3.2 is included in the Read-out/Read-in v1.0 software package, which was retrieved from `www.molecular technologies.org/supp/ReadoutReadin1.0.zip`. A single 304 x 510 pixel ROI was placed at the x, y coordinates (1,3) in the image, and the voxel size was 3x3x1 pixels as in the default setting. Voxel intensities are normalized per-channel by default in the Read-out/Read-in workflow. Rather than perform scatterplot analysis in the user interface, the data was exported and used in our k-spaces analysis.

k-spaces clustering

k-spaces was fit with $d_i = 1 \forall i$ for k ranging from 1 to 12, with 100 random initializations each and a tolerance setting of tol = .0005. This was repeated with $d_i = 3 \forall i$ (a standard GMM in 4-D space) and with $d_i = 0 \forall i$ and assignment =

'closest' (standard k-means) rather than the default of 'soft', as in all other portions of this study. For coloring the voxels by cluster in the figures as well as coloring points on the scatterplots, final soft cluster assignment was converted to a hard assignment by selecting the space with the maximum responsibility for each point.

Scatter plots

In the plots, line segments (the basis vector for each 1-D affine subspace) were centered at the cluster mean and extended for one standard deviation in each direction along the latent space before projection onto the plane of the plot. Line lengths therefore reflect absolute variation in expression for each cluster and each gene, and relative lengths for the same cluster across different plots reflect relative influence of each gene on the direction of the space.

Visualization of spatial gradients in Fig. 3.2d

In the plot of y coordinate vs gene circuit expression, for each cluster, the mean L_2 norm per y coordinate ($\frac{1}{N_y}\sum^{N_y}||\text{space}_{\text{proj}}(\mathbf{v})||_2$) was plotted if there were at least 5 voxels with that label at that y coordinate. 5 voxels was chosen because it corresponds to 10 μm (the approximate cell dimension). In the image, pixel intensity was determined by $\frac{||\text{space}_{\text{proj}}(\mathbf{v})||_2}{\max_{\mathbf{v}_c \in \text{cluster}}(||\text{space}_{\text{proj}}(\mathbf{v}_c)||_2)}$, where voxels of each cluster were normalized to a maximum of 1 separately. Without this normalization, clusters with overall low gene expression would otherwise appear dim, and both hue and gradients would be difficult to discern. An alternative would be to use the displacement along the cluster latent space (the line) from the cluster mean and to scale values from 0 to 1 for saturation. However, this would result in a saturation of 0 for some voxels despite the presence of nonzero gene expression and may be a misleading visualization.

Differential Methylation

Simulation

Simulation was performed as described in the ReFACTor study (Rahmani et al., 2016). Let $X^0_{N^0 \text{x} J^0 \text{x} K^0}$ be a matrix of observed methylation values for each individual at each site and for each cell type, $M^0_{J^0 \text{x} K^0}$ be a matrix of methylation values for each cell type at each site, averaged across the 6 individuals, and $R^0_{N^0 \text{x} K^0}$ be a matrix of each individual's cell type proportions in the reference data ($N^0 = 6$, GSE35069) (Reinius et al., 2012). Notation here is chosen for consistency with (Rahmani et al.,

2016) and does not refer to variables of the same name in our presentation of the k-spaces model. The simulation procedure used is as follows:

1. Determine a unique variance for each site, σ_j^2: An empirical distribution of site-specific variances was determined via $\sigma_j^2 = \frac{1}{(N^0K^0-1)} \sum_{i=1}^{N^0} \sum_{k=1}^{K^0} ((X_{ijk}^0 - M_{jk}^0)^2)$. However, $\sigma_j \sim$ Half-normal$(0, \sigma^* = 0.03)$ was used in Fig. 3.30 and prior to bootstrapping variances from the above empirical distribution for the final simulation as noted in the text.

2. For each site, sample μ_j from some distribution: This step is not particularly relevant as we center the data, but we used Norm$(0.5, 0.15)$.

3A. Randomly select 15% of sites to be DMRs: For each of these sites, sample K values for the row M_j via $M_{jk} \sim$ Norm$(\mu_j, \tau^2 = 0.07)$. τ was determined in the ReFACTor study.

3B. For non DMR sites, set $M_{jk} = \mu_j$ for all k.

4. For each individual and each cell type, at each site $X_{ijk} \sim$ Norm(M_{jk}, σ_j).

5. Sample each individual's cell type proportions $\mathbf{R}_i$ from a Dirichlet distribution or bootstrap from reference data.

6. Obtain each O_{ij} by taking $\sum_{k=1}^{K} X_{ijk}R_{ik} + \epsilon_{ij}$ where $\epsilon_{ij} \sim$ Norm$(0, 0.01)$ and truncate values to [0,1].

500 individuals and 100,000 sites were simulated with 5 cell types. k-spaces with a 4-D and 0-D spaces was fit to the data using a shared noise parameter. Sites were called as DMRs if their assignment responsibility to the 4-D space was greater than 0.5.

GALA II analysis

For the analyses in supplementary Figs. 3.15, 3.31, and 3.32, filtering of samples and sites prior to using the ReFACTor software was done as described in the ReFACTor study (Rahmani et al., 2016). Normalized and batch-corrected data in the series GSE77716 were downloaded from GEO. First, samples were projected onto the top two PCs. Any sample greater than two standard deviations away from the mean on either PC was removed. Then, all sites with means greater than 0.8 or less than 0.2 were removed. Prior to running ReFACTor, we did not manually remove sites with standard deviations less than 0.02, nor did we center and scale the data to unit variance, as these steps are part of the ReFACTor algorithm and are done by its

software. ReFACTor v1.0 was run with $k = 6$ as described in its paper with default settings, t = 500 and standard deviation threshold = 0.02. For the k-spaces clustering in the analyses in Figs. 3.3 and 3.17, the data from GEO was used directly and only centered. In neither case did we scale the data, and all samples and sites were included. Clustering was performed with two 4-D subspaces (to model signal and structured noise) and a 0-D subspace (to model unstructured noise) with a shared noise parameter unless noted otherwise. Ancestry labels used in Fig. 3.19 were taken from the "ethnicity" column of the GALA II metadata.

Preprocessed subsets of GALA II data for Fig. 3.15

The ReFACTor software itself removes low-variance sites; the numbers below reflect the dataset sizes that were inputted to k-spaces and ReFACTor. Note that the numbers in A will not match (Rahmani et al., 2016) exactly as only 525 samples were available at the time of their study. For example, the number of individuals with cell proportion measurements after preprocessing in (A) was 80, while in (Rahmani et al., 2016) there were 78. Similarly, there were 102,503 probes used in their analysis to our 102,470 in (A).

A: the data produced by the preprocessing procedure described in (Rahmani et al., 2016) (subset of 546 individuals and 102,470 sites). ReFACTor removed 1,142 low-variance sites downstream.

B: filtering out sites on the basis of mean > 0.8 or < 0.2 (all 573 individuals, subset of 103,024 sites). ReFACTor removed 640 low-variance sites downstream.

C: filtering invdividuals on basis of PCs (subset of 546 individuals, all 473878 sites). ReFACTor removed 263,620 low-variance sites downstream.

Full: all data (all 573 individuals and 473,838 sites). ReFACTor removed 226,847 low-variance sites downstream.

Transcriptomics

***C. elegans* embryogenesis**

A preprocessed single-cell RNA sequencing atlas of C. elegans embryogenesis was from GEO (GSE 126954), comprised of 86,024 cells from gastrulation to terminal cell differentiation that passed quality control (Packer et al., 2019). We multiplied counts by the size factors determined by Packer and coworkers to depth normalize and log1p transformed the data. We then removed 1,172 genes with 0 variance and

used SVD for a linear dimensionality reduction to 100 dimensions, a typical step performed before nonlinear dimensionality reduction methods like t-SNE (Laurens van der Maaten and Geoffrey Hinton, 2008) and UMAP (McInnes, Healy, and Melville, 2020). We then separately produced a 2-D UMAP visualization of the data with default parameters and ran k-spaces clustering with 20 3-D spaces to obtain a piecewise set of linear dimensionality reductions for the data. $k = 20$ was chosen simply because the tab20 color palette has 20 colors, and the goal was not necessarily to produce an anatomically- or developmentally-coherent clustering to match a ground truth but rather to partition the data into a manageable number "viewing windows." Timing for annotated embryogenesis events in the timeline of Fig. 3.4 were obtained from WormAtlas (Zeynep F Altun and David H Hall, 2009a; Zeynep F Altun and David H Hall, 2009b).

After choosing spaces 6, 9, 12,13, and 18 due to the discrepancy in their clustering with the visible near-loop in the UMAP, we projected cells assigned to all 5 spaces onto each of the 5 spaces to confirm a continuous trajectory and view gene expression changes represented by each space. As projections can obscure separation between points in other dimensions and each space maximizes variance within only its own cells and is not necessarily a good dimensionality reduction to show separation of cells assigned to other spaces, we encoded the distance for each cell from each subspace by the size of points. We scaled the size of the points by the cube of their distance in the ambient space from the subspace. We chose to raise them to the power of 3 because it provided a better dynamic range.

After selecting basis vectors (100-dimensional) that appeared to represent key parts of the trajectory, we mapped them back to gene expression space (19,050-dimensional) by multiplication with the original SVD components. We retrieved a list of genes to represent each PC based on a cutoff of +/- 0.05 for loadings and performed complete gene ontology enrichment analysis for biological process, molecular function, and cellular component with PANTHER release 19.0 (overrepresentation test/Fisher's exact test with FDR multiple testing correction) (Thomas et al., 2003; Mi et al., 2019). The cutoff of 0.05 was chosen based on the empirical cumulative distribution functions of the basis loadings (coefficients denoting the direction of the basis vector in gene-expression space).

Barnyard Plot

The Cell-ranger preprocessed 10k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 Cells, Chromium GEM-X Single Cell 3' dataset was downloaded from the 10X Genomics website. The number of UMIs mapping to the GRCh38 and GRCm39 genome assemblies were summed for each cell. No further preprocessing was performed.

Gene-gene relationships

The 10K PBMC 3' nextgem Chromium X filtered feature by cell matrix and 5K Human PBMCs (Donor 4) were downloaded from the 10X Genomics website and preprocessed with scanpy (Wolf, Angerer, and Theis, 2018). Cells with counts in less than 100 genes and less than 3000 total counts were removed, as well as genes expressed in less than 200 cells. Doublets were then removed with scrublet (Wolock, Lopez, and Klein, 2019). Counts were then depth normalized to the median number of counts and log1p (natural logarithm as per the default in scanpy.pp.log1p) transformed.

After this, further filtering of cells and genes were performed to avoid clustering on genes with apparent coexpression due to normalization artifacts such as those in perceptible SRSF5 vs ATP5MG in Fig. 3.28, as structure in the data for genes with low counts is dominated by artifacts of depth normalization for count data. This effect is much more pronounced in non highly expressed genes. To produce random examples of clustering in Fig. 3.28 with the 10K PBMC dataset, genes with less than 10,000 total counts were excluded from further analysis, as well as cells with less than 5,000 or greater than 15,000 total counts, resulting in 8,219 cells and 213 genes. 100 pairs of genes were then chosen randomly.

Limitations

Like GMMs and k-means clustering, selection of k is challenging in the case of non-Gaussian distributed data without a Bayesian prior. We found ICL performed better than BIC, but it was still insufficient on its own, although combination with an automated elbow method (Ville Satopaa et al., 2011) improves performance. There is also the problem of selecting the proper dimension d_i for each cluster, a known challenge in PCA. Scree plots or cross validation are typically used for PCA, but a proposed Bayesian method using a prior may be a viable approach to implement (Minka, 2000). Until methods for selection of k or subspace dimension

are validated, we emphasize caution in correlating the latent variable(s) of a cluster with observed data points or using the explained variance in a Scree plot as a method for validating clustering. For example, data generated on the surface of a 2-D square could be (mis)partitioned into clusters from fitting the data with 2 lines. Unlike BIC, the ICL criterion discourages this, but should it occur, each cluster will indeed appear more line-like when viewed in isolation as the clustering method itself implies a single latent variable per cluster. This is reminiscent of a statistical "double dipping" problem with differential expression analysis after clustering addressed by Gao, Bien, and Witten in (Gao, Bien, and Witten, 2022). Viewing clusters together will be of benefit in this scenario.

Supplement

1: Exposition on PCA and pPCA

1.1: Review of PCA

Principal component analysis, or PCA, is a common dimensionality reduction method used in biology, and is often specifically used to visualize data in 2-D plots. These visualizations are the projections of the data onto principal components—typically the first two. These principal components themselves are linear combinations of features in the data (i.e. directions in feature space). They are perpendicular directions in the feature space chosen such that the first component is the direction of maximum variance, the second is that with the most variance after removing the first component, and so on. Essentially, PCA defines a new coordinate system in our feature space, and it could be viewed as a rotation of our starting axes (maintaining the orthogonality) followed by dropping "less important" axes that contain less variance from our visualization. As an analogy for 2-D PCA of 3-D data, one could consider entering a rectangular room and seeking the best angle to take a photograph where all subjects can be seen. The walls of the room and the doorway provide a natural set of axes and "origin" from which to frame the photograph, but some subjects may block each other from view so finding a new vantage point may be better. Depth perception is lost in the photograph, and this is akin to discarding variance from our 3-D space to produce the 2-D projection.

Mathematically, determining this coordinate system is equivalent to finding a subspace that is "closest" to the data, meaning it minimizes the Frobenius norm or sum of squared distances (L_2 norms) from the data points to the subspace. (Note the similarity to linear regression, which uses residuals in a dependent variable while

this PCA procedure uses orthogonal residuals, implying errors in all variables.) It is the optimal linear dimensionality reduction method in the sense that it retains the maximum variance in the data, or "spaces out" the plotted points as much as is possible without a nonlinear distortion (Eckart and Young, 1939). Unlike tSNE and UMAP, both local and global distances are meaningful, though points that appear next to each other in a plot may be further apart in another, discarded dimension (Chari and Pachter, 2023; Laurens van der Maaten and Geoffrey Hinton, 2008; McInnes, Healy, and Melville, 2020).

The process of minimizing the sum of squared residuals and the goal of retaining the maximum variance (implying that we are minimizing the discarded variance) suggest a relationship with the Normal or Gaussian distribution and the covariance matrix. Indeed, one way to obtain PCA is to construct the covariance matrix of the data and then use singular value decomposition (SVD) to decompose it into its eigenvalues and eigenvectors, also called the spectral decomposition. The eigenvectors with the largest corresponding eigenvalues (or variances of projections of the data along the eigenvectors) are the principal components. Dropping lower PCs is equivalent to ignoring the components of our data along the directions of our lower eigenvectors. Thus, we can also arrive at PCA by fitting a Gaussian to the data first and then determining how the distribution is "stretched" in different directions to find those with the greatest variance.

1.2: The relationship between PCA and pPCA

Despite the above, PCA is not a probabilistic model for subtle reasons, and this is hinted by the slight discrepancy between our first definition- the optimal linear dimensionality reduction that maximizes retained variance- and our second definition- the basis vectors of the affine subspace closest to the data point. The second definition is subtly wrong as it is too restrictive. Finding the closest subspace produces PCA, but PCA does not necessarily find the closest subspace. This is because the notion of a "closest" affine subspace implies a specific translation of our subspace. In our photograph analogy, if we already have determined the optimal angle to take the snapshot, moving the camera forwards or backwards while maintaining the angle of view will not change the relative positions of subjects in our photograph. (This analogy neglects the fact that a camera in real life takes an image of subjects facing the focal point of the curved lens, and moving the camera into the midst of the subjects will change their relative positioning.) Likewise, any affine subspace

that shares the same basis vectors (or relative orientation) with the closest affine subspace will produce the same projection. Translation of the subspace along the discarded coordinates affects the projection in neither an absolute nor relative way. There are infinitely many affine subspaces that we can find to obtain PCA, though despite this degeneracy they all share the same basis. Of note, if the data is already centered prior to PCA, then we can amend our second definition to state it is the vector subspace (as opposed to an affine subspace) that is closest to the data, and our two definitions will be equivalent.

However, despite resolving the degeneracy of translation with a stricter definition, it is still not a probabilistic model despite the following probabilistic interpretations for PCA. Definition 3 is that PCA is the result of maximizing the likelihood of residuals for a set of points with isotropic noise (i.e., a generalization of linear regression to errors-in-variables that directly follows from Definition 2) (Collins, Dasgupta, and Schapire, 2002). Definition 4 for PCA is that it is the result of fitting a Gaussian and dropping lower dimensions. First, Definition 3 does not consider the latent space distribution; therefore, it is not fully probabilistic. It may seem this can be resolved with Definition 4, with the reasoning that as Gaussians are probabilistic models and PCA results from the process of fitting a Gaussian, PCA should be probabilistic. However, one can show that PCA can be produced by finding the correct rotation matrix to apply to any arbitrary diagonal covariance matrix with distinct entries (meaning the eigenvalues are fixed beforehand) to fit a Gaussian to the data and then obtaining the top eigenvectors of the resulting covariance matrix. Therefore, PCA does not actually describe the process for generating the data, as fitting many Gaussians can produce the same principal components. This emphasizes a key point of confusion about PCA. The axes are indeed interpretable unlike UMAP and tSNE (Chari and Pachter, 2023), and it can be viewed as the result of a probabilistic inference procedure for a subspace. However, they are the result of infinitely many probabilistic models and thus PCA does not define any one probabilistic model. Of the many Gaussian probabilistic models, there are a subset that produce the optimal marginal distribution for maximizing the data likelihood.

One such optimal model is probabilistic PCA (pPCA), which models isotropic noise added to Gaussian latent variables and produces a posterior distribution for inference of the latent points. Another is ours, which assumes isotropic noise in the complementary dimension, in which the maximum likelihood latent points are the projections. The marginal distributions for the observed data are the same (adding

Gaussian noise to a Gaussian variable produces a new variable that is also Gaussian-distributed) and thus the clustering of data and fitting of subspaces is the same in the mixture case.

1.3: The generative model for pPCA

Of note, there may be some confusion as well regarding the isotropic noise in pPCA, and it is particularly important to understanding how the noise is modeled in k-spaces as well, as the marginal distribution is the same. pPCA, as posed by Tipping and Bishop (Michael E Tipping and Christopher M Bishop, 1999a), has Gaussian-distributed latent points perturbed by isotropic Gaussian noise, meaning equal, normally-distributed noise in all directions. However, the data is *not* generated by sampling a latent point from a Gaussian distribution, picking a direction uniformly at random, and then sampling a distance from a Half-normal distribution to perturb the latent point to obtain an observed data point (we will refer to this as the radial Gaussian). The correct generative process is to sample a latent point from a Gaussian distribution, and then for each of the D dimensions of the observed data, perturb the data by distances independently sampled from the same Gaussian (or equivalently from an isotropic D-dimensional multivariate Gaussian). The D-dimensional multivariate Gaussian is rotationally symmetric, as is the radial Gaussian by construction. Both *are* isotropic. However, they produce data with distinct distributions. Consider the 2-D case of each, centered at the origin. Start with the radial Gaussian. If we generate data and want to know the distribution of points lying exactly along the x axis, this would be the subset of points with exactly $\theta = 0$ or 180. As $y = 0$ for this set of points, their x coordinate is exactly equal to the distance from the origin, or the orthogonal residual length from our 0-D subspace. The conditional distribution is Gaussian. Rotational symmetry and our described generative process generalizes this argument to any angle of θ, so since θ is sampled uniformly at random, the marginal distribution of residual lengths would be the same Gaussian. Now consider the bivariate isotropic Gaussian, and again examine the set of points lying exactly on the x axis. The residuals of these points have a Gaussian conditional distribution as well. Their residuals are merely the first of the two independently sampled Gaussian variables that together give us each point. This distribution is rotationally symmetric as well. It would seem that this would produce a Gaussian marginal distribution of residual lengths. However, it does not. If we forgo using a conditional distribution and integrating and instead directly compute our marginal distribution based on independence of the x and y

values of the data, we see that the orthogonal residuals are $\sqrt{x^2 + y^2}$, where X and Y are Gaussian. This is a Rayleigh distribution.

The distinction in generative models (picking a direction at random and then perturbing vs. perturbing independently in orthogonal directions) is subtle. The key issue is the limiting process for obtaining the set of points along the x axis in each scenario. This is known as the Borel-Kolmogorov paradox (Andrey N. Kolmogorov, 1933), and it arose because we must ensure that our conditional density is derived in the same coordinate system as we intend to integrate in to obtain the marginal distribution. In our bivariate Gaussian from pPCA, we picked an x coordinate and a y coordinate separately, simultaneously determining angle and distance. Our conditional density of points along the x axis was obtained as $y \to 0$, i.e., as we squished a rectangle onto the x axis. In contrast, squishing a wedge onto the x axis as $\theta \to 0$ will produce a different conditional density- the Rayleigh distribution. Mentally "integrating" out the angle by using rotational symmetry mixed a conditional density from Cartesian coordinates with an integration in polar coordinates, producing our paradox.

2: Derivation of variance as a function of residuals

In ordinary least squares, minimizing the residual sum of squares is equivalent to maximizing the likelihood of a model with normally distributed residuals. PCA corresponds to multidimensional orthogonal linear regression, meaning that it minimizes the sum of squared orthogonal residuals, or squared distances of points to their projections on the subspace. PCA could also be viewed as maximizing $f_{X_\perp}$ for observing a set of N points in $\mathbb{R}^D$, where the observed points are generated from some latent distribution and then given noise from multivariate Gaussians $\mathcal{N}(\text{proj}_s(\mathbf{x}), \sigma^2\mathbf{I}_D)$ (Collins, Dasgupta, and Schapire, 2002). One can rotate the covariance matrix of these Gaussians such that d axes of our new coordinate system match the d principal components. Thus, the observed point's complementary component of its likelihood is given by an isotropic multivariate Gaussian in the complementary space $\mathcal{N}(\text{proj}_s(\mathbf{x}), \sigma^2 \cdot \mathbf{I}_{D-d})$:

$$f_{X_\perp}(\mathbf{x}|s) = \prod_{a=d+1}^{D} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x_a - proj_s(\mathbf{x})_a)^2}{\sigma^2}}. \tag{3.7}$$

Let $n_i = D - d_i$ for convenience, and define r as the length of the orthogonal residual from $\mathbf{x}$ to its projection on space s. Let j index individual points $\mathbf{x}_j$ and a index

dimensions for a space, such that x_a refers to the value of point $\mathbf{x}$ in the ath dimension of a space's coordinate system. Rotational symmetry allows the likelihood of a point to be written as a function of its residual's length:

$$f_{X_\perp}(r|s) = \frac{1}{(2\pi)^{n_i/2}\sigma^{n_i}} e^{-\frac{1}{2}\frac{r^2}{\sigma^2}}. \tag{3.8}$$

σ_i^2 is the mean variance over the complementary dimensions, as noted in (Michael E Tipping and Christopher M Bishop, 1999a), and we show this below for completeness. Once the spaces have been fitted, the expected variance for each space, σ_i^2, over its complementary dimensions is given by:

$$\begin{aligned}\sigma_i^2 &= \mathbb{E}_{a\in\perp}[(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)^2] - \mathbb{E}_{a\in\perp}[(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)]^2, \qquad (3.9)\\ &= \sum_{j=1}^{N} P(s_i|\mathbf{x}_j)\frac{1}{n_i}\sum_{a=1}^{n_i}(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)^2 - \left(\sum_{j=1}^{N_i} P(s_i|\mathbf{x}_j)\frac{1}{n_i}\sum_{a=1}^{n_i}(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)\right)^2.\end{aligned}$$

In the case of hard assignment where N_i indicates the number of points assigned to space i, this becomes:

$$\sigma_i^2 = \frac{1}{N_i}\sum_{i=1}^{N_i}\frac{1}{n}\sum_{a=1}^{n}(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)^2 - \left(\frac{1}{N_i}\sum_{i=1}^{N_i}\frac{1}{n}\sum_{a=1}^{n}(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)\right)^2.$$

As minimizing the residuals means the second term is 0,

$$\begin{aligned}\sigma_i^2 &= \sum_{j=1}^{N_i} P(s_i|\mathbf{x}_j)\frac{1}{n_i}\sum_{a=1}^{n_i}(x_{ja} - \text{proj}_{s_i}(\mathbf{x}_j)_a)^2 - 0,\\ &= \frac{1}{N_i n_i}\sum_{j=1}^{N_i} r_{ij}^2. \qquad (3.10)\end{aligned}$$

In soft, probabilistic assignment, this becomes

$$= \frac{1}{N\pi_i n_i}\sum_{j=1}^{N} P(s_i|\mathbf{x}_j) r_{ij}^2. \tag{3.11}$$

In some cases, we may also want to assume noise-generating process is the same for all spaces; that is, σ_i^2 is equal for all spaces. While noise within the latent spaces is not explicitly modeled, one can simply subtract the corresponding variance from the covariance matrix in $f_{X_\parallel}(\mathbf{x}|s)$ to make an isotropic $f_{X_{\text{noise}}}(\mathbf{x}|s)$ that includes both the latent and complementary dimensions. This relies on the assumption that the per-dimension estimated shared noise parameter is not greater than the per-dimension

variances within any of the latent spaces, and in our software we enforce this constraint. In their paper introducing probabilistic PCA, where Tipping and Bishop note that the noise parameter corresponds to the average per-dimension variance discarded by a pPCA model, it is known how many dimensions are discarded (Michael E Tipping and Christopher M Bishop, 1999a). Here, where we have a single shared σ^2 rather than a σ_i^2 unique to each space and additionally allow d_i to vary between mixture components, this quantity is unknown because d_i varies between mixture components. σ^2 therefore corresponds to the *expected* variance discarded, and the denominator is the *expected* number of direct observations of our noise process. Let π_{ij} be a shorthand denoting the responsibility of space s_i for point $\mathbf{x}_j$ ($\pi_{ij} = P(s_i|\mathbf{x}_j)$), and use the identity $\sum_{j=1}^{N} \pi_{ij} = N\pi_i$.

$$\sigma^2 = \frac{1}{N \sum_{i=1}^{k} \pi_i n_i} \sum_{j=1}^{N} \sum_{i=1}^{k} \pi_{ij} r_{ij}^2. \tag{3.12}$$

Intuitively, we arrive at this formula because our model specifies each complementary dimension of each space contains a conditionally independent observation of the same noise process controlled by parameter σ^2, so rather than jointly fitting the dimensions of a $\sum_{i=1}^{k} d_i$-dimensional isotropic multivariate distribution, we can treat the problem as a univariate distribution of a single random variable for noise, Z, where each entry $x_{ja}^{\perp}$ in each $\mathbf{x}_j^{\perp}$ from each space is an independent sampling of Z. If we knew which space generated each point or performed hard assignment, we would have $\sum_{j=1}^{N} \sum_{i=1}^{k} n_i 1_{\mathbf{x}_j \in \mathcal{S}_i}$ observations of Z; however, the number of total observations of Z depends on which of the spaces generated each point $\mathbf{x}_j$, as potentially each space generates a vector $\mathbf{x}^{\perp}$ with different dimension. This can also be shown with a maximum likelihood approach for the optimal choice of σ^2 in the M step. We begin with the expected complete data log-likelihood:

$$\begin{aligned} \mathcal{L} &= \sum_{j=1}^{N} \sum_{i=1}^{k} \pi_{ij} \log \left[f_{\parallel}(\mathbf{x}_j^{\parallel} \mid s_i) \cdot f_{\perp}(\mathbf{x}_j^{\perp} \mid s_i) \right], \\ &= \sum_{j=1}^{N} \sum_{i=1}^{k} \pi_{ij} \log \left[f_{\parallel}(\mathbf{x}_j^{\parallel} \mid s_i) \right] + \sum_{j=1}^{N} \sum_{i=1}^{k} \pi_{ij} \log \left[f_{\perp}(\mathbf{x}_j^{\perp} \mid s_i) \right]. \end{aligned} \tag{3.13}$$

We can discard the first term, as it is independent of the second during optimization

and will be 0 when we later take the derivative of $\mathcal{L}$ with respect to σ.

$$\begin{aligned}
\mathcal{L}^{\perp} &= \sum_{i=1}^{k}\sum_{j=1}^{N} \pi_{ij} \log(\prod_{a=1}^{n_i} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{(\mathbf{x}_j - \text{proj}_{s_i}(\mathbf{x}_j))_a^2}{\sigma^2}), \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N} \left[\pi_{ij} n_i \log(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{\pi_{ij}}{2} \sum_{a=1}^{n_i} \frac{(\mathbf{x}_j - \text{proj}_{s_i}(\mathbf{x}_j))_a^2}{\sigma^2} \right], \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N} -\pi_{ij} n_i \frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{N} \pi_{ij} \frac{r_{ij}^2}{\sigma^2}, \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N} -\pi_{ij} n_i \frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{N} \pi_{ij} r_{ij}^2, \\
&= \sum_{i=1}^{k}\sum_{j=1}^{N} -\pi_{ij} n_i (\frac{1}{2} \log(2\pi) + \log(\sigma)) - \frac{1}{2}\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{N} \pi_{ij} r_{ij}^2.
\end{aligned}$$

Setting the derivative to zero,

$$\frac{d\mathcal{L}^{\perp}}{d\sigma} = -\frac{1}{\sigma}\sum_{i=1}^{k}\sum_{j=1}^{N} \pi_{ij} n_i + \frac{1}{\sigma^3}\sum_{i=1}^{k}\sum_{j=1}^{N} \pi_{ij} r_{ij}^2 = 0,$$

and solving yields equation 12.

3: Analytical consideration of the methylation simulation

The hierarchical generative model for simulating methylation data is considerably more complex than the k-spaces model used to fit the data. Here, we study the realized distribution of the observed data. First, we show how centering allows for a convenient decomposition of the data matrix into the sum of (k-1)-D and 0-D signal and noise matrices by eliminating the effects of variation in the underlying site-specific means from the hierarchical model. Both matrices contribute to DMRs, while only the noise matrix contributes to non DMRs, and k-spaces fits these sites separately with different dimension subspaces.

Then, we examine the noise for each site across individuals, which are our features once the matrix is transposed, and find that it is Gaussian and approximately isotropic with regards to individuals. We next turn to the distribution over all entries in our noise matrix where varying σ_j^2 means that it is not Gaussian. However, under the assumption that the noise is isotropic, this distribution is monotonically decreasing with residual length. This result is our justification for approximating it in the E-step with Gaussians of equal variance for all spaces. Finally, we discuss some conditions under which EM with the k-spaces model will begin to fail.

Let i always index rows corresponding to individuals, j always index columns representing sites, and k index cell types. There are N individuals, J sites, and K cell types. In the simulation, $N = 500$, $J = 100,000$, $K = 5$. Our N by J methylation matrix is $\mathbf{O}$, and the centered matrix is $\mathbf{O'}$, but we will cluster on its transpose $\mathbf{O'}^{\mathrm{T}}$.

3.1: Discussion of the generative process

Before examining the effect of centering in the decomposition, we discuss the generative process to set up our analysis. The hierarchical generative model, described in Methods, is reproduced below for convenience.

$$
\begin{aligned}
\mu_j &\sim \text{some distribution},\\
M_{jk} &\sim \mathrm{Norm}(\mu_j, \tau^2),\\
\sigma_j &\sim \mathrm{Halfnorm}(0, \sigma_*^2),\\
X_{ijk} &\sim \mathrm{Norm}(M_{jk}, \sigma_j^2),\\
\mathbf{R_i} &\sim \mathrm{Dir}(\alpha, K),\\
O_{ij} &= \sum_{k=1}^{K} X_{ijk} R_{ik} + \epsilon \text{ where } \epsilon \sim \mathrm{Norm}(0, 0.01), \text{ truncated to } [0, 1].
\end{aligned}
$$

We will show that the precise distributions of μ_j and M_{jk} do not matter. The scale of variation in M_{jk} does matter, however, because technical and biological noise are added. Some distribution on the simplex is needed to acquire individual cell type proportions R_{ik}. The precise distribution is again not critical, but as the variation in cell type proportions from one individual to another decreases, the signal-to-noise ratio shrinks. With no variation in proportions, DMRs will not be detected regardless of τ. We will use a symmetric Dirichlet distribution here for this analysis. In simulation where site-specific variances $\sigma_1^2, ..., \sigma_M^2$ are sampled from Halfnormal$(0, 0.03^2)$, even as the concentration parameter, α, becomes quite large (~ 100), k-spaces clustering still works, although accuracy decreases (Fig. 3.30a). However, it fails around $\alpha = 40$ (Fig. 3.30b) when using σ_j^2 sampled from the empirical distribution in the reference data (Reinius et al., 2012), which is slightly heavier tailed than the Half-normal and contains a very small fraction of extreme variances (Fig. 3.30). We note that the reference data is comprised of only

6 individuals and 5 measured cell types, and that the small subset of sites may be highly variable due to processes different from those governing methylation noise at the vast majority of sites. Eliminating all $\sigma_j > 0.2$ from the empirical distribution (the top 0.29% most variable sites) restores the robustness (Fig. 3.30c). Potential reasons for this will be discussed under "Complementary space likelihood of each point $\mathbf{O}_j^{(\mathbf{n})}$ decreases monotonically with residual length and classification is biased depending on distance from origin."

There is one more factor to consider, truncation of final methylation values to the range [0,1]. Sites with μ_j close to 0 or 1 could be problematic in two ways: 1) sites that are supposed to be DMRs may in practice not be DMRs if τ is too small or if by chance all 5 sampled M_{jk} for that site are either out of bounds or within bounds but still very close to μ_j 2) biological and technical noise for these sites will not be distributed in the same manner as the rest due to the truncation. Moving forward, we will assume that μ_j is distributed in such a way that most sites have μ_j far enough away from the boundaries to avoid the second problem with the distribution of noise, and the first problem is merely a semantic issue of whether a site is a DMR because M_{jk} "could" vary in principle or because the realizations of M_{jk} vary "enough." Note that in this simulation, μ_j and σ_j^2 are independent; if σ_j^2 were smaller near the bounds in real data, that could reduce this issue. We also ignore ϵ for analytical simplicity but acknowledge it could be important in some regimes.

3.2: Centering the data eliminates the effects of variation in site-specific means to allow decomposition into (k-1)-D and 0-D signal and noise matrices

Broadly, we want to examine the covariance structure of sites away from the vector $< 1, 1, 1, ..., 1 >$, or, in other words, correlation in "errors" off the value in the average individual. DMRs (features in $\mathbf{O}^{\mathrm{T}}$) will exhibit a covariance across individuals (dimensions in $\mathbf{O}^{\mathrm{T}}$), while non DMRs will have random noise. However, the underlying mean μ_j varies from site to site, interfering with this, so we center.

Our centered matrix is $\mathbf{O}'$. It is helpful to view $\mathbf{O}'$ as a sum of cell type proportion-driven variation and biological noise. To get there, we begin by decomposing $\mathbf{O}'$:

$$
\begin{aligned}
O'_{ij} &= O_{ij} - \frac{1}{N}\sum_{a=1}^{N} O_{aj}, \\
&= \sum_{k=1}^{K} X_{ijk}R_{ik} - \frac{1}{N}\sum_{a=1}^{N}\sum_{k=1}^{K} X_{ajk}R_{ak}, \\
&= \sum_{k=1}^{K}[X_{ijk}R_{ik} - \frac{1}{N}\sum_{a=1}^{N} X_{ajk}R_{ak}].
\end{aligned}
$$

Define $X'_{ijk} = X_{ijk} - M_{jk}$ noting that $X'_{ijk} \sim \text{Norm}(0, \sigma_j^2)$. Then we obtain the following:

$$
\begin{aligned}
O'_{ij} &= \sum_{k=1}^{K}[R_{ik}(X_{ijk} - M_{jk} + M_{jk}) - \frac{1}{N}\sum_{a=1}^{N} R_{ak}(X_{ajk} - M_{jk} + M_{jk})], \\
&= \sum_{k=1}^{K}[R_{ik}X'_{ijk} + R_{ik}M_{jk} - \frac{1}{N}\sum_{a=1}^{N}(R_{ak}X'_{ajk} + R_{ak}M_{jk})], \\
&= \sum_{k=1}^{K}[R_{ik}X'_{ijk} + R_{ik}M_{jk} - \frac{1}{N}\sum_{a=1}^{N} R_{ak}X'_{ajk} - \frac{1}{N}\sum_{a=1}^{N} R_{ak}M_{jk}].
\end{aligned}
$$

$\mathbb{E}[\frac{1}{N}\sum_{a=1}^{N} R_{ak}X'_{ajk}] = 0$ because $\mathbb{E}[X'_{ajk}] = 0$ and X'_{ajk} and R_{ak} are independent, and in practice, $\frac{1}{N}\sum_{a=1}^{N} R_{ak}X'_{ajk}$ will be close to 0. It is a (randomly weighted) sample mean, and as $N \to \infty$, this converges to 0. Thus,

$$
O'_{ij} \approx \sum_{k=1}^{K} R_{ik}X'_{ijk} + \frac{1}{N}\sum_{a=1}^{N}\sum_{k=1}^{K}[R_{ik}M_{jk} - R_{ak}M_{jk}].
$$

We can then define a noise matrix $\mathbf{O^{(n)}}$, denoting biological noise resulting from variation in cell type methylation profiles between individuals, and a signal matrix $\mathbf{O^{(s)}}$, denoting the signal driven by the variation in cell type proportions:

$$
\begin{aligned}
O_{ij}^{(\mathbf{n})} &= \sum_{k=1}^{K} R_{ik}X'_{ijk}, \\
\text{and } O_{ij}^{(\mathbf{s})} &= \frac{1}{N}\sum_{a=1}^{N}\sum_{k=1}^{K}[(R_{ik} - R_{ak})M_{jk}]. \\
\mathbf{O'} &\approx \mathbf{O^{(n)}} + \mathbf{O^{(s)}}.
\end{aligned}
$$

It is important to note that our k-spaces model instead considers the decomposition $\mathbf{x}' = \mathbf{x}^{\perp} + \mathbf{x}^{\parallel}$ and does not explicitly model noise within the latent subspace, while our generative model for the DMR problem has noise in all directions as modeled explicitly by pPCA. However, this difference between k-spaces components and pPCA components is relevant only for inference of the latent points' positions and not for classification of points, as the resulting marginal distribution is the same Gaussian (and a k-spaces model can be converted to a pPCA model with isotropic noise in both the latent space and complementary space by subtracting the noise from the latent variable's covariance matrix), meaning that classification of points by the spaces will be the same. Our task is then to partition the rows of $\mathbf{O'}^{T}$ into a set of sites that are differentially methylated and therefore are modeled by a (k-1)-dimensional space (signal + noise) and a second set of sites that are not differentially methylated and are better modeled by a 0-D space (pure noise). This could be extended to model other signals, such as a disease status or a batch effect that correlates with another subset of sites.

3.3: The distribution of noise at each site is approximately isotropic

Here, we show columns of our noise matrix $\mathbf{O}_j^{(\mathbf{n})} \sim \mathcal{N}_N(\mathbf{0}, \sigma_j^2 \cdot \mathrm{diag}(||\mathbf{R}_i||_2, ..., ||\mathbf{R}_i||_2, ..., ||\mathbf{R}_N||_2))$ and the degree of anisotropy is therefore bounded by the constraints $\sum_{k=1}^{K} R_{ik} = 1, R_{ik} \geq 0$. Each $O_{ij}^{(\mathbf{n})}$ is the sum of normal random variables, making it normally distributed. Specifically, each $O_{ij}^{(\mathbf{n})}$ is the weighted mean of K independent samples $X'_{ij1}...X'_{ijK}$ from the same normal distribution.

First we consider our rows to understand where the anisotropy arises. Entries within a row $\mathbf{O}_i^{(\mathbf{n})}$ (values for different sites but the same individual) are uncorrelated but not independent. While all X'_{ijk} are i.i.d. for a given j, the realizations of $\mathbf{R}$ for each individual are shared by entries within rows of $\mathbf{O}^{(\mathbf{n})}$, so there are systematic differences between dimensions (individuals) when we cluster on $\mathbf{O}^{(\mathbf{n})\,\mathrm{T}}$. Consider a simple case examining site j in individual A, who has uniform proportions of cell types, and individual B, whose blood is composed of only one cell type. Individual B's variance for noise at site j is σ_j^2, while individual A's is the variance of the sample mean of K realizations with variance σ_j^2: $\frac{\sigma_j^2}{K}$. Individuals with more imbalanced cell proportions will have greater magnitude for entries in their rows of the noise matrix $\mathbf{O}^{(\mathbf{n})}$ because $\mathrm{var}(O_{i1}^{(\mathbf{n})}), ..., \mathrm{var}(O_{ij}^{(\mathbf{n})}), ..., \mathrm{var}(O_{iJ}^{(\mathbf{n})})$ are not independent.

In contrast, entries within a column $\mathbf{O}_j^{(\mathbf{n})}$ (noise values for the same site across individuals), are independent as no information is shared. We can then treat a column as

a realization of a multivariate normal distribution with a diagonal covariance matrix. $O_{ij}^{(\mathbf{n})}$ is the sum of random variables $R_{i1}X'_{ij1}...R_{iK}X'_{ijK}$. $\text{cov}(R_{ia}X'_{ija}, R_{ib}X'_{ijb}) = 0$ for $a \neq b$, because R_{ik} and X'_{ijk} are independent, and $\mathbb{E}[X'_{ijk}] = 0$. (Alternatively, one could observe that while $R_{ia}X'_{ija}$ and $R_{ib}X'_{ijb}$ are not independent— their magnitudes are related, because as R_{ia} increases, R_{ib} is more likely to be smaller— they are uncorrelated, as the sign of X'_{ijk} is independent for each entry and it is symmetric about 0.) Thus, the variance for $O_{ij}^{(\mathbf{n})}$ is the sum of variances for the underlying $R_{i1}X'_{ij1}...R_{iK}X'_{ijK}$. Once the cell proportion vectors $\mathbf{R}_i$ have been realized, $\text{var}(O_{ij}^{(\mathbf{n})}) = \sum_{k=1}^{K} R_{ik}^2 \text{var}(X'_{ijk}) = \sigma_j^2 \sum_{k=1}^{K} R_{ik}^2$ because each R_{ij} is fixed in $R_{i1}X'_{ij1}...R_{iK}X'_{ijK}$. This can be simplified to $\text{var}_{ij} = \sigma_j^2 ||\mathbf{R}_i||_2$, and because $R_{ik} \geq 1 \forall i, k$, we can bound $||\mathbf{R}_i||_2$ to be between 1 and $\frac{1}{k}$.

k-spaces clustering fits a model of isotropic Gaussian noise. Our consideration of entries within individual columns shows that the distribution of noise is approximately isotropic. While the M-step only requires that the principal components of variation correspond to the signal from cell proportion variation, assignment of points to spaces in the E-step could be affected if the variation in the scale of noise between individuals, $||\mathbf{R}_i||_2$, becomes extreme and noise is significantly non-isotropic. However, $||\mathbf{R}_i||_2$ is bounded between 1 and $\frac{1}{K}$, and K is small ($K = 5$) in this study. This non-isotropic noise would be properly modeled by a factor model, but as α grows and the imbalance in cell proportions between individuals reduces, noise will be approximately isotropic. Furthermore, it is the degree of imbalances between individuals' cell proportions $\mathbf{R}_i$ that that matters for non-isotropy rather than the absolute differences between individuals' cell proportions, because that is what drives variation in $||\mathbf{R}_i||_2$. In practice, we do not need the unrealistic assumption that cell types are present in equal proportions (as in the symmetric Dirichlet distribution parameterized by α)— only that cell proportions are imbalanced to approximately the same extent across individuals. Moving forward, we assume noise is approximately isotropic so that we can more easily study the effect of site-specific variances to produce non-Gaussian noise.

3.4: Complementary space likelihood of each point $\mathbf{O}_j^{(\mathbf{n})}$ decreases monotonically with residual length and classification is biased depending on distance from origin

Under the assumption that noise is isotropic with respect to individuals ($c^2 = ||\mathbf{R}_1||_2 = ... = ||\mathbf{R}_N||_2$), $\mathbf{O}_j^{(\mathbf{n})} \sim \mathcal{N}_N(\mathbf{0}, \sigma_j^2 c^2 \mathrm{I})$. Fig. 3.14 shows that we can

approximate the empirical distribution of σ_j fairly well with a Half-normal distribution. Combining equation (8) with a Half-normal prior on σ_j, we get the following distribution over columns: $f_{\mathbf{O}_j^{(\mathbf{n})}}(r) = \int_0^\infty \frac{1}{(2\pi)^{n/2}\sigma_j^n} e^{-\frac{1}{2}\frac{r^2}{\sigma_j^2}} \cdot \frac{2}{\sigma_*\sqrt{2\pi}} e^{-\frac{1}{2}\sigma_j^2/\sigma_*^2} d\sigma_j$, which monotonically decreases with r. Therefore, provided noise is approximately isotropic, it would seem appropriate to use a classification scheme that assigns a point to its nearest subspace. However, there is the caveat that assuming both spaces correctly pass through the origin, any point other than the origin will be closer to the $(k-1)$-dimensional space, so a latent distribution is needed. Thus, we set our noise parameter to be equal between the spaces to exploit the monotonicity rather than a pure, non probabilistic generalization of k-means that assigns points to the closest affine subspace.

To understand the impact of the latent distribution in the $(k-1)$-D space, we focus on the vast majority of non DMRs that are located generally close to the origin because they lack variance from cell type proportions. For assignment in k-spaces clustering to work on the not-linearly separable DMRs and non DMRs, likelihood for non DMRs given by the $(k-1)$-D space must drop sufficiently due to spreading of probability density over the latent dimensions by the latent space Gaussian versus the tighter isotropic noise Gaussian. This also implies DMRs with very low overall variance between cell types ($M_{j1} \approx M_{jk} \approx ... \approx M_{jK}$ and σ_j is small) will be biased towards assignment to the 0-D space. If the EM procedure translates the spaces to be centered at the origin and oriented correctly, and density mixing proportions π_i of the spaces are equal, the switch towards misclassification of points should occur at the intersection between the PDF of the $(k-1)$-D Gaussian of the latent space and the PDF for noise along those same dimensions given by the 0-D isotropic Gaussian. In contrast, non DMRs may be misclassified when the realized σ_j is large by the same argument: beyond this boundary, the more spread out latent distribution provides a higher density in those dimensions. This could imply a limit on classification accuracy for k-spaces clustering as opposed to a specific model designed to detect DMRs.

We hypothesize this balance between spread out latent distribution density of the latent space and tighter isotropic density of the noise space is what produces the observed sudden failure of the model when α becomes large and σ_j is bootstrapped from the heavier-tailed observed distribution, as curiously, sensitivity rather than specificity drops to 0 when the model is initialized with approximately correct spaces and allowed to converge to a local minima (Fig. 3.30b) (For this analysis, the model

was initialized by separately fitting each space to the entire dataset, which provides no input information about the partitioning of sites into DMRs and non DMRs but does provide a reasonable starting point for the basis vectors of the $(k-1)$-D space). Of note, this is not the maximum likelihood solution as assigning all points to the $(k-1)$-D space (sensitivity = 1, specificity = 0) is a higher (or equal) likelihood model than assigning all points to the 0-D space. Removing sites with extreme variance (sites with $\sigma_j > 0.2$, or the top 0.29‰) led to the disappearance of this behavior (Fig. 3.30c).

4: Note on merged qHCR image hues

k-spaces may naively seem to be similar to counting the number of visually identifiable colors when the 4 channels of the image are overlaid. Indeed, examination of Fig. 4a in (Trivedi et al., 2018) shows distinct colors in different regions of the image, where pixel intensities across the 4 channels (*myod1*, *tpm3*, *her1*, *her7*) are mixed together as violet, blue, green, and red, respectively. However, there are two caveats to this. First, and more trivial, is that identifying hues (i.e. by taking a 3 channel qHCR image and assigning the three channels to RGB values) assumes that the lines pass through the origin. In such a case, this could be mathematically represented by hierarchical clustering using the cosine similarity index. However, more critically, even in the 4-channel case, depending on how pixel intensities across channels are mixed to produce colors, different anatomical regions may appear to be of the same hue as one another in one rendering of the data but of different hues in another version of the same image. Thus, even in the 4-dimensional case, our eyes already become unreliable.

5: Note on overfitting

A study reexamining ReFACTor's performance on the GALA II dataset pointed out that ReFACTor was trained and tested on the same data and was overfitting (Zheng et al., 2017). We note that just as ReFACTor experienced large drops in R^2 when the linear regression was done with train-test partitions, so does *k*-spaces (Fig. 3.33) (Zheng et al., 2017). However, the eigenvalues (0.266, 0.227, 0.134, 0.120) of *k*-spaces 4-D latent space (i.e. the latent $\sigma_1 ... \sigma_4$) suggest that the observed signal is largely coming from only be three detectable cell types, mirroring the ground truth cell type variation. The ground truth cell composition data in the GALA II dataset indicates that the two cell types demonstrating severe overfitting, monocytes and basophils, have interquartile ranges from 5.7 to 7.9% and 0.4 to 0.5% respectively,

near or below the limit of detection in simulation (Fig. 3.3b, Table 3.1). In fact, *k*-spaces clustering with only a 2-D latent space rather than a 4-D one performs just as well (Fig. 3.34).

6: *k*-spaces clustering identifies mixed cells in single-cell species-mixing experiments

We also show that *k*-spaces is a useful approach for quantifying the separation in mixtures of cells from different species used to assess the fidelity of single-cell RNA-seq assays. Typically, such mixtures are studied heuristically via a "barnyard plot" comparing the number of unique molecular identifiers (UMIs) that map to the human genome to the number that map to the mouse genome for each barcode. While estimating the rate of doublets or multiplets can be done with manual thresholding, using *k*-spaces, we deconvolve these data by fitting two lines and a point, corresponding to human cells, mouse cells, and mixed cells, which include droplets contaminated by nearby lysed cells as well as doublets. Setting the noise parameter equal for all spaces produces an estimation of a mixed cell rate of 1.1% (mainly doublets), but it underestimates the rate of contamination (Fig. 3.27a). Allowing each space to have its own noise parameter yields a 6.7% rate of mixed cells, including both doublets and other contaminating processes, and an estimated standard deviation of 32 and 47 cross-species UMIs per droplet for pure human and pure mouse cells, respectively (Fig. 3.26a).

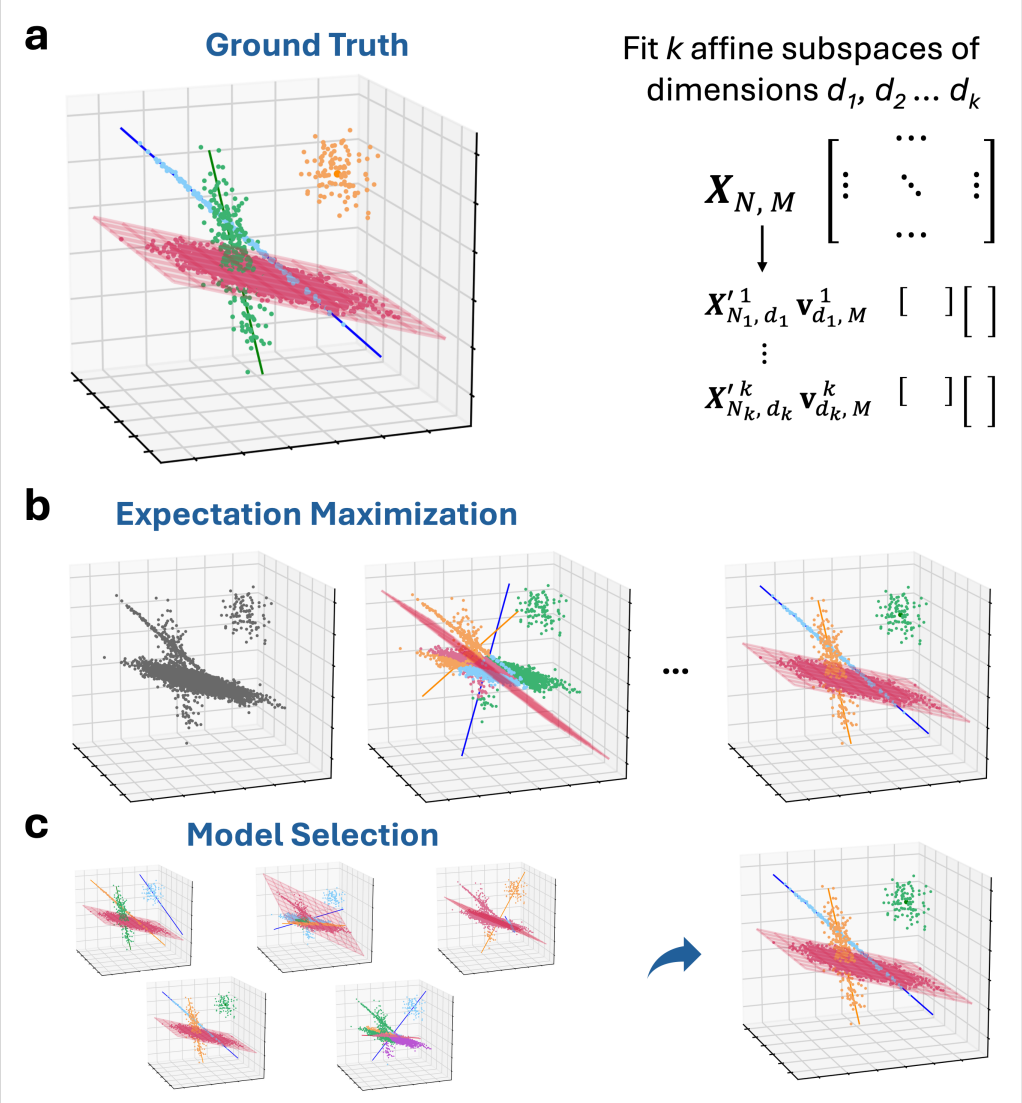


Figure 3.1: **Overview of *k*-spaces clustering on simple synthetic data. a,** Data are generated with isotropic noise from 1 plane, 2 lines, and 1 point. The *k*-spaces clustering algorithm will simultaneously partition and reduce the dimension of the data. **b,** Once a set of spaces is specified to cluster with (here, 1 plane, 2 lines, and 1 point), an expectation maximization algorithm is used to maximize the data likelihood. **c,** Model selection with the Integrated Completed Likelihood (ICL) or Bayesian information criterion (BIC) could be used to select the best model, both between models with spaces of the same dimension or models with different dimensions. Here, BIC was used to select between competing models. From left to right: 1 plane and 3 lines (yielding a correct classification of points but incorrect dimension of clusters); the correct solution from (b); a local minimum with 1 plane, 2 lines, and 1 point; 2 lines and 3 points; and another local minimum with 1 plane, 2 lines, and 1 point.

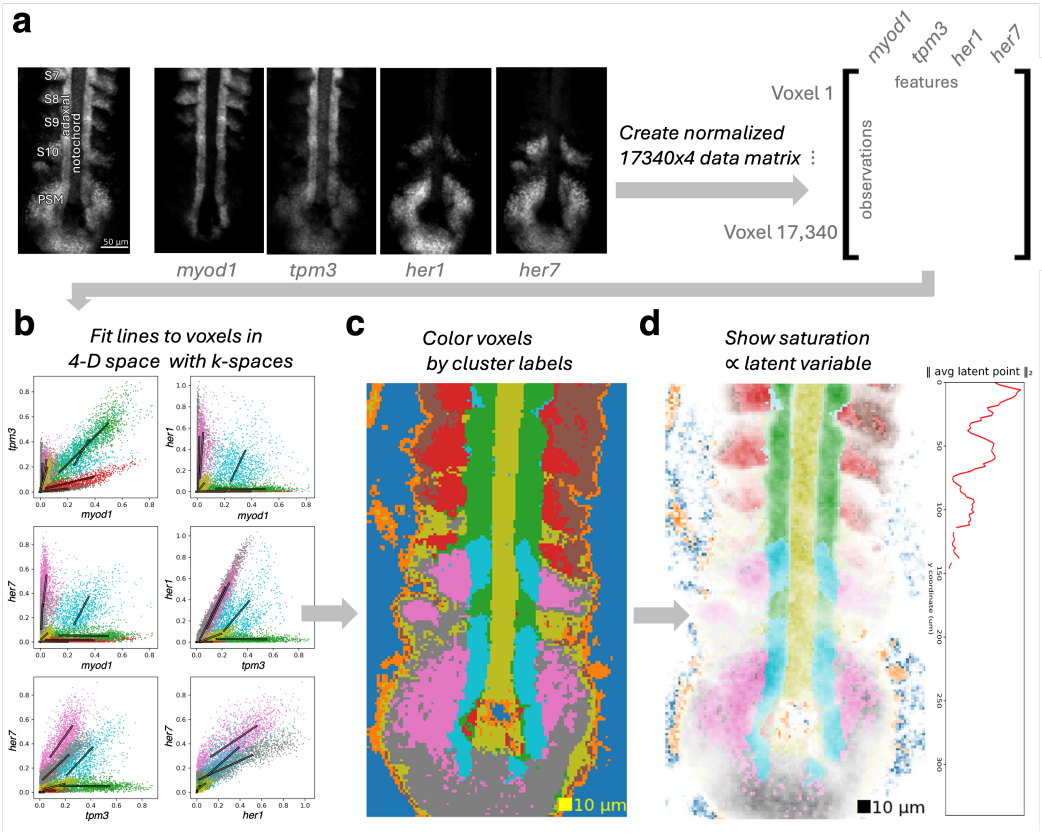


Figure 3.2: **Automated read-out/read-in analysis of a qHCR image of mRNA expression during zebrafish somitogenesis.** **a,** Anatomical features labeled in a grayscale merge (left) of a 4-channel quantitative image of four target mRNAs (middle); 4-channel data was extracted and normalized (right) using the Read-out/ Read-in software with default settings (Trivedi et al., 2018) to produce a 4-D scatter plot of 17,340 voxel intensities. **b,** Clusters were then automatically identified in 4-D using k-spaces to fit lines ($k = 9$, $d_i = 1 \forall i$); these 4-D clusters can be conveniently visualized as six 2-D scatter plots, corresponding to projections of the 4-D data and spaces onto 2-D planes. **c,** Each voxel in the qHCR image is then shaded with its k-spaces expression cluster index to reveal the anatomical distribution of each circuit state (e.g., red program, brown program, etc). 10 μm is a typical cell size. **d,** Visualization of spatial distribution in expression level for gene circuit states using pixel saturation (left; all programs) or average L_2 norm of latent points versus y coordinate (right; red program). See Figs. 3.5 and 3.7 for additional data.

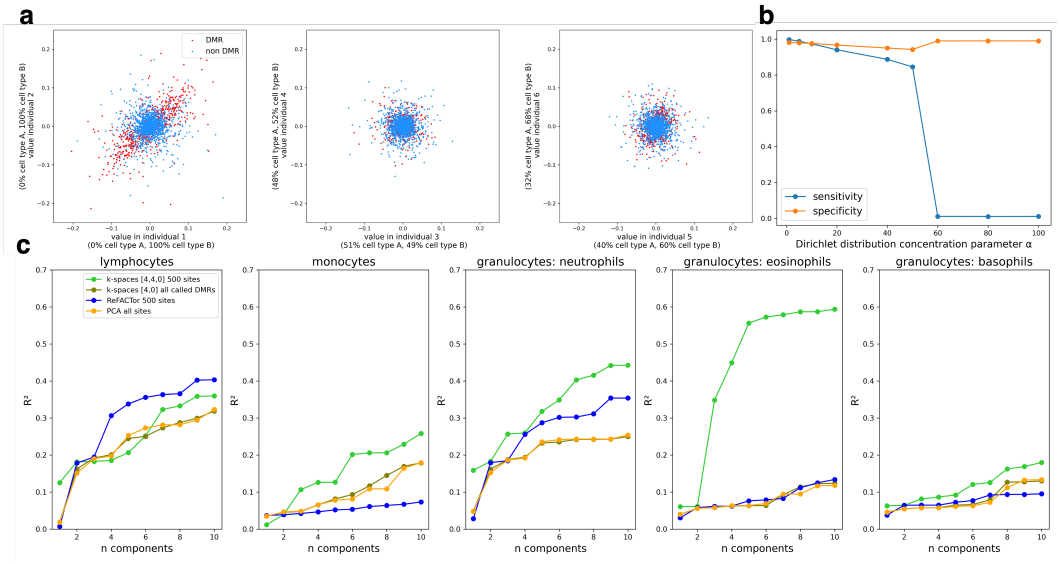


Figure 3.3: **DMR detection in simulated and real data.** **a,** Visualization of centered and transposed data of a toy example with only two cell types in six individuals. In the transposed data space, each axis is an individual while each point is a methylation site. DMRs are correlated between individuals with overlapping extreme cell-type proportions (left) while no cell-type proportion signal is observed in dimensions corresponding to individuals with prototypical cell-type proportions, defined as 50% cell type A and 50% cell type B in this simulation (center). In individuals with mild variation from the average proportions (right), a subtle signal is visible to distinguish DMRs from non DMRs, which is used by *k*-spaces to infer DMRs by examining sites across many individuals in high dimensional space **b,** Sensitivity and specificity of DMR detection in the simulation of five cell types using a symmetric Dirichlet distribution. When cell-type proportions varied by $> \sim$ 1-2%, DMRs were detectable, but when the variation in cell-type proportions becomes too low, the model failed catastrophically. **c,** Regression performance on cell-type proportions in the GALA II dataset.

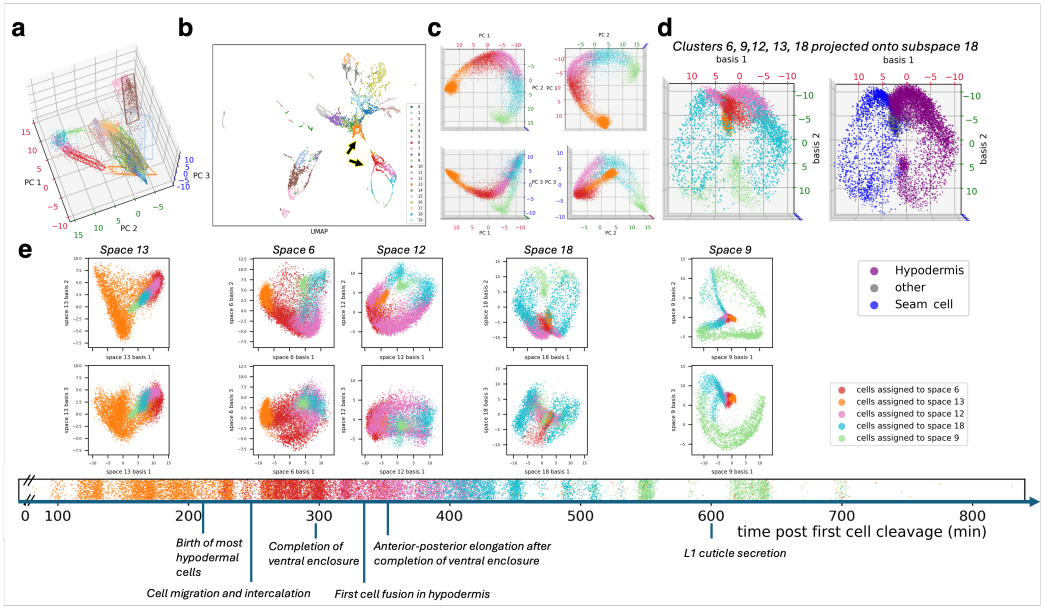


Figure 3.4: **Projection onto multiple subspaces allows visualization of the hypodermis and seam cell trajectories.** **a,** 3-D PCA visualization of developing cells with projections of *k*-spaces's affine subspaces. **b,** UMAP visualization of the same cells from (a), colored by *k*-spaces cluster index. **c,** 3-D PCA of the subset of cells in the orange, red, pink, blue, and green clusters showing a different topology from the UMAP in (b). **d,** Projection of the same cells from (c) onto subspace 18, optimized for viewing gene expression changes in the blue cluster. Alternate coloring by cell type (right) shows that space 18 captures the bifurcation between the seam cell and hypodermis trajectories. **e,** Timeline of key events in epidermis and general embryonic development along with visualization of the cells from (c) and (d) using the dimension reduction yielded by each subspace. Cells are colored by *k*-spaces cluster index, but all cells are projected onto each subspace to visualize the full trajectory. The horizontal strip plot displays the time elapsed from first cell cleavage for cells from each cluster. Points are "jittered" by Gaussian random noise with a standard deviation of 3 minutes for visibility for illustrative purposes only. Sizes of points in (d) and (e) decrease with distance from the 3-D subspace they are projected onto. Thus, points with nearby projections but different sizes are separated in other dimensions of gene expression space. The "unjittered" timeline plot is shown in Fig. 3.25. Plots are reproduced with alternative coloring by cell type and embryo time in Figs. 3.20 and 3.21. Both cell types trace the same path over time in basis 2 vs 3 of space 9 (see Fig. 3.21 for coloring by time), suggesting parallel gene expression changes in both seam cell and hypodermis cell trajectories. There are also seam cell specific (space 9 basis 1 versus 2) and hypodermis cell specific (space 12 basis 1 versus 2) expression changes over time.

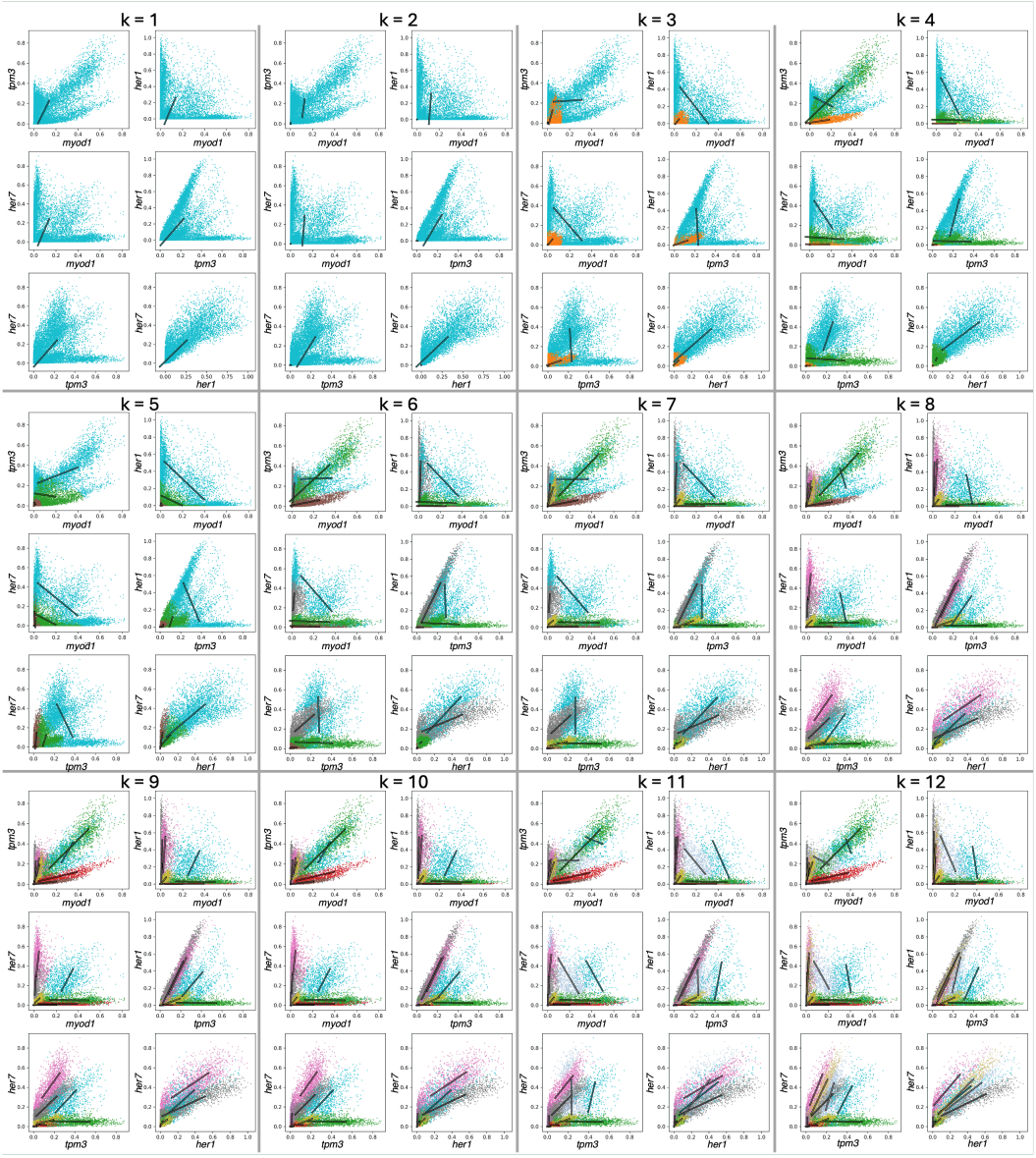


Figure 3.5: **Pairwise voxel intensity scatter plots and projections of fitted lines using k-spaces with $k = \{1, 2, \ldots, 12\}$ on the qHCR zebrafish somitogenesis image of Fig. 3.2.**

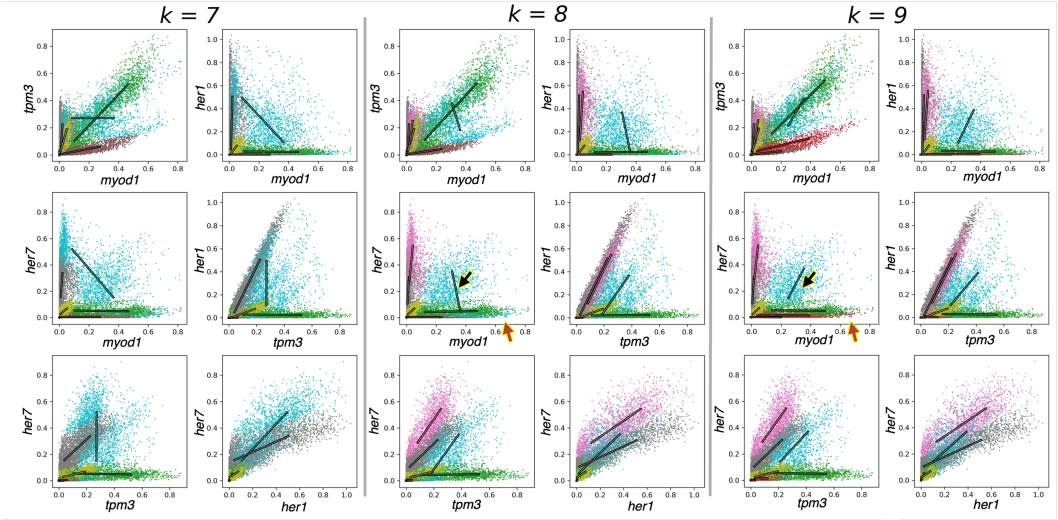


Figure 3.6: **Pairwise voxel intensity scatter plots and projections of fitted lines for $k = \{7, 8, 9\}$ to aid in model selection.** The six pairwise plots for full data, colored by cluster assignment, along with projections of fitted lines as described in Fig. 3.2b. Arrowheads indicate evidence of underfitting in $k = 8$.

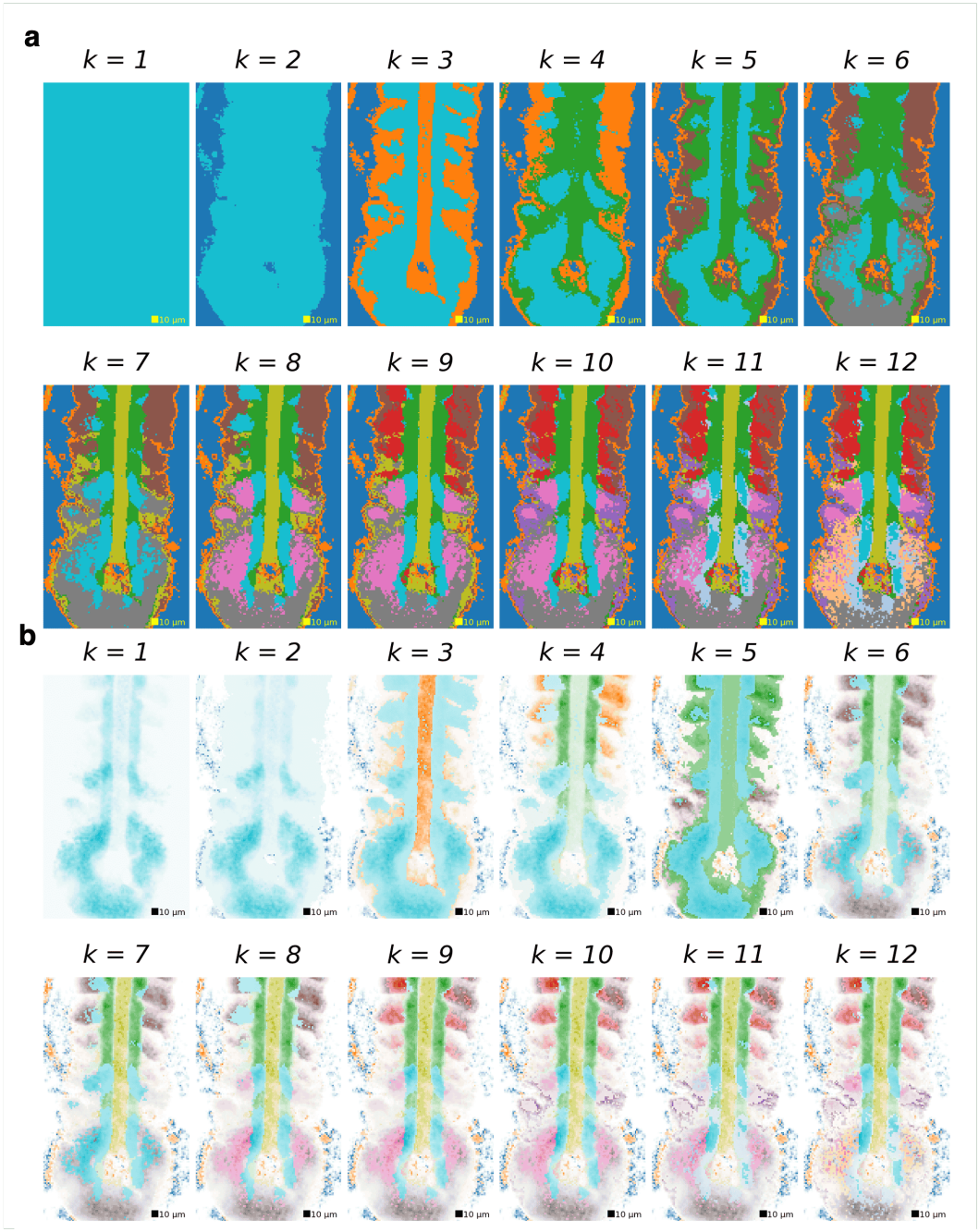


Figure 3.7: **Automatic segmentation of the qHCR zebrafish somitogenesis image of Fig. 3.2 using *k*-spaces with $k = \{1, 2, \ldots, 12\}$.** Each panel is the highest likelihood result out of 100 *k*-spaces runs. *k*-spaces runs for a given value of k do not build off of the $k - 1$ solution. **a,** Each voxel is shaded by its *k*-spaces cluster index.**b,** *k*-spaces clustering with color saturation proportional to latent variable values as described in Methods.

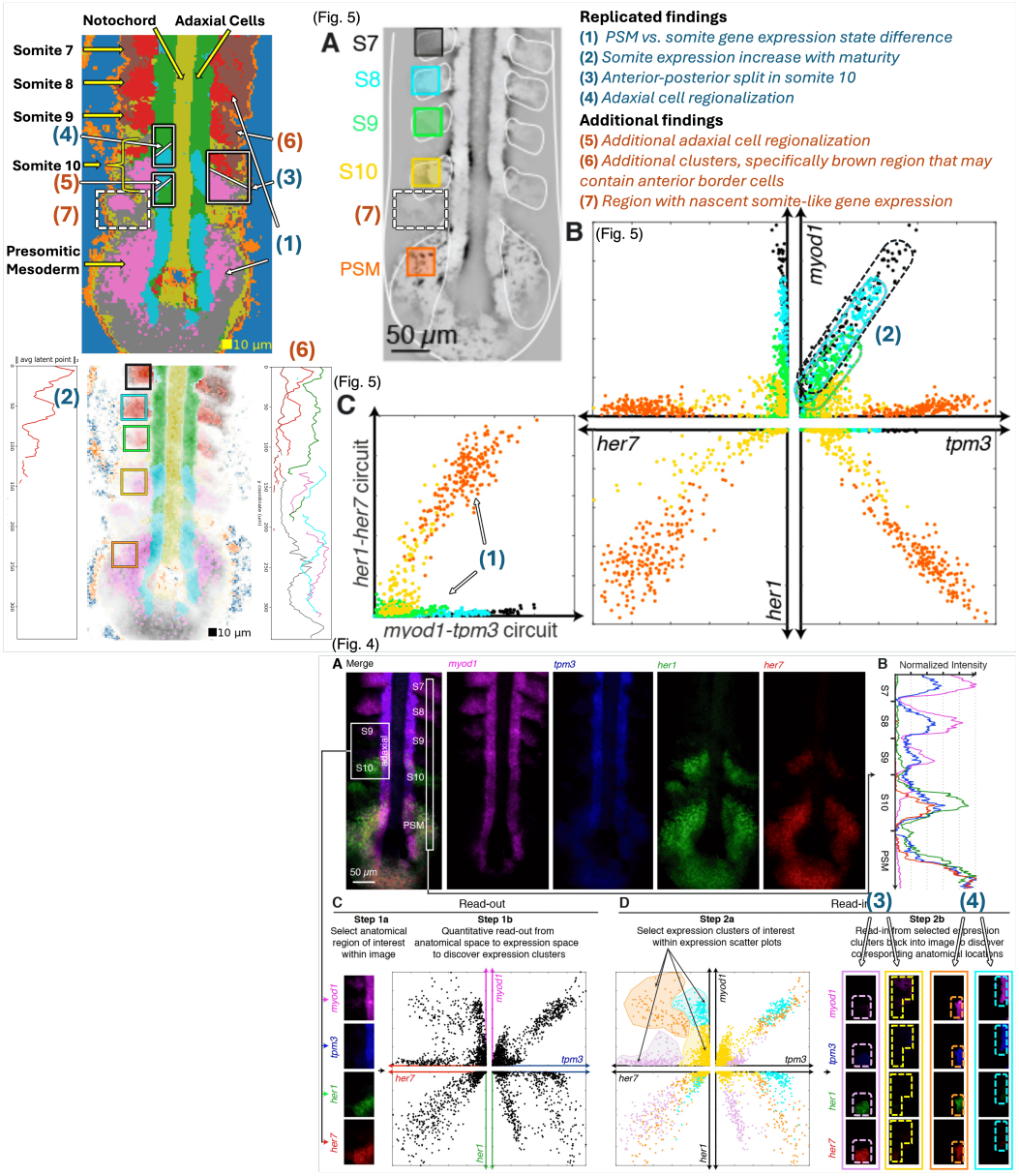


Figure 3.8: **qHCR image analysis: annotated comparison to Figs. 4 and 5 of Trivedi et al., 2018 (Trivedi et al., 2018).** Fig. 4 of Trivedi et al., 2018 demonstrates read-out/read-in: read-out from one ROI encompassing somites S10 and S9 as well as adjacent adaxial cells to pairwise expression scatter plots (panel C), read-in from manually selected expression clusters back into the anatomical context of the image (panel D). Fig. 5 of Trivedi et al., 2018 demonstrates read-out for five ROIs (PSM, S10, S9, S8, S7) with voxel intensities from panel A displayed in the pairwise expression scatter plots of panel B, and the circuit scatter plot of panel C; note that image cropping differs slightly from the present work. Replicated and additional findings are annotated with colored numbers.

Figure 3.8: **Replicated findings: (1)** *k*-spaces identifies pink and gray programs in the presomitic mesoderm (PSM) vs red and brown programs in the maturing somites (S9, S8, S7); in Fig. 5C, the PSM co-expresses the *her1-her7* circuit and the *myod1-tpm3* circuit while the maturing somites (S9, S8, S7) only express the *myod1-tpm3* circuit. **(2)** *k*-spaces identifies that the red program strength increases as somites mature (S9→S8→S7), as shown by increasing saturation and vertical profile of the red latent variable; in Fig. 5B this can be seen in the increasing amplitude and constant slope of the green, cyan, and black expression clusters.**(3)** *k*-spaces identifies anterior/posterior regionalization of the youngest somite (S10) with the posterior region featuring the pink program found in the adjacent PSM and the anterior region featuring the red program found in the adjacent maturing somites (S9, S8, S7); Fig. 4D displays similar regionalization of S10 (anterior: yellow program; posterior: lavender program). **(4)** *k*-spaces identifies corresponding regionalization of the adaxial cells adjacent to the nascent somite (anterior: green program; posterior: cyan program); Fig. 4D displays corresponding regionalization (anterior: cyan program; posterior: orange program). **Additional findings: (5)** *k*-spaces identifies additional segmentation of the adaxial cells between the nascent somite (S10) and the PSM (alternating between the green and cyan programs); Fig. 4 did not include this region in the ROI so this spatial oscillation of adaxial circuit states was not identified. **(6)** *k*-spaces identifies additional expression clusters corresponding to possible detection of distinct cell types within the maturing somites (red vs brown programs) and the PSM (pink vs gray programs); Fig. 4 did not identify these subdivisions of the maturing somites and the PSM using manual cluster selection. **(7)** *k*-spaces detected a structure between the PSM and the left nascent somite (S10) that deviates from bilateral symmetry and features the same pink program featured in the PSM and the posterior region of S10; Figs. 4 and 5 did not include this region in the manually selected ROIs. Figs. 4 and 5 adapted with permission of the Company of Biologists and N.P., senior author of: "Multidimensional quantitative analysis of mRNA expression within intact vertebrate embryos". en. In: Development 145.1, dev156869. issn: 1477-9129, 0950-1991. doi: 10.1242dev.156869. url: https://journals.biologists.com/dev/article/145/1/dev156869/48765/Multidimensional-quantitative-analysis-of-mRNA.

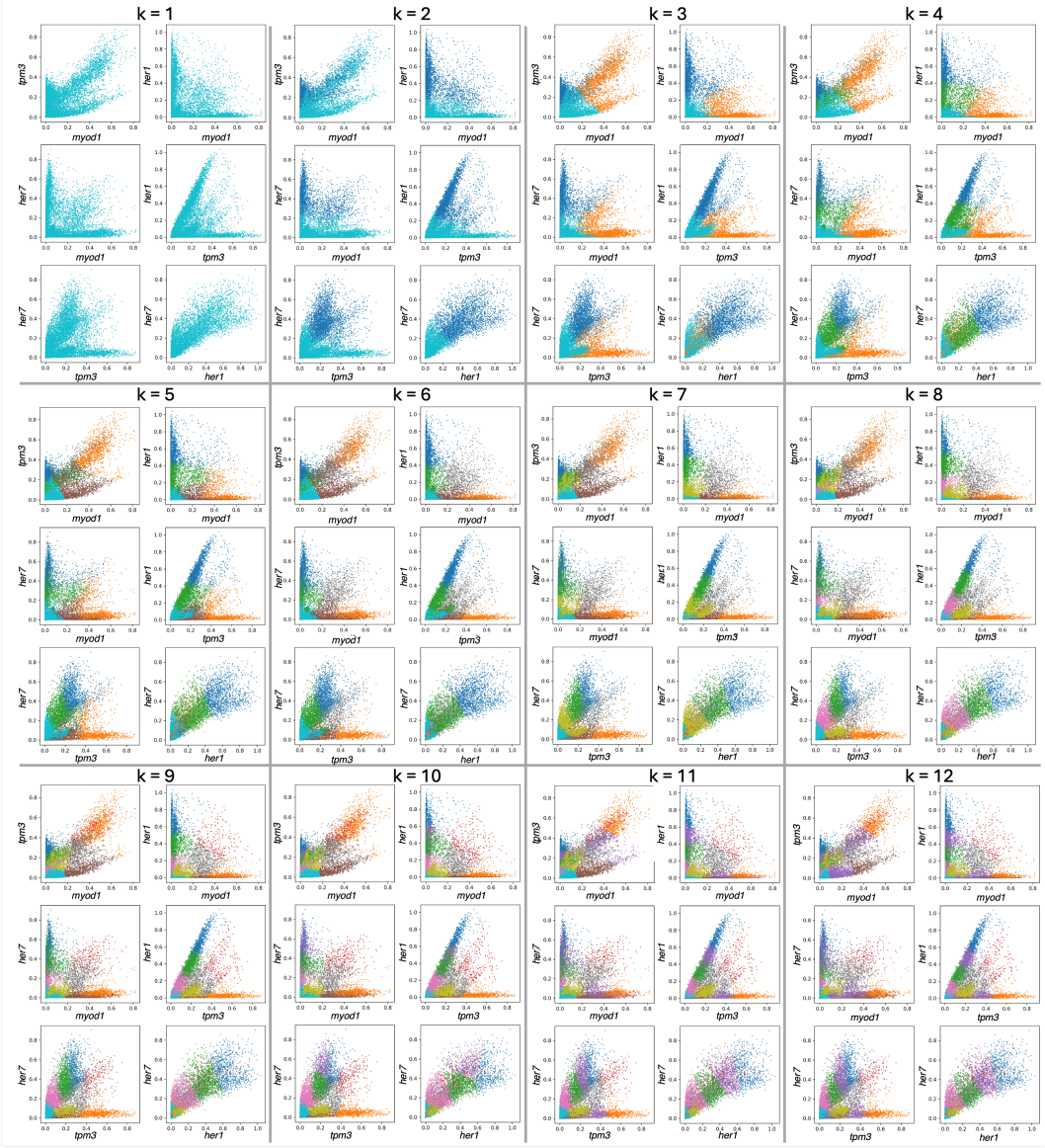


Figure 3.9: **Pairwise voxel scatter plots colored by cluster for k-means with $k = \{1, 2, \ldots, 12\}$.**

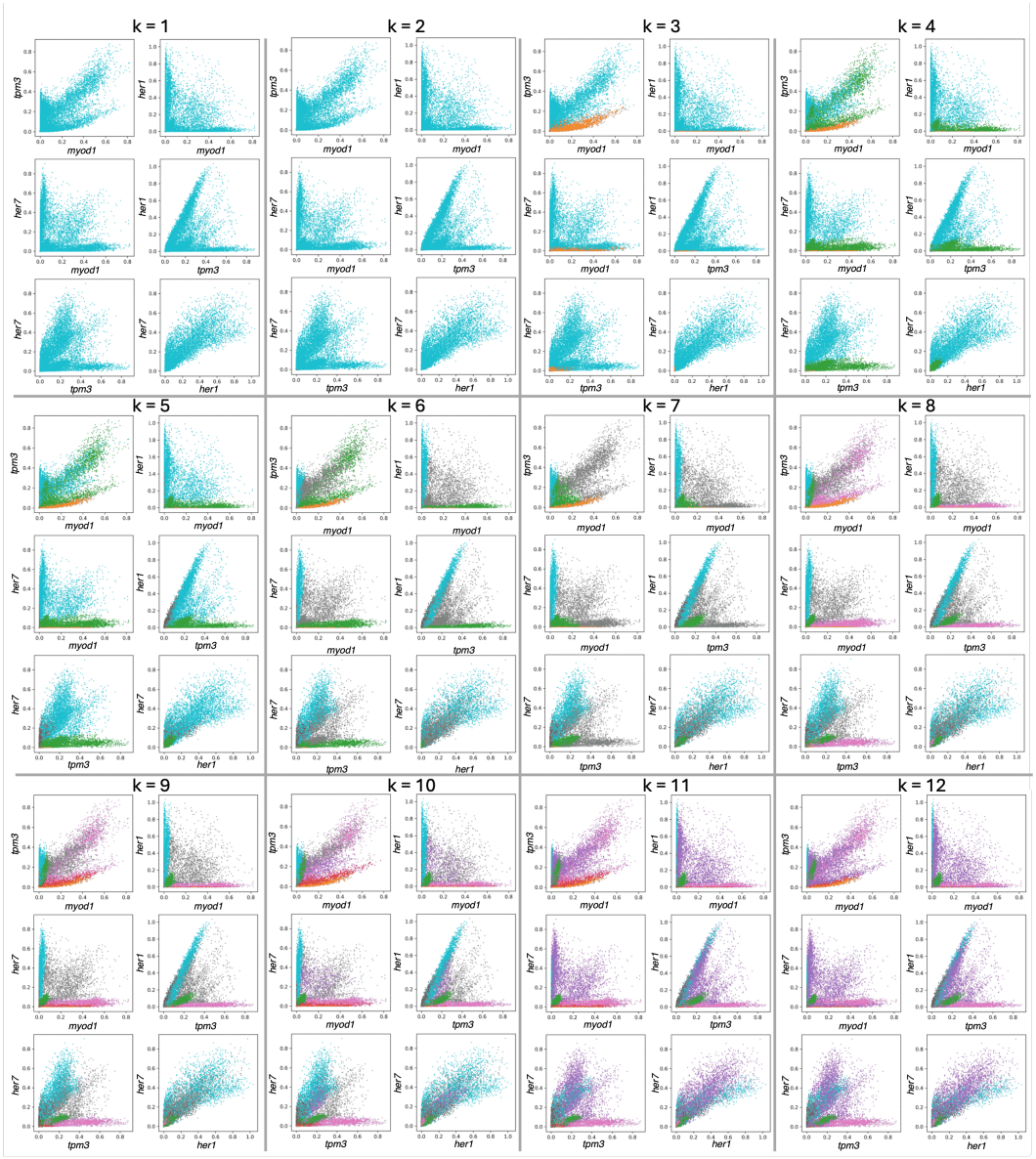


Figure 3.10: **Pairwise voxel scatter plots colored by cluster for GMMs with $k = \{1, 2, \ldots, 12\}$.**

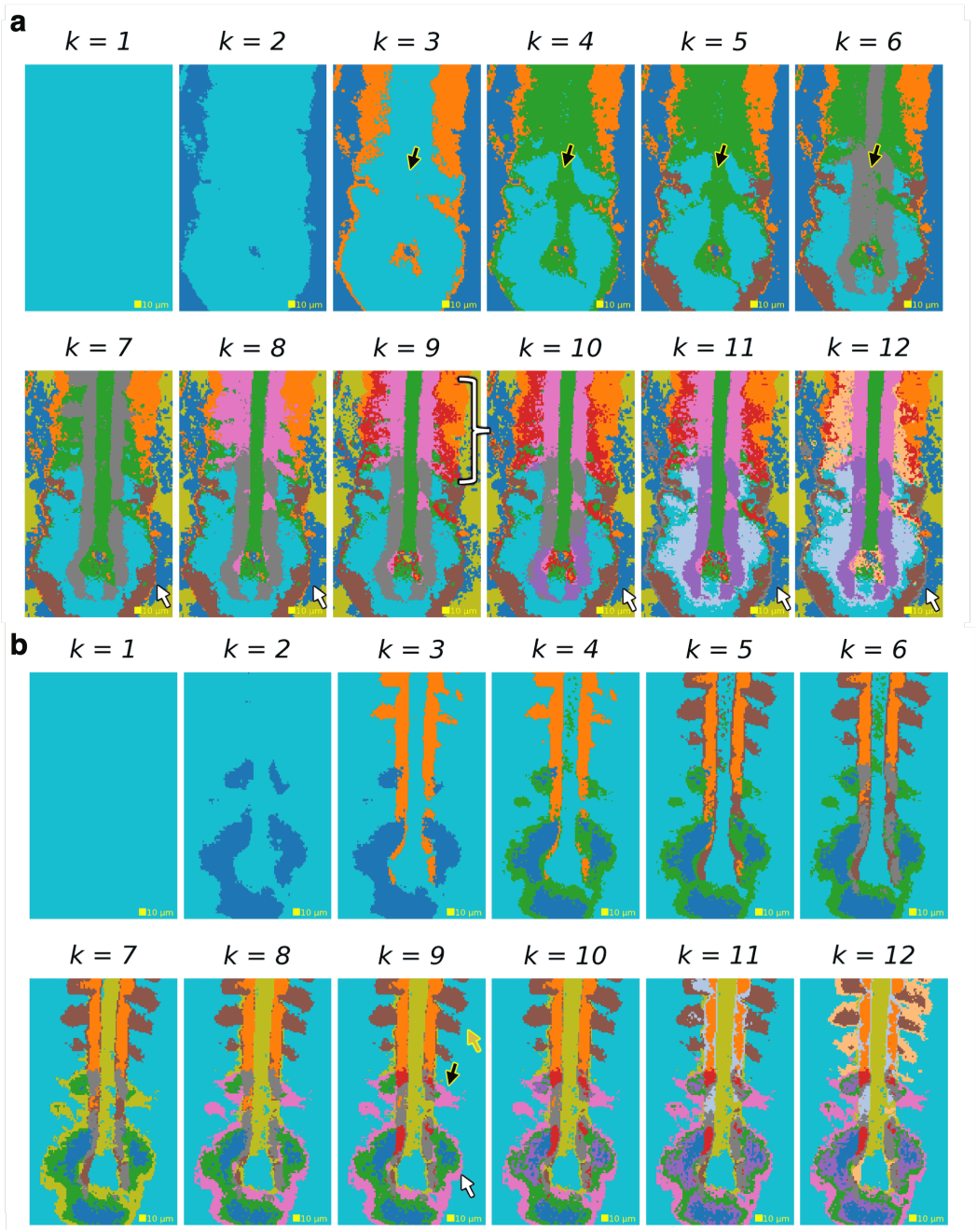


Figure 3.11: **Comparison of Gaussian Mixture Models and *k*-means performance on zebrafish embryo data (comparison for Fig. 3.7a using *k*-spaces).** Each image is the highest likelihood result out of 100 *k*-spaces runs. *k*-spaces runs for a given value of k do not build off of the $k-1$ solution. **a,** GMM clustering. Black arrowheads indicate undetected notochord until $k = 7$, white arrowheads indicate over-segmentation of background space, and white braces indicate somite region which is poorly segmented across most values of k. **b,** *k*-means clustering. Black arrowhead indicates misclassified anterior portion of somite 10 as background, gold arrowhead indicates misclassification of cells on edge of all somites as background, and white arrowhead marks concentric segmentation of presomitic mesoderm

Figure 3.11: likely because the 0-dimensionality of clusters in k-means underfits linear clusters.

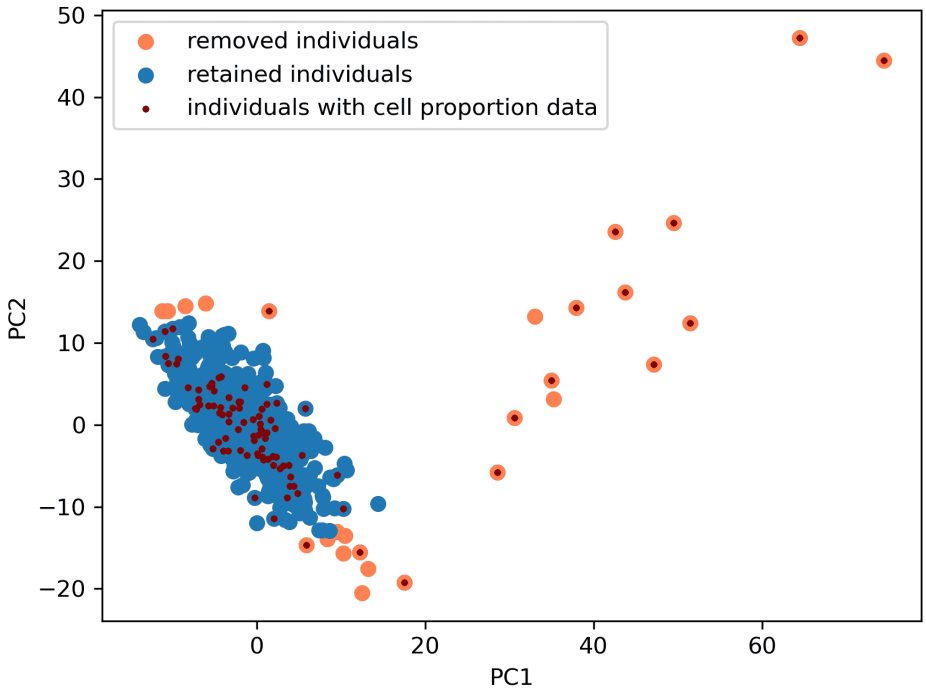


Figure 3.12: **Preprocessing as described in the ReFACTor study**. Individuals more than two s.d. from the mean on the first two PCs are shown in orange.

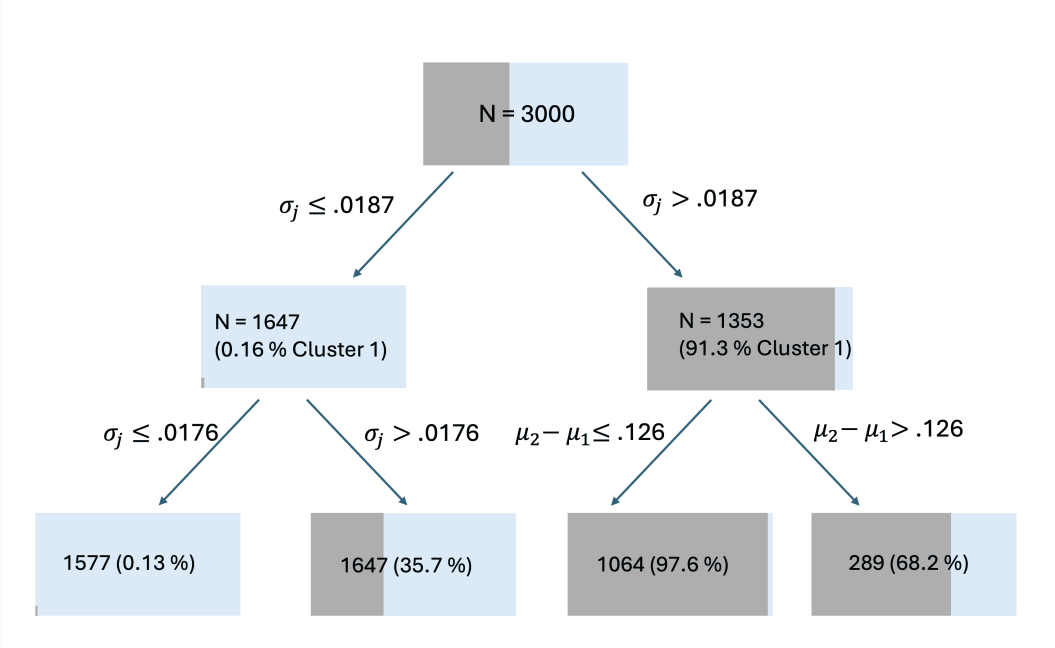


Figure 3.13: **Decision tree trained to predict *k*-spaces DMR calls using ground truth difference in μ_j between cell types and site-specific noise σ_j in simulation.** Simulation was done with the same process described in Methods with only 2 cell types, 500 individuals from a Dirichlet distribution with $\alpha = 1$, 10,000 total sites, 30% DMRs, no added technical noise, and σ_j drawn from a Half-normal with a s.d. of 0.025. $\tau = 0.07$ as in the simulation in the main text. Nodes are shaded gray according to the percentage of sites in that node assigned by *k*-spaces to cluster 1 (called DMRs). The high prediction accuracy at the second level of the tree using only σ_j reveals that poor performance of *k*-spaces in calling DMRs with unconstrained noise parameters was due to fitting the variation in site-specific variance rather than variation in cell type proportions.

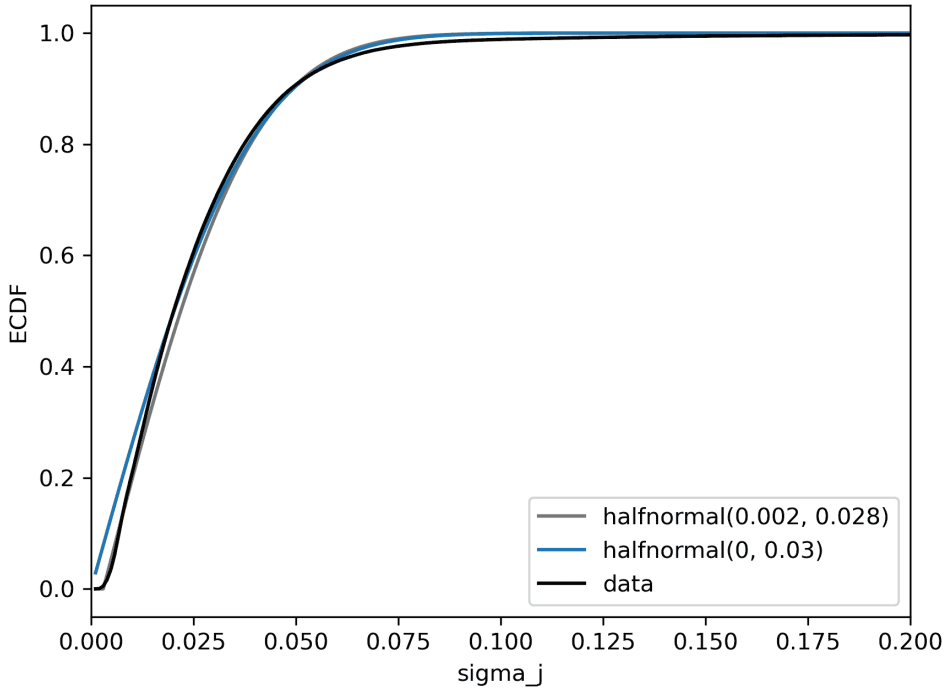


Figure 3.14: **The empirical distribution of site-specific variances in (Reinius et al., 2012) is well approximated by a Half-normal distribution**. We use location = 0,σ = 0.03 in our simulation, as a nonzero lower bound seems unrealistic, but location = 0.002, σ = 0.0028 is shown as well. While the Half-normal centered at 0 more noticeably underestimates the empirical distribution of sample variances for the lowest 20% of sites, it is the slight discrepancy at the tail that is responsible for the lower robustness to large α in the simulation when using the empirical distribution of variances instead of a Half-normal. Removing the 0.29% of sites with $\widehat{\sigma_j} > 0.2$ restores the robustness to sudden failure on α up to at least 100 (Fig. 3.30).

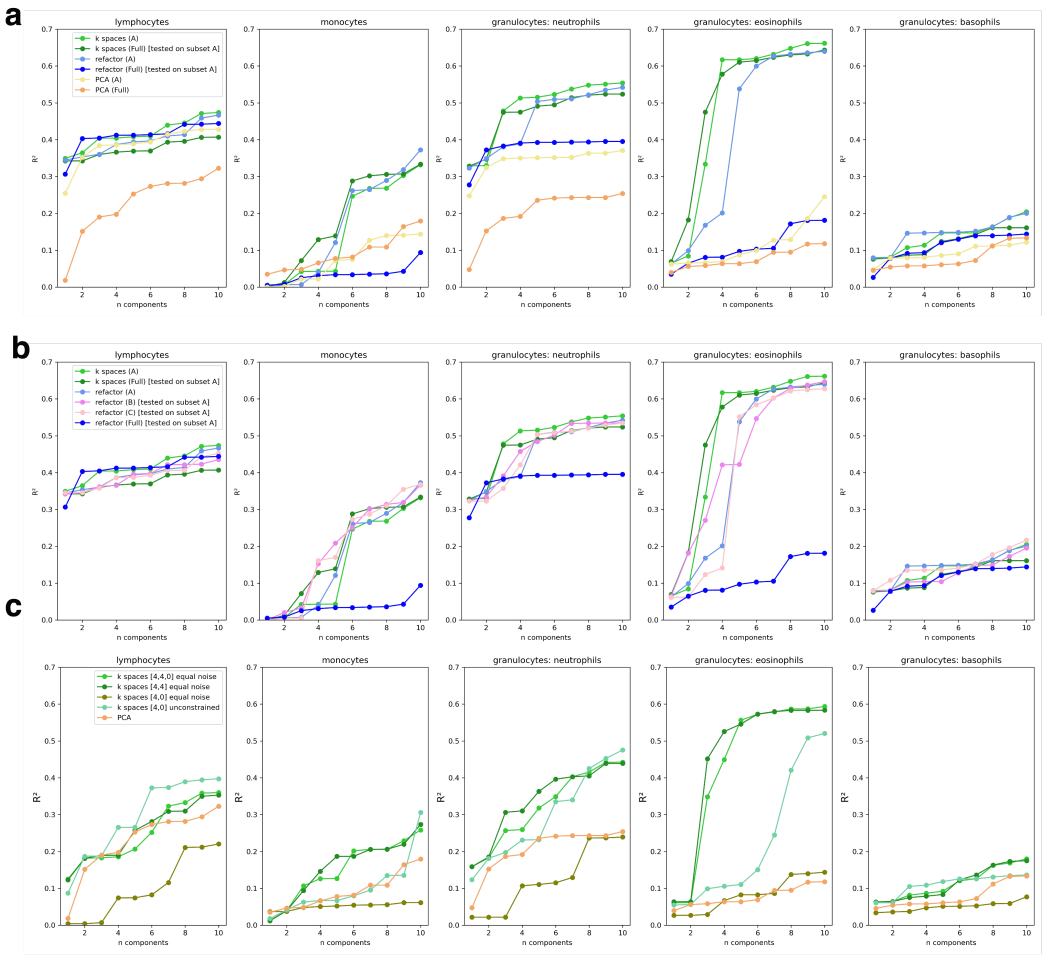


Figure 3.15: ***k*-spaces clustering can perform better than ReFACTor on the full GALA II dataset. a,** *k*-spaces retains similar performance between the preprocessed and the full data (the top 500 closest sites by angle are selected for the preprocessed data in addition to thresholding by the noise of the 4-D structured noise space) while ReFACTor does not. Note that while PCA on the full data and PCA on the preprocessed data are quite different in lymphocytes and neutrophils, ReFACTor has the biggest changes in performance on monocytes and eosinophils, where PCA has no performance change. **b,** Eliminating either preprocessing step (B and C) does not result in significant difference, while eliminating both (Full) leads to failure of ReFACTor. **c,** *k*-spaces clustering with only one 4-D and one 0-D space performs poorly as well but can be corrected with the addition of another 4-D space to model structured noise, as shown in (a), or with replacement of the 0-D space with another 4-D space. Incidentally, *k*-spaces clustering with only one 4-D and one 0-D space but without the shared noise parameter performs well.

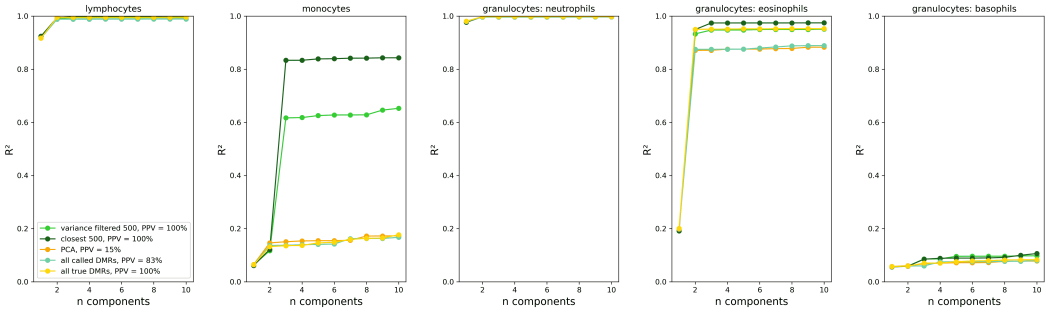


Figure 3.16: **Good DMR detection and good cell type proportion prediction are not equivalent in simulation**. Simulated data is the same used to generate Fig. 3.3b using bootstrapped cell type proportions from GALA II (Esteban Burchard, 2016) and cell type profiles and site specific variances from (Reinius et al., 2012) as described in Methods. PCA and regression was performed as in 3.3b. Despite good performance in simulation (78% sensitivity, 96% specificity, and 83% positive predictive value (PPV), regression using PCA on all called DMRs fails (teal) and is no better than PCA on all data (orange). Regression analysis on the ground truth DMRs (gold) fails to outperform standard PCA as well. Post-clustering filtering strategies (light green, dark green) improve performance.

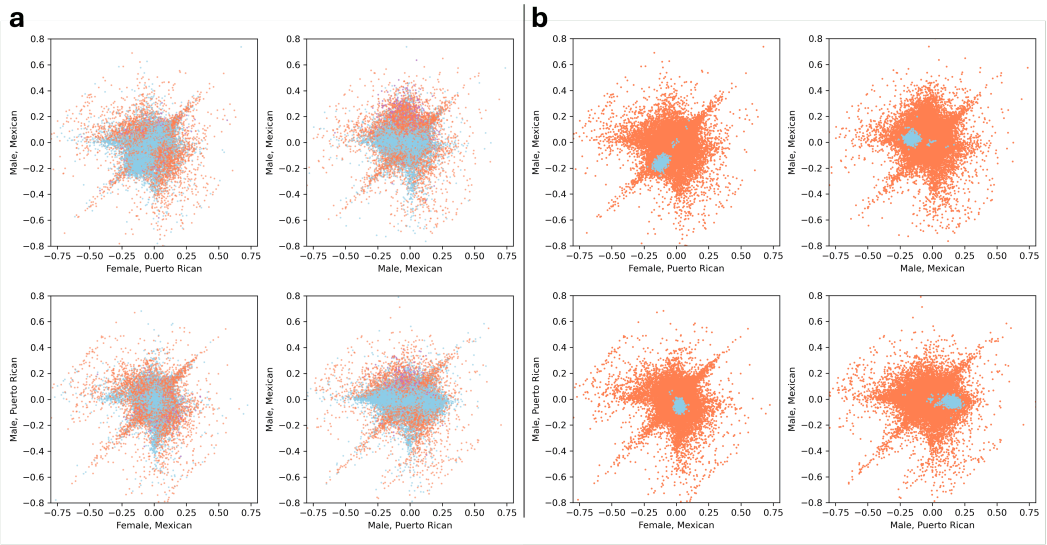


Figure 3.17: **Called DMRs and selected informative DMRs shown in four pairs of individuals from the full GALA II dataset. a,** All sites assigned to the first 4-D latent space (capturing structured noise with respect to the cell proportion problem) are shown in orange. All sites assigned to the second 4-D latent space (containing DMRs) are shown in blue. Remaining sites clustered onto 0-D space are purple. Noise is evidently not isotropic, and a streak persists independent of apparent cell proportion-driven covariation in DMRs. **b,** 500 DMRs (blue) selected from the 4-D DMR space on the basis of having a variance greater than that of the structured noise 4-D space capturing sites perturbed by random noise. All other sites are orange.

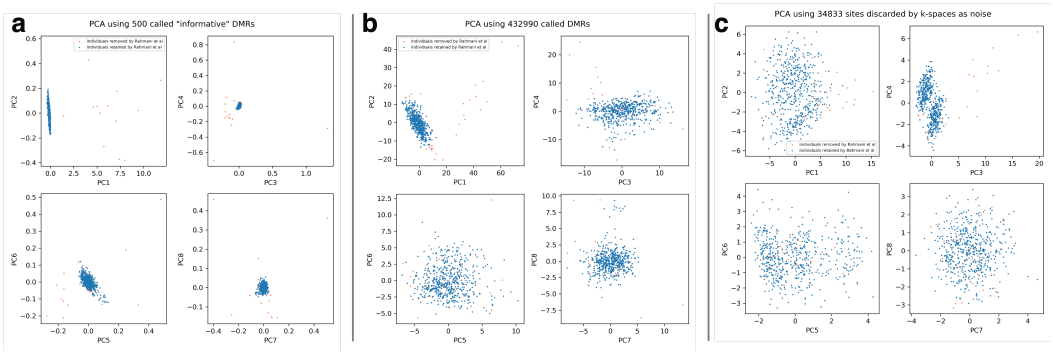


Figure 3.18: **Structure of 'signal' and 'structured noise' discarded by *k*-spaces [4,4,0] on the GALA II data.** Orange points are those removed in (Rahmani et al., 2016) prior to applying ReFACTor. **a,** Top 8 PCs of the data using the selected "informative" DMRs. **b,** Top 8 PCs of the data using all sites captured by the first 4-D subspace (all called DMRs). **c,** Top 8 PCs of the data using all sites captured by the second 4-D subspace intended to capture and discard structured noise.

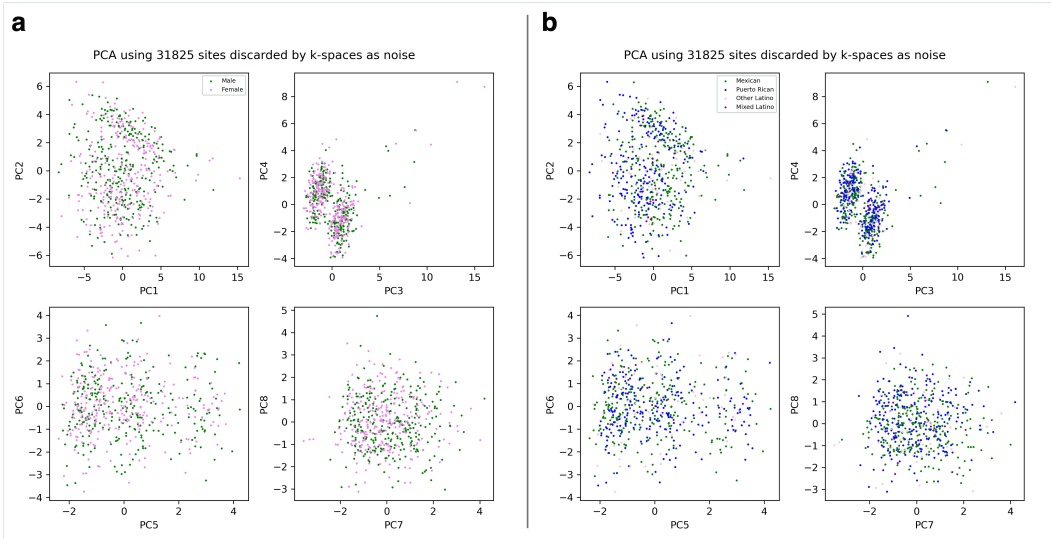


Figure 3.19: **'Structured noise' discarded by *k*-spaces in GALA II does not appear to correlate with available metadata. a,** PCA of samples using the 31,825 sites assigned to the 4-D 'noise' subspace when clustering with two 4-D subspaces and a 0-D subspace colored by sex. **b,** The same data colored by ancestry.

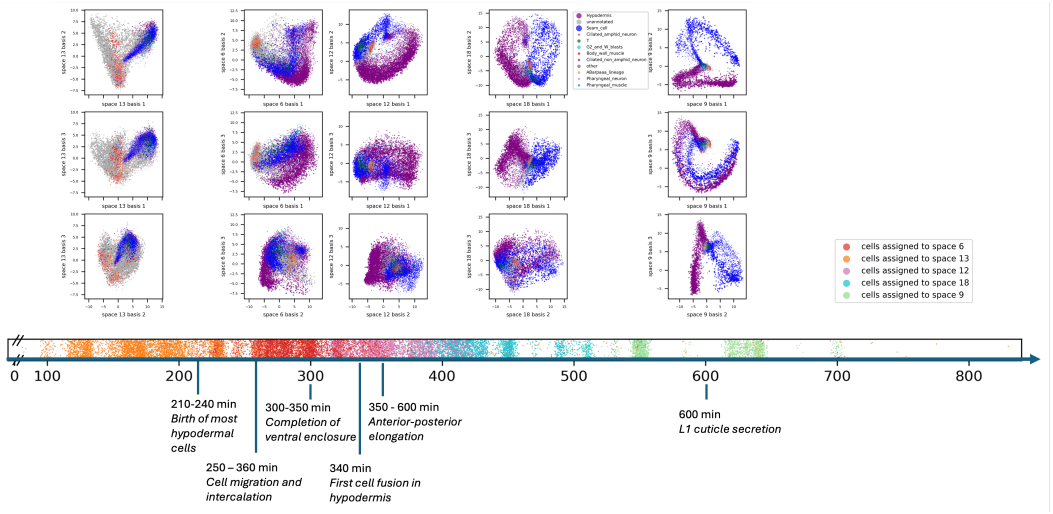


Figure 3.20: **Clusters 6,9,12,13, and 18 in *C. elegans* embryogenesis (Packer et al., 2019) visualized with cell type annotations.** Neurons, hypodermis, and seam cells are all ectoderm derived lineages. Space 13 shows the divergence from neurons early in development, which is is depicted as a break in the orange UMAP cluster in Fig. 3.4b.

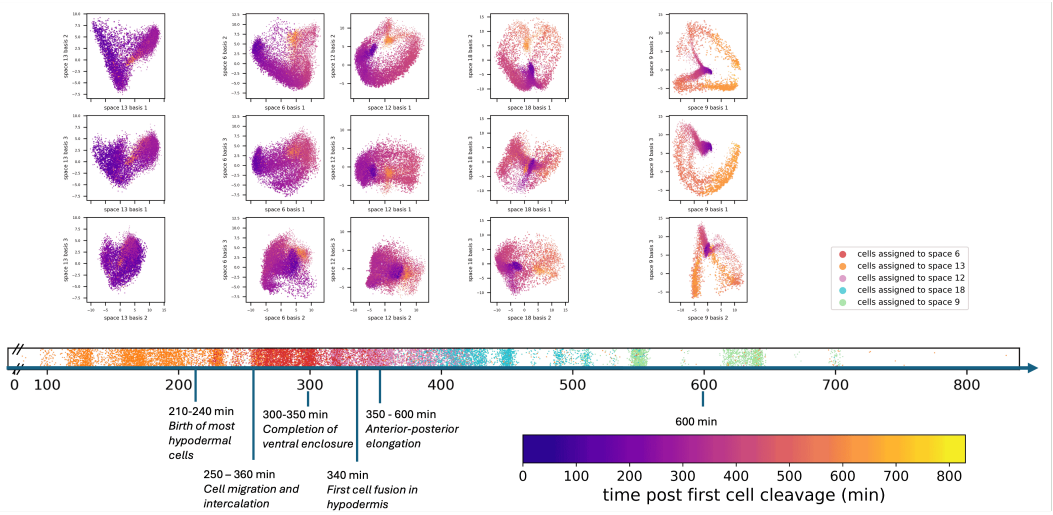


Figure 3.21: **Clusters 6,9,12,13, and 18 in *C. elegans* embryogenesis (Packer et al., 2019) visualized with embryo time.** Early and late cells can be compared within the subspaces to determine if cells return to the same state as their progenitors with respect to the gene expression programs represented in a given axis or if there is continuous progression with maturity.

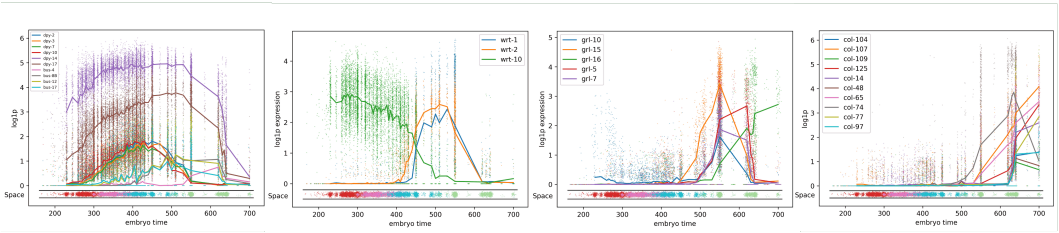


Figure 3.22: **Selected gene families represented in *k*-spaces basis vectors.** While these genes typically strongly influenced the basis vectors (i.e., appeared among the top ranked genes by coefficients) in only one to two spaces, expression levels of the genes can subsequently be tracked across all cells to show when they are expressed. Cells from the orange cluster (cluster 13) are not included as that cluster includes neuronal progenitors.

	Embryonic time (min) 10th-90th percentile	Enrichment analysis interpretation	GO Biological Process	GO Molecular Function	GO Cellular Component	Notes
3-Space 13	130 – 260	Events: end of rapid proliferation and birth of most hypodermal cells				
Basis 1			none	none	none	none
Basis 2			• negative regulation of DNA recombination • nucleosome assembly • protein refolding • chromosome condensation • translation elongation	• nucleosomal DNA binding • translation elongation factor activity • structural constituent of chromatin	• nucleosome	
3-Space 6	260-330	Events: cell migration and intercalation				
Basis 1		remodeling of ECM (lipocalin family)	• tetrahydrobiopterin biosynthetic process • cuticle development involved in collagen and cuticulin based molting cycle • post-embryonic body morphogenesis	• structural constituent of collagen and cuticulin-based cuticle • structural constituent of chromatin	• nucleosome • gap junction • collagen trimer • extracellular region	• lots of dpy genes
3-Space 12	320 – 410	Events: end of intercalation, completion of ventral closure, first cell fusion in hypodermis, elongation				
Basis 1		molt	• notch signaling pathway • molting cycle, collagen and cuticulin-based cuticle	• notch binding	• ER lumen • extracellular region	• warthog • some dpy genes • nhr
Basis 2			• lysine biosynthetic process via aminoadipic acid • tetrahydrobiopterin biosynthetic process • lipid transport • carboxylic acid catabolic process	none	none	• nhr-127 • nhr-270 • lipocalin family
3-Space 18	400 - 500	Events: cell fusions in hypodermis, elongation				
Basis 1		some cuticle genes	none	none	• extracellular matrix/ external encapsulating structure	• cut-2 • cut-5 • nhr-270
Basis 2		some cuticle genes	none	none	• cytosolic large ribosomal subunit	
3-Space 9	550 – 640	Events: cuticle secretion				
Basis 1		cuticle synthesis	• unclassified • *depleted for "cellular process" and "regulation of cellular process"*	• structural component of cuticle • *depleted for "binding"*	• collagen trimer • extracellular matrix • unclassified • *depleted for "cytoplasm" and "nucleus"*	• grl family • wrt family • lots of collagen
Basis 2		cuticle synthesis	none	• structural component of cuticle • glycosyltransferase activity • *depleted for "binding"*	• extracellular matrix • collagen trimer • extracellular region	• turns off in seam cell arm
Basis 3		cuticle synthesis	• *depleted for "cellular process"*	• structural component of cuticle • *depleted for "binding"*	• collagen containing ECM • collagen trimer • extracellular region • *depleted for "cytoplasm" and "nucleus"*	• on/off in both arms • grl inverted sign with respect to Basis 1

Figure 3.23: **Enrichment analysis results on genes with coefficients greater than** 0.05 **or less than** −0.05 **in basis vectors from *k*-spaces.**

3-Space 13 B1	3-Space 13 B2	3-Space 6 B1	3-Space 12 B1	3-Space 12 B2	3-Space 18 B1	3-Space 18 B2	3-Space 9 B1	3-Space 9 B2	3-Space 9 B3
130–260 min	130–260 min	260–330 min	320–410 min	320–410 min	400–500 min	400–500 min	550–640 min	550–640 min	550–640 min
		Remodeling ECM	*Molt*				*Cuticle secretion*	*Cuticle secretion*	*Cuticle secretion*
May be a bifurcation. Increases rapidly but in two directions	Decreases with maturation	Increases with maturation	Off/on for hypodermis, increases slightly for seam	Increases for hypodermis	Bifurcation, increases in seam cells vs hypodermis	Increases	Increase with maturation	Turning off in seam cells	On/off
		Decreases in late stage				Decreases at later time points	Decrease in early time points	Turned on from 400–500 min	Final state is same as early cells
- B0281.5 - C08F1.6 - C50F4.6 - F14B6.3 - F39F10.4 F53B3.5 F54D5.7 - F59C12.4 - K02B12.2 - T05D4.2 - T25E12.6 W04A8.4 Y110A2AL.4 Y65A5A.1 Y7A9D.1 act-4 - aldo-2 - ant-1.1 - cec-8 cey-1 cht-1 clec-196 cpg-20 cpg-24 crt-1 - cyb-3 dpy-14 - dsl-2 dsl-3 - eef-1A.1 eef-1A.2 - eef-1B.1 - eef-1G enpl-1 erd-2.1 fasn-1 - fbxb-37 - fbxb-41 - fbxb-84 - fbxb-91 gmap-1 gst-20 - hil-3 - hil-4 idh-1 igcm-3 lbp-1 lbp-3 lpr-3 - mcm-7 mps-2 noah-1 - pcn-1 pdi-2 - plag-15 pmt-2 - rbm-3.1 ret-1 rml-3 - rsp-1 sec-61.A - skr-7 sqt-3 srlf-1 sto-1 sym-1 tba-1 tbb-2 tct-1 - tir-1 ttr-50 wrt-10	B0281.5 C01G6.3 - C08F1.10 C35E7.5 F14B6.3 F39F10.4 - F44E5.4 - F44E5.5 F59C12.4 R06C1.4 R11A5.3 T05D4.2 - W04A8.4 - Y71F9AL.7 - Y75B8A.18 Y82E9BR.3 - ZK973.11 ahcy-1 aldo-2 - asp-4 atp-6 btb-6 cht-1 clec-196 - cnt-2 ctc-3 cyb-3 dpy-14 dsl-2 dsl-3 eef-1A.1 eef-1B.1 eef-1G epg-2 - fbxb-66 gmap-1 - hil-2 - hil-3 hil-4 - his-24 - hmg-1.1 - hsp-16.2 - hsp-16.41 - hsp-16.48 hsp-3 - knl-2 lbp-3 - mix-1 - nurf-1 pab-1 pdi-2 phdh-1 plag-15 pmt-2 rbm-3.1 skr-7 sqt-3 srlf-1 - sto-1 sym-1 tba-1 tbb-2 tct-1 - tir-1 ttr-50 wrt-10	- C01G6.3 C26B9.3 C29E4.15 - C35E7.5 F48D6.4 F57B9.1 K02E10.4 K09G11.1 - T04D3.5 - T05H4.11 - W04A8.4 - Y37E3.30 Y41D4B.6 - Y65A5A.1 - Y7A9D.1 - act-4 - ant-1.1 aqp-2 - atz-1 calu-1 - cav-1 ceeh-1 - chdp-1 - cht-1 - clec-266 col-121 col-76 cpg-20 - cpi-2 - crt-1 cutl-2 cyn-6 dao-2 dpy-10 dpy-14 dpy-17 dpy-2 dpy-3 dpy-7 dsl-6 - enpl-1 - epg-2 fkb-4 fkb-7 hch-1 - hil-2 - his-24 - his-72 - hmg-1.1 - hrpa-1 - htz-1 inx-12 inx-13 - inx-3 lam-3 - lbp-1 - lbp-3 - lgg-1 - lpr-3 lron-8 mlt-11 mlt-8 - nbet-1 nlp-31 noah-1 noah-2 nsy-4 - pab-1 - pab-2 pcbd-1 pck-2 pgph-3 - pmt-2 qdpr-1 - rbm-3.2 sqt-3 - sto-1 - ttr-50 - ubq-1 ugt-60	B0205.4 B0365.9 B0393.5 C06C3.4 C06E7.88 C26B9.3 C26F1.1 - C47E12.13 C48E7.1 C52G5.2 C53A5.2 F11E6.5 F15B9.8 F26E4.3 F31C3.6 - F33H2.8 F44E2.4 F46F11.7 F46G11.6 - F54D5.7 - F55C9.3 - F55C9.5 F57B9.1 K02E10.4 K08B12.1 M03F4.6 T01D1.8 - T04G9.4 T05H10.3 T19C4.1 - W04H10.6 Y102A11A.5 Y110A2AL.4 Y23H5B.8 - Y37E3.30 Y41D4B.6 Y45F10B.59 Y58A7A.2 - Y65A5A.1 ZC449.2 acn-1 aldo-2 - asns-2 atf-8 best-14 - clec-266 clec-78 col-121 - cpg-20 - cpg-24 - crt-1 - cutl-2 dao-2 dpy-2 dpy-3 dpy-7 - dsl-3 dsl-6 - enpl-1 far-3 - fasn-1 - fbn-1 - gln-3 - gst-20 - hch-1 hgo-1 hhat-1 hil-7 hsp-25 - hsp-3 - idh-1 ifa-3 - lbp-1 lpr-1 lpr-3 lpr-4 lpr-5 lpr-6 lpr-7 nab-1 - nhr-127 nlp-33 pcbd-1 - pdi-6 pgph-3 - pmt-2 ptr-25 - sams-1 spp-13 srm-1 - sto-1 - sym-1 tag-273 ttr-2 ttr-3 - txdc-12.1 unc-52 - wrt-10	C35A5.11 - D1005.2 F18A11.2 F33H2.8 - F48D6.4 F54B11.10 - F57B9.1 - F58H10.1 - H40L08.2 - K01A2.4 - R148.5 T03G6.1 T04G9.4 T19B10.2 - W04A8.4 - Y110A2AL.4 Y23H5B.8 - Y41D4B.6 - Y43F8B.2 Y6D1A.2 - aass-1 atf-8 - atp-6 bus-12 bus-17 - calu-1 - ceeh-1 - cey-1 - clec-196 - cpg-20 - cpg-8 cpi-2 - ctc-1 - cutl-2 - cyn-6 - dao-2 dhs-5 - dsl-3 - dsl-6 - eef-1A.2 - elt-3 - fkb-4 - gmap-1 - hacl-1 - lam-3 lpr-1 lpr-3 lpr-4 lpr-5 lpr-6 - lron-8 - mab-21 - maoc-1 mlt-11 - mps-2 - ndk-1 nhr-127 nhr-270 - nlp-31 nmy-1 - pah-1 - pcbd-1 - pck-2 - pgph-3 - qdpr-1 - rabr-1 - ret-1 - rmd-2 rml-3 spp-13 - srlf-1 sym-2 - ttr-20 - ugt-60 wrt-10	- B0365.9 - B0393.5 C02E7.6 - C06C3.4 - C06E7.88 C29A12.6 C35A5.11 - C48E7.1 - C53A5.2 C54C8.4 D1014.5 - F11E6.9 F13A7.12 - F15B9.8 F15H10.8 F17B5.6 F18A11.2 - F26E4.3 - F31C3.6 F33H2.8 F35G2.5 F37A4.3 - F46E10.2 - F46F11.7 - F46G11.6 F49D11.6 F54B11.10 F58E6.13 - F58H1.2 K01D12.1 - K02E10.4 - K08B12.1 - T01D1.8 T03G6.1 - T05H10.3 T06A1.1 - T19C4.1 - Y102A11A.5 - Y41D4B.6 - ZC449.2 abu-13 act-4 - aqp-2 arf-5 asns-2 - best-14 bgnt-1.2 bre-1 bus-12 bus-17 bus-8B - calu-1 - clec-78 cope-1 - cpg-8 ctg-1 cut-2 cut-5 cyp-42A1 - dao-2 dex-1 dhs-5 - dsl-6 dtmk-1 eef-1A.1 eef-2 - far-3 glf-1 nbet-1 nhr-127 nhr-270 - nlp-33 nmy-1 nstp-3 oac-51 pst-2 - ptr-25 rla-1 rml-3 rocf-1 rpl-12 rpl-7A samt-1 - srlf-1 - sym-1 sym-2 - tag-297 - ttr-2 - ttr-59 - unc-52 - vha-2 wrt-2	C01G6.9 C02E7.6 C02F4.4 C14A4.9 - C29E4.15 C35A5.11 C52G5.2 F01D5.6 F09F9.2 F15H10.8 F18C5.5 F41E3.3 F49D11.6 - F53B3.5 F58H1.2 K01D12.1 K08B12.1 M03B6.3 - R06C1.4 R07E3.6 T03D8.6 T05H10.3 T19A5.3 Y102A11A.5 - Y110A2AL.4 Y37A1B.7 Y54G2A.76 - Y58A7A.2 Y65B4BL.1 Y73E7A.8 ZC449.1 ZK154.1 abu-13 - act-4 - ant-1.1 - asns-2 - atp-6 bus-8B cnc-3 - cpg-24 - ctc-3 - ctg-1 cut-2 dpy-17 - dsl-6 - eef-1A.1 - eef-1B.1 - eef-2 glf-1 grl-15 grl-5 grl-7 - hch-1 - inx-12 - inx-13 mab-7 - mlt-11 - mlt-8 mlt-9 - mps-2 - ndk-1 - nduo-6 - nmy-1 - noah-1 - noah-2 nstp-3 osm-8 - pab-1 - pcbd-1 - pdi-2 - pdi-6 - rla-1 - rml-3 - rpl-12 - rpl-27 - rpl-7A - rps-16 - sma-1 spp-13 sups-1 - sym-1 - sym-2 - tag-273 - tct-1 txdc-12.1 wrt-1 - wrt-10 wrt-2	- C01G6.9 C02E7.7 - C02F4.4 C10F3.7 - C14A4.9 - C26B9.3 - C35A5.11 - C52G5.2 - F09F9.2 F10D11.6 - F15B9.8 - F18C5.5 - F30A10.2 - F33D4.6 F41F3.3 F53F1.4 - F58H1.2 - H03E18.1 H10E21.4 K01D12.9 - K08B12.1 - M03B6.3 - R07E3.6 R07E5.4 R12E2.14 R12E2.15 R12E2.7 R74.2 - T19B10.2 - T26C5.2 - Y102A11A.5 - Y11D7A.5 - Y37A1B.7 Y47D7A.13 - Y65B4BL.1 - Y71G12B.25 - Y73E7A.8 - ZC123.1 - ZC449.1 ZK180.5 - abu-13 - aqp-2 - cbn-1 cht-5 - cnc-3 col-104 col-107 col-109 col-125 col-14 col-48 col-65 col-74 col-77 col-97 cut-3 - cut-6 cutl-14 - dpy-14 - dpy-17 - glf-1 - grl-10 - grl-15 grl-16 - grl-5 - grl-7 - hhat-1 - lpr-3 - mab-7 - mlt-9 - osm-8 ram-2 - rml-3 - spp-13 - sqt-3 - sups-1 - txdc-12.1 - wrt-1 - wrt-2	- C01G6.9 C02E7.6 C10F3.7 C15C6.1 C29A12.6 C35A5.11 - C38C6.3 - C53A5.2 D1014.5 - F09F9.2 F11A10.6 F13A5.5 F13A7.12 F15H10.8 F23H12.7 - F26E4.3 - F30A10.2 F35G2.5 F36F12.1 F37A4.3 F47B8.5 F49D11.6 F58E6.13 - F58H1.2 H23N18.5 K01D12.1 - K01D12.9 - K08B12.1 R05G9R.1 - R12E2.15 - R74.2 T01D1.4 T06A1.1 T08G5.3 W08G11.1 - Y102A11A.5 - Y37A1B.7 Y47D7A.13 - Y64H9A.2 Y75B12B.11 ZK381.2 ZK829.3 abu-13 arf-1 bgnt-1.2 bgnt-1.3 bre-1 bus-12 bus-17 bus-4 bus-8B - clec-230 cnc-3 - col-165 col-65 - col-74 cpg-23 cpg-7 cut-2 - cut-3 cut-5 dex-1 dpy-14 dpy-17 dtmk-1 - grd-5 - grl-16 lips-3 - lpr-7 mlt-9 - nlp-31 - nlp-33 nstp-3 ntp-1 oac-51 pst-2 rocf-1 samt-1 - sqt-3 swt-5 - unc-52 - wrt-1	- B0365.9 - C06E7.88 - C10F3.7 - C18E9.2 C38C6.3 - C48E7.1 C50D2.1 - F11E6.9 - F26E4.3 F26G1.5 F30A10.2 - F31D4.5 F33A8.7 F41F3.3 - F46G11.6 - F52A8.1 F56D3.1 - F58H1.2 - F59B10.5 K01D12.1 - K08B12.1 - M03B6.3 - R07E3.6 R74.2 T05H10.3 - Y102A11A.5 - Y54F10AM.6 - Y54G2A.76 Y64H9A.2 - Y73E7A.8 ZC123.1 - ZC449.2 abu-13 - cav-1 - clec-180 clec-230 - col-125 - col-14 col-165 - col-65 col-74 - col-77 - ctc-3 cut-2 cut-3 - dao-2 dpy-14 - gln-3 gmap-1 grl-10 grl-15 - grl-16 grl-5 grl-7 - hmg-1.1 lgc-22 lips-3 lon-8 - lpr-1 - lpr-3 - lpr-4 - lpr-5 - lpr-6 lpr-7 mlt-9 pqn-32 - spp-13 sqt-3 - tag-297 - ttr-59 - wrt-1 wrt-10 - wrt-2

Figure 3.24: **Gene lists from selected basis vectors of 3-D spaces 6, 9, 12, 13, and 18.** Genes with negative loadings are indicated with a minus sign. Time ranges are the 10th to 90th percentiles of time points represented by cells assigned to each space. Genes involved in morphology, molting, and the cuticle are highlighted in yellow (based on expert knowledge of PWS verified by annotations in WormBase (Sternberg et al., 2024) and a table of molt-related genes in (Lažetić and Fay, 2017)), and genes from the nuclear hormone receptor, warthog, and groundhog-like families are highlighted in blue.

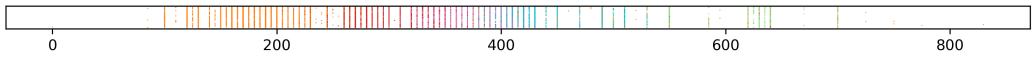


Figure 3.25: **Strip plot of embryo time colored by cluster assignment.** The strip plot of embryo time shown in Fig. 3.4 was randomly perturbed by a Gaussian noise with a standard deviation of 3 minutes for visibility of overlapping points. The unmodified plot is shown here.

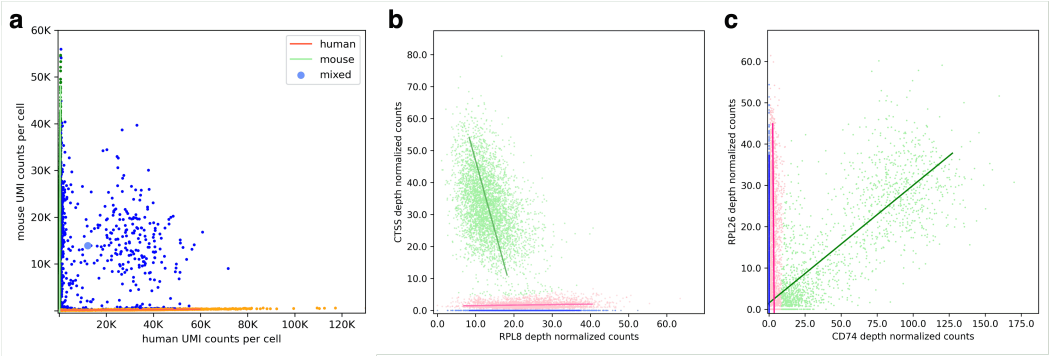


Figure 3.26: **Further applications in RNA sequencing. a,** Analysis of a human-mouse single-cell mixture experiment (10x Genomics, 2024a) with two lines and a point. **b,** Scatterplot of RPL8 and CTSS expression in (10x Genomics, 2022). $R^2 = 0.37$ overall. **c,** Scatterplot of CD74 and RPL26 expression in (10x Genomics, 2024b). $R^2 = 0.113$ overall, implying a little to no linear relationship measured by simple correlation despite clear structure in the data.

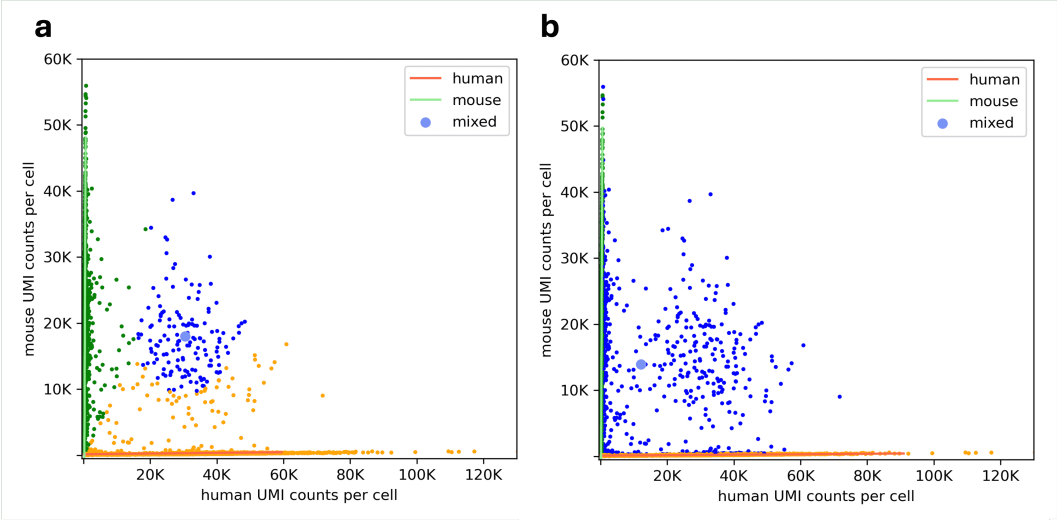


Figure 3.27: **Closest and probabilistic assignment yield different outcomes in a barnyard plot.** Data is from (10x Genomics, 2024a). **a,** Setting the noise parameter equal for all spaces yields an estimate of 1.1% mixed cells. **b,** Probabilistic assignment yields 6.7% mixed cells (reproduced from Fig. 3.26a).

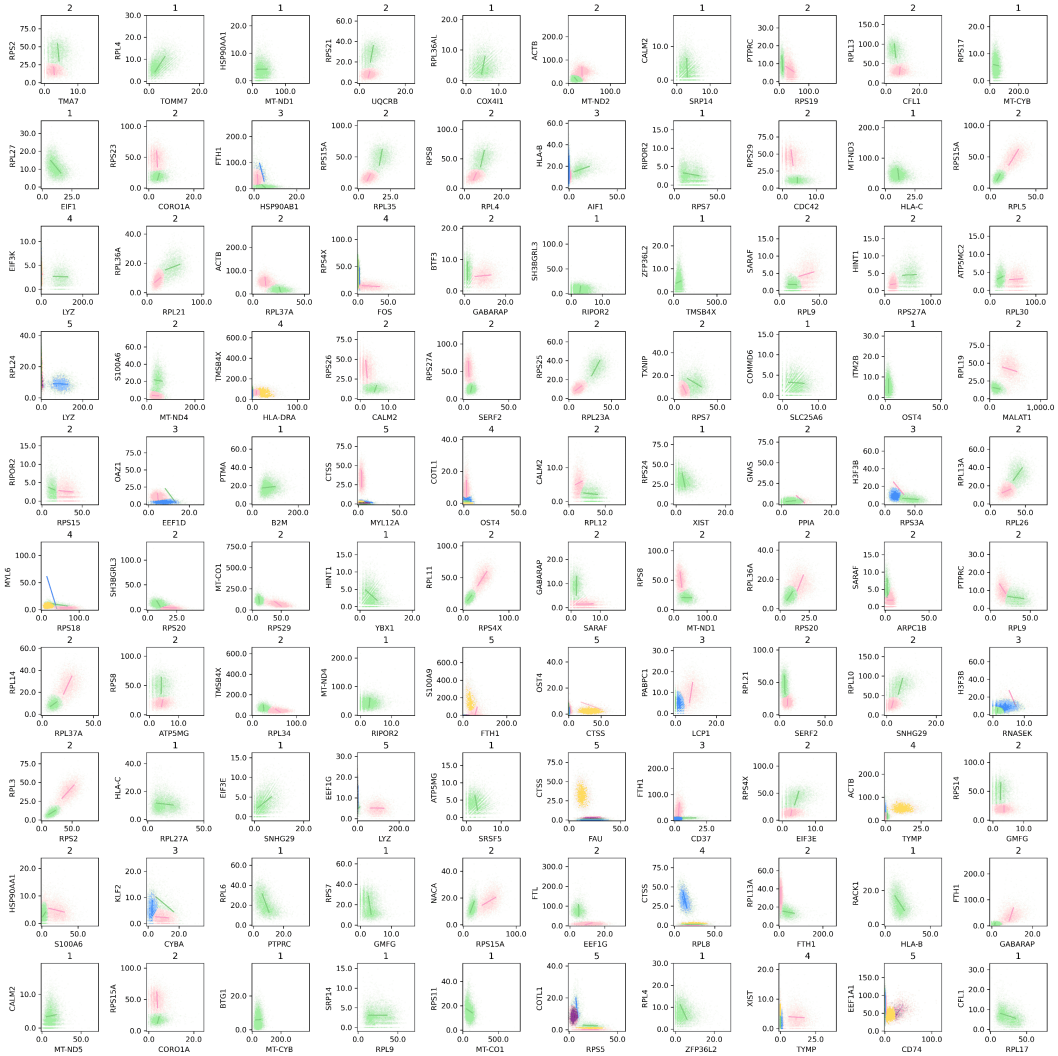


Figure 3.28: **100 random examples of clustering with selection of k via the ICL.** Starting from $k = 1$, k was incremented if the ICL improved, with a maximum of $k = 5$. The subplot title for each gene pair indicates the selected value of k.

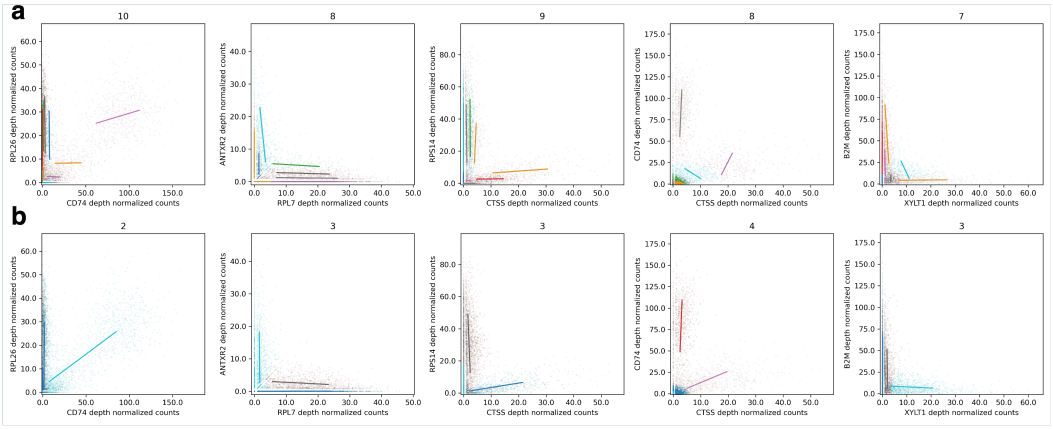


Figure 3.29: **Use of ICL on 5 selected gene pairs from (10x Genomics, 2024b). a,** ICL still leads to some overclustering. Starting from $k = 1$, k was incremented if the ICL improved, with a maximum of $k = 20$. The subplot title for each gene pair indicates the selected value of k. **b,** Automated elbow detection using the kneedle algorithm (Ville Satopaa et al., 2011) on the ICL values from (a) provides better model selection. Subplot title again indicates the selected value of k. Remaining overclustering is due to fitting dropout.

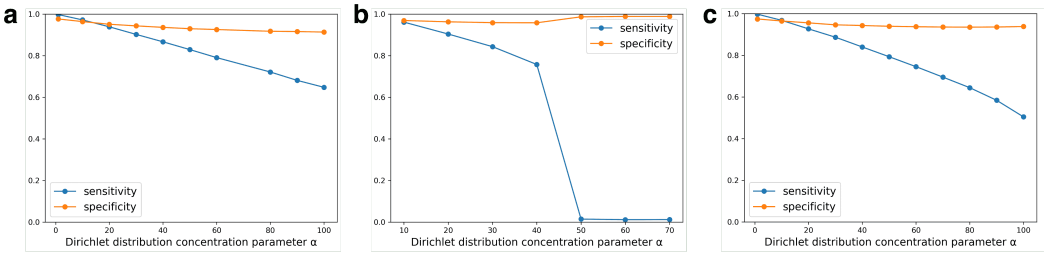


Figure 3.30: **Robustness to variation of Dirichlet concentration parameter. a,** $\sigma_j \sim \text{Halfnorm}(0, 0.03^2)$. **b,** Empirical distribution of site-specific variances (reproduced from Fig. 3.3). **c,** Empirical distribution of site specific variances with the top 0.29% of variances ($\sigma_j > 0.2$) removed.

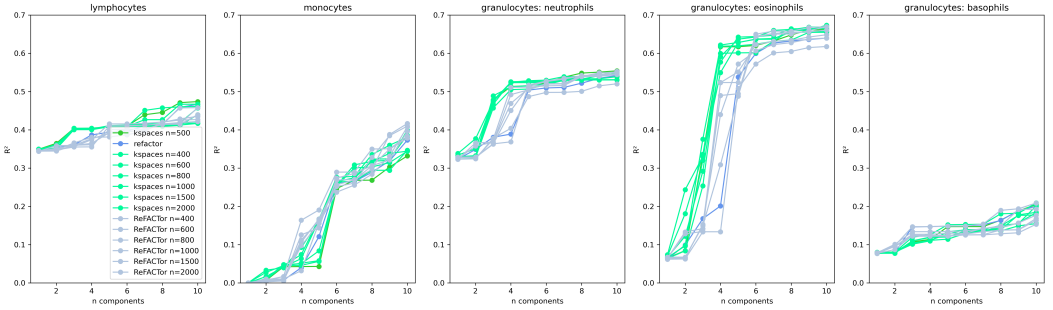


Figure 3.31: **Varying n, the number of selected sites, on the GALA II data.** The threshold for selecting informative sites is the standard deviation of noise, 0.053.

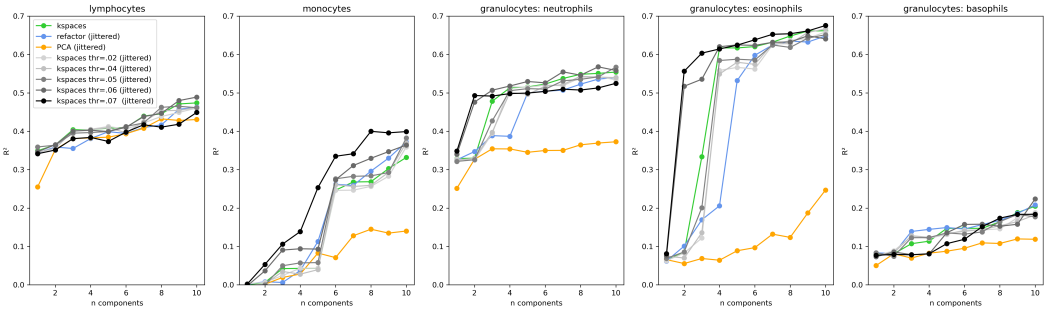


Figure 3.32: **Varying the minimum standard deviation threshold for site selection in *k*-spaces.** n is held constant at 500 sites. Our default choice is the standard deviation of noise, which is 0.053 here.

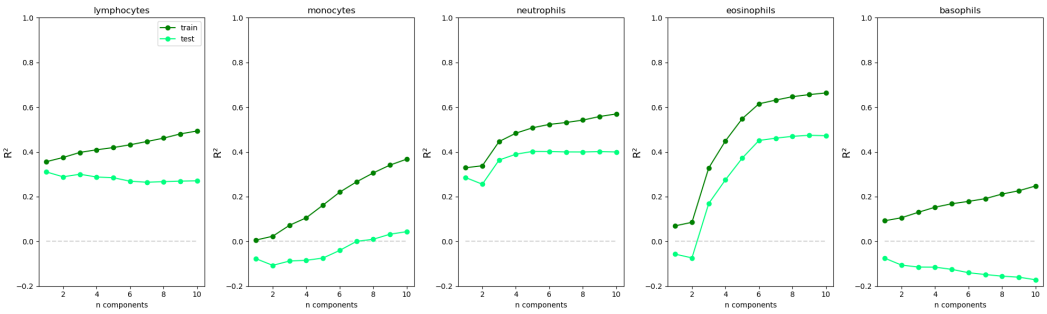


Figure 3.33: **Cross validation as in (Zheng et al., 2017).** R^2 values from linear regression are reported as a function of the number of principal components used, averaged over 50 runs. For each run, the subset of the GALA II data with cell proportion measurements passing the preprocessing filters in (Rahmani et al., 2016) was randomly split into training (n = 50) and test data (n = 30). *k*-spaces clustering with a 4-D and 0-D space with equal noise parameters was run on the training data to obtain informative sites as well as the principal components of those sites in the training data. For the test sets, the same sites and the principal components obtained from the training data were used for regression.

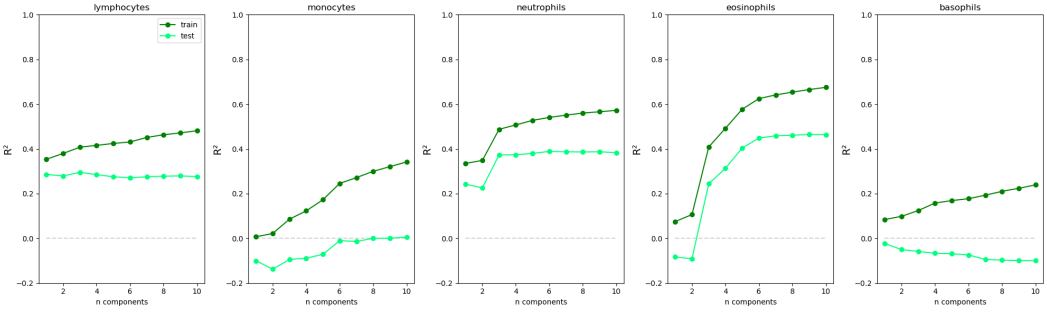


Figure 3.34: **Cross validation as in (Zheng et al., 2017) with a 2-D space for DMRs.** The same procedure was used as in Fig. 3.33 but *k*-spaces clustering with a 2-D space was used instead of 4-D.

Cell Type	10th percentile	25th	50th	75th	90th
Lymphocytes	21.8	27.8	32.7	40.2	43.9
Monocytes	4.7	5.7	6.6	7.9	9.5
Neutrophils	40.0	47.4	54.8	61.3	68.5
Eosinophils	1.2	1.8	4.0	6.5	9.9
Basophils	0.3	0.4	0.5	0.5	0.7

Table 3.1: **GALA II cell type variation n = 80.**

References

10x Genomics (Mar. 2022). *10k Human PBMCs, 3' v3.1 Chromium X (without intronic reads), Single Gene Expression Dataset analyzed using Cell Ranger v6.1.2*. Version 1.0. Accessed on October 2, 2024. 10x Genomics.

– (Apr. 2024a). *10k 1:1 Mixture of Human HEK293T and Mouse NIH3T3 Cells, Chromium GEM-X Single Cell 3', Single Gene Expression Dataset analyzed using Cell Ranger v8.0.0*. Version 8.0.0. Accessed on October 2, 2024. 10x Genomics.

– (Nov. 2024b). *5k Human PBMCs (Donor 4), Universal 3' Gene Expression dataset analyzed using Cell Ranger 9.0.0*. Version 1.0. Accessed on June 23, 2024. 10x Genomics.

Agarwal, Pankaj K and Nabil H Mustafa (June 2004). "k-Means Projective Clustering". en. In: *PODS '04: Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Paris, France: Association for Computing Machinery, pp. 155–165. ISBN: 978-1-58113-858-0.

Andrey N. Kolmogorov (1933). "Erkältung eines Borelschen Paradoxons". German. In: *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Julius Springer, pp. 44–45.

Archambeau, Cédric, Nicolas Delannay, and Michel Verleysen (Mar. 2008). "Mixtures of robust probabilistic principal component analyzers". en. In: *Neurocomputing* 71.7-9, pp. 1274–1282. ISSN: 09252312. DOI: 10.1016/j.neucom.2007.11.029.

Bertoletti, Marco, Nial Friel, and Riccardo Rastelli (May 2015). *Choosing the number of clusters in a finite mixture model using an exact Integrated Completed Likelihood criterion*. en. arXiv:1411.4257 [stat]. DOI: 10.48550/arXiv.1411.4257.

Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (July 2000). "Assessing a mixture model for clustering with the integrated completed likelihood". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.7, pp. 719–725. ISSN: 01628828. DOI: 10.1109/34.865189.

Bishop, Christopher M. and Michael E. Tipping (Mar. 1998). "A hierarchical latent variable model for data visualization". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.3, pp. 281–293. ISSN: 01628828. DOI: 10.1109/34.667885.

Bonnet, Aline et al. (Sept. 2017). "Quaking RNA-Binding Proteins Control Early Myofibril Formation by Modulating Tropomyosin". en. In: *Developmental Cell* 42.5, 527–541.e4. ISSN: 15345807. DOI: 10.1016/j.devcel.2017.08.004.

Bradley, P S and O L Mangasarian (2000). "k-Plane Clustering". en. In: *Journal of Global Optimization* 16, pp. 23–32. DOI: 10.1023/A:1008324625522.

Chari, Tara and Lior Pachter (Aug. 2023). "The specious art of single-cell genomics". en. In: *PLOS Computational Biology* 19.8. Ed. by Jason A. Papin, e1011288. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011288.

Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (Dec. 2018). "Variational inference for probabilistic Poisson PCA". en. In: *The Annals of Applied Statistics* 12.4. ISSN: 1932-6157. DOI: 10.1214/18-AOAS1177.

Choi, Harry M. T. et al. (2018). "Third-Generation in Situ Hybridization Chain Reaction: Multiplexed, Quantitative, Sensitive, Versatile, Robust". In: *Development* 145, dev165753. DOI: 10.1242/dev.165753.

Collins, Michael, Sanjoy Dasgupta, and Robert E. Schapire (Nov. 2002). "A Generalization of Principal Component Analysis to the Exponential Family". en. In: *Advances in Neural Information Processing Systems 14*. Ed. by Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani. The MIT Press, pp. 617–624. ISBN: 978-0-262-27173-8. DOI: 10.7551/mitpress/1120.003.0084.

Devoto, Stephen H. et al. (Nov. 1996). "Identification of separate slow and fast muscle precursor cells in vivo, prior to somite formation". en. In: *Development* 122.11, pp. 3371–3380. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.122.11.3371.

Eckart, Carl and Gale Young (Sept. 1939). "The approximation of one matrix by another of lower rank". en. In: *Psychometrika* 1, pp. 211–218. DOI: 10.1007/BF02288367.

Esteban Burchard (Feb. 2016). *Differential DNA methylation in Latino population*. Version GSE77716. GEO.

Gao, Lucy L., Jacob Bien, and Daniela Witten (Oct. 2022). "Selective Inference for Hierarchical Clustering". en. In: arXiv:2012.02936 [stat]. DOI: 10.48550/arXiv.2012.02936.

Geeves, Michael A., Sarah E. Hitchcock-DeGregori, and Peter W. Gunning (Apr. 2015). "A systematic nomenclature for mammalian tropomyosin isoforms". en. In: *Journal of Muscle Research and Cell Motility* 36.2, pp. 147–153. ISSN: 0142-4319, 1573-2657. DOI: 10.1007/s10974-014-9389-6.

Ghahramani, Zoubin and Geoffrey E. Hinton (May 1996). "The EM Algorithm for Mixtures of Factor Analyzers". en. In: *Technical Report CRG-TR-96-1*.

Gitlin, Andrew et al. (Dec. 2018). "Improving K-Subspaces via Coherence Pursuit". en. In: *IEEE Journal of Selected Topics in Signal Processing* 12.6, pp. 1575–1588. ISSN: 1932-4553, 1941-0484. DOI: 10.1109/JSTSP.2018.2869363.

Hinton, Geoffrey E, Michael Revow, and Peter Dayan (1995). "Recognizing Handwritten Digits Using Mixtures of Linear Models". en. In.

Hollway, Georgina E. et al. (Feb. 2007). "Whole-Somite Rotation Generates Muscle Progenitor Cell Compartments in the Developing Zebrafish Embryo". en. In: *Developmental Cell* 12.2, pp. 207–219. ISSN: 15345807. DOI: 10.1016/j.devcel.2007.01.001.

Houseman, Eugene Andres et al. (Dec. 2012). "DNA methylation arrays as surrogate measures of cell mixture distribution". en. In: *BMC Bioinformatics* 13.1, p. 86. ISSN: 1471-2105. DOI: 10.1186/1471-2105-13-86.

John Aldrich (Apr. 1998). "Doing Least Squares: Perspectives from Gauss and Yule". In: *International Statistical Review* 66.1, pp. 61–81. DOI: https://doi.org/10.2307/1403657.

Kametani, Harunobu et al. (Dec. 2024). "Twisted cell flow facilitates three-dimensional somite morphogenesis in zebrafish". en. In: *Cells & Development* 180, p. 203969. ISSN: 26672901. DOI: 10.1016/j.cdev.2024.203969.

Keenan, Samuel R. and Peter D. Currie (June 2019). "The Developmental Phases of Zebrafish Myogenesis". en. In: *Journal of Developmental Biology* 7.2, p. 12. ISSN: 2221-3759. DOI: 10.3390/jdb7020012.

Kobak, Dmitry and Philipp Berens (Nov. 2019). "The art of using t-SNE for single-cell transcriptomics". en. In: *Nature Communications* 10.1, p. 5416. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13056-x.

Langfelder, Peter and Steve Horvath (Dec. 2008). "WGCNA: an R package for weighted correlation network analysis". en. In: *BMC Bioinformatics* 9.1, p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559.

Laurens van der Maaten and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9, pp. 2579–2605.

Lažetić, Vladimir and David S. Fay (Jan. 2017). "Molting in *C. elegans*". en. In: *Worm* 6.1, e1330246. ISSN: 2162-4054. DOI: 10.1080/21624054.2017.1330246.

Lipor, John et al. (Mar. 2021). "Subspace clustering using ensembles of K -subspaces". en. In: *Information and Inference: A Journal of the IMA* 10.1, pp. 73–107. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaaa031.

Listgarten, Jennifer et al. (June 2012). "Improved linear mixed models for genome-wide association studies". en. In: *Nature Methods* 9.6, pp. 525–526. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2037.

Livingston, Jadzia et al. (June 2022). "Visualization of mRNA Expression in Pseudomonas aeruginosa Aggregates Reveals Spatial Patterns of Fermentative and Denitrifying Metabolism". en. In: *Applied and Environmental Microbiology* 88.11. Ed. by Maia Kivisaar, e00439–22. ISSN: 0099-2240, 1098-5336. DOI: 10.1128/aem.00439-22.

McInnes, Leland, John Healy, and James Melville (Sept. 2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. en. DOI: 10.48550/arXiv.1802.03426.

Mi, Huaiyu et al. (Mar. 2019). "Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0)". en. In: *Nature Protocols* 14.3, pp. 703–721. ISSN: 1754-2189, 1750-2799. DOI: 10.1038/s41596-019-0128-8.

Minka, Thomas (Dec. 2000). "Automatic choice of dimensionality for PCA". In: *MIT Media Laboratory Perceptual Compution Section Technical Report No. 514*.

Nguyen, Phong Dang et al. (Aug. 2014). "Haematopoietic stem cell induction by somite-derived endothelial cells controlled by meox1". en. In: *Nature* 512.7514, pp. 314–318. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13678.

Nguyen-Chi, Mai E. et al. (Oct. 2012). "Morphogenesis and Cell Fate Determination within the Adaxial Cell Equivalence Group of the Zebrafish Myotome". en. In: *PLoS Genetics* 8.10. Ed. by Mary C. Mullins, e1003014. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1003014.

Packer, Jonathan S. et al. (Sept. 2019). "A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution". en. In: *Science* 365.6459, eaax1971. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aax1971.

Pieples, Kathy and David F. Wieczorek (July 2000). "Tropomyosin 3 Increases Striated Muscle Isoform Diversity". en. In: *Biochemistry* 39.28, pp. 8291–8297. ISSN: 0006-2960, 1520-4995. DOI: 10.1021/bi000047x.

Preissl, Sebastian, Kyle J. Gaulton, and Bing Ren (Jan. 2023). "Characterizing cis-regulatory elements using single-cell epigenomics". en. In: *Nature Reviews Genetics* 24.1, pp. 21–43. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/s41576-022-00509-1.

Price, Alkes L. et al. (July 2010). "New approaches to population stratification in genome-wide association studies". en. In: *Nature Reviews Genetics* 11.7, pp. 459–463. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg2813.

Rahmani, Elior et al. (May 2016). "Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies". en. In: *Nature Methods* 13.5, pp. 443–445. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3809.

Reinius, Lovisa E. et al. (July 2012). “Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility”. en. In: *PLoS ONE* 7.7. Ed. by Angela H. Ting, e41361. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0041361.

René Vidal (2010). “A tutorial on subspace clustering”. Johns Hopkins University. URL: https://www.cis.jhu.edu/~rvidal/publications/SPM-Tutorial-Final.pdf.

Robert J. Adcock (Mar. 1978). “A Problem in Least Squares”. In: *The Analyst* 5.2, pp. 53–54. DOI: https://doi.org/10.2307/2635758.

Robin L. Plackett (Dec. 1949). “A Historical Note on the Method of Least Squares”. In: *Biometrika* 36.3/4, pp. 458–460. DOI: https://doi.org/10.2307/2332682.

Roweis, Sam (July 1998). “EM Algorithms for PCA and SPCA”. en. In: Denver, Colorado, USA, pp. 626–632.

Schulte, Samuel J., Mark E. Fornace, et al. (2024). “HCR Spectral Imaging: 10-Plex, Quantitative, High-Resolution RNA and Protein Imaging in Highly Autofluorescent Samples”. In: *Development* 151, dev202307.

Schulte, Samuel J., Boyoung Shin, et al. (2024). “Multiplex, Quantitative, High-Resolution Imaging of Protein:Protein Complexes via Hybridization Chain Reaction”. In: *ACS Chem. Biol.* 19.2, pp. 280–288.

Schwarzkopf, Maayan et al. (2021). “Hybridization Chain Reaction Enables a Unified Approach to Multiplexed, Quantitative, High-Resolution Immunohistochemistry and in Situ Hybridization”. In: *Development* 148.22, dev199847. ISSN: 0950-1991, 1477-9129. DOI: 10.1242/dev.199847.

Stellabotte, Frank et al. (Apr. 2007). “Dynamic somite cell rearrangements lead to distinct waves of myotome growth”. en. In: *Development* 134.7, pp. 1253–1257. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.000067.

Sternberg, Paul W et al. (May 2024). “WormBase 2024: status and transitioning to Alliance infrastructure”. en. In: *GENETICS* 227.1. Ed. by V Wood, iyae050. ISSN: 1943-2631. DOI: 10.1093/genetics/iyae050.

Su, Ting and Jennifer G. Dy (2004). “Automated hierarchical mixtures of probabilistic principal component analyzers”. en. In: *Twenty-first international conference on Machine learning - ICML ’04*. Banff, Alberta, Canada: ACM Press, p. 98. DOI: 10.1145/1015330.1015393.

Teng Zhang, Arthur Szlam, and Gilad Lerman (Sept. 2009). “Median K-flats for hybrid linear modeling with many outliers”. In: *Proceedings of 2nd IEEE International Workshop on Subspace Methods*. Kyoto, Japan: IEEE, pp. 234–241. DOI: 10.1109/ICCVW.2009.5457695.

Thomas, Paul D. et al. (Sept. 2003). “PANTHER: A Library of Protein Families and Subfamilies Indexed by Function”. en. In: *Genome Research* 13.9, pp. 2129–2141. ISSN: 1088-9051. DOI: 10.1101/gr.772403.

Tipping, Michael E and Christopher M Bishop (Feb. 1999a). “Mixtures of Probabilistic Principal Component Analyzers”. en. In: *Neural Computation* 11.2, pp. 443–482. DOI: 10.1162/089976699300016728.

– (Sept. 1999b). “Probabilistic Principal Component Analysis”. en. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61.3, pp. 611–622. ISSN: 1369-7412, 1467-9868. DOI: 10.1111/1467-9868.00196.

Trivedi, Vikas et al. (Jan. 2018). “Multidimensional quantitative analysis of mRNA expression within intact vertebrate embryos”. en. In: *Development* 145.1, dev156869. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.156869. URL: https://journals.biologists.com/dev/article/145/1/dev156869/48765/Multidimensional-quantitative-analysis-of-mRNA.

Ville Satopaa et al. (June 2011). “Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior”. In: *ICDCSW ’11: Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops*. Minneapolis, MN, USA, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20.

W. Edwards Deming (1943). “The General Problem In Least Squares”. In: *Statistical adjustment of data*. New York: John Wiley & Sons, Inc, pp. 49–51.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis (Dec. 2018). “SCANPY: large-scale single-cell gene expression data analysis”. en. In: *Genome Biology* 19.1, p. 15. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0.

Wolock, Samuel L., Romain Lopez, and Allon M. Klein (Apr. 2019). “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data”. en. In: *Cell Systems* 8.4, 281–291.e9. ISSN: 24054712. DOI: 10.1016/j.cels.2018.11.005.

Yuen, Michaela et al. (Nov. 2015). “Muscle weakness in *TPM3* -myopathy is due to reduced Ca^{2+} -sensitivity and impaired acto-myosin cross-bridge cycling in slow fibres”. en. In: *Human Molecular Genetics* 24.22, pp. 6278–6292. ISSN: 0964-6906, 1460-2083. DOI: 10.1093/hmg/ddv334.

Zeynep F Altun and David H Hall (2009a). “Epithelial system, hypodermis”. In: *WormAtlas*. URL: doi:10.3908/wormatlas.1.13.

– (2009b). “Epithelial system, seam cells”. In: *WormAtlas*. URL: doi:10.3908/wormatlas.1.14.

Zheng, Shijie C et al. (Mar. 2017). “Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses”. en. In: *Nature Methods* 14.3, pp. 216–217. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4187.

Appendix

Here, I show that to fit our model when all spaces share a single variance parameter, σ^2, for noise, neither the E nor the M step of our algorithm needs to be modified and σ_i^2 can simply be corrected after clustering for each space at the end of each M step. Including this constraint does not affect the fit of latent Gaussian variables (in other words, neither the basis vectors nor latent variance parameters). The only exception is if an independently fitted latent variable has less variance than the estimated shared noise parameter, in which case both variance parameters will be updated.

E step:

No E step is performed with unequal noise in the complementary space, so the constraint is satisfied.

M step:

To update the parameters of space s_i in the M-step, we maximize its contribution to the expected complete-data log-likelihood, weighted by the responsibilities π_{ij}. Letting $\mathbf{x}_j = \mathbf{x}_j^{\parallel} + \mathbf{x}_j^{\perp}$ denote the decomposition of each observation, the contribution to the expected log-likelihood from component s_i is:

$$\mathcal{L}_i = \sum_{j=1}^{N} \pi_{ij} \log \left[f_{\parallel}(\mathbf{x}_j^{\parallel} \mid s_i) \cdot f_{\perp}(\mathbf{x}_j^{\perp} \mid s_i) \right],$$

$$= \sum_{j=1}^{N} \pi_{ij} \log \frac{1}{(2\pi)^{d/2} \prod_{a=1}^{d} \sigma_{ia}} - \frac{\pi_{ij}}{2} \sum_{a=1}^{d} \frac{\mathrm{disp}_{s_i}(\mathbf{x}_j)_a^2}{\sigma_{ia}^2} + \pi_{ij} \log \frac{1}{(2\pi)^{n/2} \sigma_i^n} - \frac{\pi_{ij}}{2} \frac{r_{ij}^2}{\sigma_i^2}.$$

However, we will consider the procedure of fitting an unconstrained multivariate Gaussian (obtained by setting $d = D$) to the assigned data to show that arbitrarily altering its lower $\sigma_{ia}^2 ... \sigma_{iD}^2$ parameters after it has been fit will affect neither the basis vectors for our decomposition nor the upper $\sigma_{i1}^2 ... \sigma_{i,a-1}^2$ parameters. Assume our Gaussian is centered at the weighted mean of the data assigned to the space, and our basis vectors $b_{i1} ... b_{ia} ... b_{iD}$ (also free parameters) are chosen to be our principal components. Then the components of the total variance of our data along the axes defined by our basis vectors are $v_{i1}^2 ... v_{ia}^2 ... v_{iD}^2$ (not parameters but values dependent on our choice of basis vectors), where, without loss of generality, our indexing for basis vectors is chosen such that v_{ia}^2 are sorted in rank order from largest to smallest. Setting the last n of our σ_{ia} parameters equal to some $\sigma \leq \sigma_{id}$ in the k-spaces algorithm preserves the rank order, and I will show it is equivalent to d-dimensional

PCA regardless of which σ is exchanged for the lowest n σ_{ia} and any additional $\sigma_{ia} < \sigma$. Let $\mathcal{L}_i^*$ be the expected log-likelihood when we obtain b_{ia}, v_{ia}^2, and σ_{ia}^2 from fitting the unconstrained Gaussian, which corresponds to classical principal component analysis with full dimension.

$$\begin{aligned}
\mathcal{L}_i^* &= \sum_{j=1}^{N}(\log \frac{1}{(2\pi)^{D/2}\prod_{a=1}^{D}\sigma_{ia}} - \frac{1}{2}\sum_{a=1}^{D}\frac{\text{disp}_{s_i}(\mathbf{x}_j)_a^2}{\sigma_{ia}^2})\pi_{ij} \qquad (3.14)\\
&= N\pi_i \log \frac{1}{(2\pi)^{D/2}\prod_{a=1}^{D}\sigma_{ia}} - \sum_{j=1}^{N}\frac{1}{2}\sum_{a=1}^{D}\frac{\text{disp}_{s_i}(\mathbf{x}_j)_a^2\pi_{ij}}{\sigma_{ia}^2},\\
&= -N\pi_i \log((2\pi)^{D/2}\prod_{a=1}^{D}\sigma_{ia}) - \frac{1}{2}\sum_{a=1}^{D}\sum_{j=1}^{N}\frac{\text{disp}_{s_i}(\mathbf{x}_j)_a^2\pi_{ij}}{\sigma_{ia}^2},\\
&= -N\pi_i \log(\prod_{a=1}^{D}\sigma_{ia}) - \frac{1}{2}\sum_{a=1}^{D}\sum_{j=1}^{N}\frac{\text{disp}_{s_i}(\mathbf{x}_j)_a^2\pi_{ij}}{\sigma_{ia}^2},\\
&= -N\pi_i \sum_{a=1}^{D}\log(\sigma_{ia}) - \frac{1}{2}\sum_{a=1}^{D}\sum_{j=1}^{N}\frac{\text{disp}_{s_i}(\mathbf{x}_j)_a^2\pi_{ij}}{\sigma_{ia}^2},\\
&= -\frac{1}{2}\sum_{a=1}^{D}(2N\pi_i\log(\sigma_{ia}) + \frac{N\pi_i v_{ia}^2}{\sigma_{ia}^2}),\\
&= -\frac{N\pi_i}{2}\sum_{a=1}^{D}(2\log(\sigma_{ia}) + \frac{v_{ia}^2}{\sigma_{ia}^2}). \qquad (3.15)
\end{aligned}$$

As our axes $a...D$ are given by orthonormal basis vectors, the variance along each axis is independent and our space parameters $\sigma_{i1}^2...\sigma_{ia}^2...\sigma_{iD}^2$ can be fit independently. Therefore, assuming no rotation of the basis vectors occurs (meaning the variance components $v_{i1}^2...v_{ia}^2...v_{iD}^2$ will be unchanged), any update setting the lower $\sigma_{i,d+1}^2...\sigma_{i,D}^2$ equal to a model-wide σ^2 after the standard M step is performed independently for each space will not affect the first $\sigma_{i1}^2...\sigma_{id}^2$ as variance in the latent dimensions contributes independently to the likelihood. Now, I will show that no rotation occurs because for any arbitrary set of variance parameters $\sigma_{i1}^2...\sigma_{iD}^2$, the optimal choice of basis vectors and assignment of variance parameters to basis vectors is given by the principal components of the data.

I will prove this by induction. Assume for our (weighted) data we choose b_{i1} to be our top principal component and that we know the optimal solution for the remaining $(D-1)$ dimensions of the data resulting from that choice as the 1^{st} dimension. We additionally have our sorted arbitrary $\sigma_{i1}^2...\sigma_{iD}^2$ and have assigned σ_{i1}^2 to our principal component because it by definition explains the most variance

and should therefore be assigned the largest variance parameter (later I prove that for any set of basis vectors and arbitrary variance parameters, the optimal choice for the likelihood is to assign them rank order). Define a function $G_i(R)$ that gives the updated expected log-likelihood if we rotate b_{i1} (meaning that we do not build off of the $(D-1)$-dimension solution to merely append the parameters b_{i1} and σ_{i1}^2) and correspondingly must rotate some subset of other basis vectors $b_{ir_1}...b_{ir_c}$ to maintain orthonormality. Define $v_i(R)_a$ as the change in v_{ia}^2 (variance along b_{ia}) as a result of the rotation denoted by R. Assume that after the rotation, b_{i1} still explains more variance than any other basis vector; if it does not, then it is suboptimal at face value. To amend this, swap its index with the new basis vector that explains more variance, such that the σ_i^2 variance parameter is now assigned to the basis vector with the largest variance, degenerately arriving at another rotation that satisfies our assumption. As the total variance of the data itself is constant, b_{i1} was pointing in the direction of greatest variance before the rotation, $v_i(R)_1 < 0$ and $v_{i1}^2...v_{iD}^2$ must sum to the total variance, the sum of $v_{ir_1}^2...v_{ir_c}^2$ therefore must increase by $-v_i(R)_1$.

$$G_i(R) = \mathcal{L}_i^* - \frac{N\pi_i}{2}(-\frac{v_i(R)_1}{\sigma_{i1}^2} + \sum_{a=2}^{D} \frac{v_i(R)_a}{\sigma_{ir}^2}). \tag{3.16}$$

$\frac{v_i(R)_1}{\sigma_{i1}^2} > \sum_{a=2}^{D} \frac{v_i(R)_a}{\sigma_{ir}^2}$ because the latter is a set of fractions with denominators all smaller than σ_{i1}^2 while the numerators all add up to $v_i(R)_1$. Therefore, any rotation will produce a lower likelihood (and I do not mean this as any incremental rotation, which would indicate a local minimum; I mean *any* final rotation, making this the global optimum). Any rotation to the orthonormal basis that involves the basis vector b_{i1} will be suboptimal no matter what the variance parameters are, and by definition any other rotation not involving b_{i1} will either be suboptimal or equally optimal. Therefore, the optimal solution to the D-dimensional problem is the optimal solution to a $(D-1)$-dimensional problem for the $(D-1)$ dimensions of the data in the lower $(D-1)$ principal components, with principal component 1 appended to the solution. By induction, the expected likelihood is maximized by choosing basis vectors equal to the principal eigenvectors of the (weighted) covariance matrix, regardless of what happens to the variance parameters.

Below I show that the optimal ordering of arbitrary variance parameters is the same rank order as the true variances. Define a function $H_i(a, b)$ that gives the updated

expected log-likelihood if parameters σ_{ia}^2 and σ_{ib}^2 are exchanged for space s_i:

$$\begin{aligned} H_i(a,b) &= \mathcal{L}_i^* - \frac{N\pi_i}{2}(-\frac{v_{ia}^2}{\sigma_{ia}^2} + \frac{v_{ia}^2}{\sigma_{ib}^2} - \frac{v_{ib}^2}{\sigma_{ib}^2} + \frac{v_{ib}^2}{\sigma_{ia}^2}), \\ &= \mathcal{L}_i^* - \frac{N\pi_i}{2}(\frac{v_{ib}^2 - v_{ia}^2}{\sigma_{ia}^2} + \frac{v_{ia}^2 - v_{ib}^2}{\sigma_{ib}^2}). \end{aligned}$$

Define $z(a,b) = v_{ia}^2 - v_{ib}^2$ and substitute:

$$H_i(a,b) = \mathcal{L}_i^* - \frac{N\pi_i}{2}(\frac{-z(a,b)}{\sigma_{ia}^2} + \frac{z(a,b)}{\sigma_{ib}^2}). \tag{3.17}$$

$z(a,b) > 0$ if our parameter exchange reassigns σ_{ia}^2 from a basis vector with greater variance to one with lower variance, and is less than zero if vice versa. Therefore, any parameter exchange that decreases the rank order alignment of $\sigma_{i1}^2 ... \sigma_{ia}^2 ... \sigma_{iD}^2$ with $v_{i1}^2 ... v_{ia}^2 ... v_{iD}^2$ decreases our likelihood, and any exchange that increases the alignment increases our likelihood, making $\mathcal{L}_i^*$ the maximum.

Lastly, we consider the matter of how to handle a latent dimension in one of the spaces that has less variance than the computed model-wide σ^2. In the spirit of being able to freely interchange between k-spaces and pPCA and assuming that noise is truly isotropic and occurs within the latent space as well, we do not allow this. σ^2 should be the minimum variance, and now those specific latent dimensions of those specific spaces should be incorporated into estimating the variance of isotropic noise. To do this correctly, we order the problematic latent variances from smallest to largest. They should be successively (or with binary search) added into the calculation of σ^2 until no latent σ_{ia}^2 violates the condition $\sigma_{ia} > \sigma \ \ \forall i \, \forall a$.

Chapter 4

LOW-DIMENSIONAL SUBSPACES REVEAL TISSUE-LEVEL DIVERSITY OF GENE COVARIANCE WITHIN PATHWAYS

This chapter is a manuscript in preparation: Nicholas Markarian, Lior Pachter, and Paul Sternberg, "Low-dimensional subspaces reveal tissue-level diversity of gene covariance within pathways."

Abstract

Pathways are regularly used to interpret RNA-sequencing results in physiological context. Gene expression comparisons across tissues in pathways are typically analyzed using mean expression profiles, ignoring tissue-dependent similarities or differences in covariance structure. On the other hand, gene expression networks built using correlations between genes typically ignore similarities or differences in expression level. Whether different expression levels of correlated genes in different tissues are dictated by a global scaling relationship or whether this itself is context-specific is not captured by these approaches. We find that correlations can be confounded by Simpson's paradox due to different sample subpopulations and therefore apply the *k*-spaces algorithm to learn gene coexpression scaling relationships and characterize their similarity or specificity across tissues. We find that in the metabolic pathways studied, gene expression is restricted to subspaces which we represent with a mixture of latent variable models, and we find that the partitioning of samples among subspaces coherently reflects known relationships between tissues.

Introduction

Over the past few decades, we have constructed qualitatively detailed models of pathways that are used to interpret transcriptomics data (Kanehisa and Goto, 1999; Milacic et al., 2024; The Gene Ontology Consortium et al., 2023; Thomas, Hill, et al., 2019; Thomas, Campbell, et al., 2003). These models serve as a valuable resource for computational methods that generate hypotheses about pathway regulatory state (Huang, Sherman, and Lempicki, 2009; Subramanian et al., 2005; Thomas, Campbell, et al., 2003) and for researchers manually examining pathway models to decide whether to pursue follow-up experiments probing those pathways.

However, these models lack a quantitative reference for gene expression variation. Typical expression ranges can be found individually for genes in a pathway from various resources (The GTEx Consortium et al., 2015; GTEx Consortium, 2017; Lachmann et al., 2018; Papatheodorou et al., 2018)), and average expression values per tissue are now displayed in Reactome alongside pathways (Milacic et al., 2024). Violin plots displaying the distribution of expression for each gene can be added to show expression ranges alongside pathway models, but they represent each gene independently, ignoring covariation between genes. Information about which genes in pathway models are correlated, in what contexts, and whether scaling relationships are conserved across tissues is lacking. Approaches for characterizing gene expression similarity in pathways across cell types or tissues (Granados, Kanrar, and Elowitz, 2024; Abu Rmaileh et al., 2020; Frost, 2021) typically rely on an averaging step first to reduce the data to a set of mean expression profiles that are subsequently compared, ignoring the covariance of the genes within a tissue or cell type.

The problem of identifying correlated subsets of genes within pathways could be addressed using methods for constructing gene expression networks for the pathways individually instead of the entire transcriptome. The question of which samples to use with what level of context-specificity, however, could be of importance. Recent efforts have uncovered context and tissue-specificity of gene coexpression networks (Piro et al., 2011; Pierson et al., 2015; Saha et al., 2017; Kunes et al., 2024). However, because gene expression networks based on the correlations between genes are inherently scale-free (Langfelder and Horvath, 2008), correlation-based approaches can miss differences in slopes of gene expression relationships when the correlation structure is still conserved. Thus, conserved correlations do not imply that a single eigengene (a dimension reduction of a set of genes to their projection onto principal component 1) (Langfelder and Horvath, 2008; Alter, P. O. Brown, and Botstein, 2000) to represent pathway gene covariation across samples from heterogenous sources is appropriate. Moreover, whether or not they start with a correlation matrix that eliminates information about the level of gene expression, approaches focused on *building* a network still focus on whether the genes covary rather than characterizing the stoichiometry that produces the covariation. The conservation and tissue-specificity of gene coexpression itself— accounting for both the stoichiometry of mRNA abundances and their covariance structure— have not been studied jointly to our knowledge owing to a lack of computational tools to do so. In initial work, we found that gene coexpression relationships, even within pathways, were influenced by tissue identity, which confounded the relationships.

To parse such context-specific effects, we developed *k*-spaces clustering in other work (Markarian, Engelhardt, et al., 2025).

In this study, we applied *k*-spaces clustering to gene coexpression for three metabolic pathways, revealing the diversity of gene expression relationships and showing that expression scaling could be well-represented with a low-dimensional latent variable models. We found that gene expression did not vary freely within independently observed expression ranges but was restricted to subspaces tied to tissue identity. Although the careful choice of genes was important to learning meaningful latent variable representations, we forced interpretability by utilizing curated pathways, and this approach is complementary to gene expression network discovery methods and matrix factorization methods by characterizing expression stoichiometry and expression level similarity across tissues. We take a step towards constructing useful references for pathway gene expression by discovering low-dimensional subspaces for modules of consistently correlated genes within three metabolic pathways, and we outline outstanding challenges and strategies for scaling this approach to our shared databases of pathways.

Results

Simpson's paradox confounds linear correlations in publicly-available gene expression data

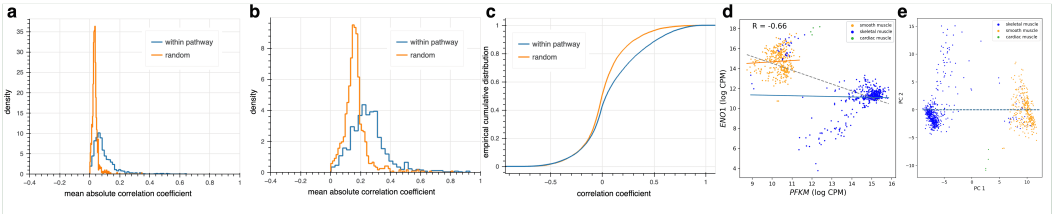


Figure 4.1: **Simpson's paradox in publicly available data.** **a,** Distribution of mean absolute correlation coefficients in the ARCHS4 dataset using samples from all tissues. **b,** Distribution of mean absolute correlation coefficients in the ARCHS4 dataset using muscle samples only. Correlations are higher but with less separation between intra-pathway means and size-matched, random sets of genes. **c,** Empirical cumulative distribution function of correlation coefficients between genes in muscle samples. **d,** Simpson's paradox illustrated in *PFKM* and *ENO1*. The overall best fit line, with $R = -0.45$, is shown in gray, with best fit lines for smooth and skeletal muscle plotted as well. **e,** PCA of glycolysis pathway genes, colored by muscle subtype. The overall best fit line, principal component 1 passing through the data mean, is shown in gray.

To determine whether a latent variable model could be appropriate to represent

gene expression in known pathways, we examined the correlation structure between genes in the human ARCHS4 dataset (Lachmann et al., 2018). ARCHS4 contains uniformly preprocessed publicly available bulk RNA-sequencing data from the Sequence Read Archive, and prior work showed GO Biological Processes could be predicted from the genome-wide correlation matrix (Lachmann et al., 2018). We found that while the distribution of correlation coefficients within pathways was positively shifted with respect to random genes, it was strikingly low (Fig 4.1a). We subsequently followed up by examining the mean correlation coefficients in specific protein complexes that reasonably would be expected to have tight regulation and high correlation between subunits: 80S ribosome, ATP synthase, electron transport chain complex I, 26S proteasome, SNARE, DNA Polymerases α, δ, γ, and ϵ, RNA Polymerases I-III, and the nuclear pore complex. Even with excluding variable subunits, these were surprisingly low as well. Though misquantification of similar transcripts could play a role, all had mean correlation coefficients below 0.2 except ATP synthase and 80S ribosome, which had means of 0.21 and 0.34, respectively.

However, when we split the samples by tissue type and examined three tissues we expected to be of relatively homogeneous composition— muscle, adipose, and liver tissues— we found much higher correlations, both within and outside of pathways, and a surprising abundance of negative correlations. Results with muscle samples are shown in Fig 4.1b-e. While it was not unexpected to have negative correlations between pathways or biological processes, the abundance of negative correlations within the pathways seemed unusual. We examined the data behind some of these negative correlations in the glycolysis pathway in muscle samples, as we were unaware of any such regulation and found that the negative correlations were actually a result of Simpson's paradox — confounding due to different gene coexpression relationships for smooth and skeletal muscle. Both were retrieved as part of our "muscle" dataset, and shifts in gene expression relationships reversed positive intra-subtype correlations (Fig 4.1d). When we visualized pathway-level data using principal components analysis for glycolysis genes, we found that PC 1, the so-called eigengene (Alter, P. O. Brown, and Botstein, 2000; Langfelder and Horvath, 2008), was nearly orthogonal to the common axis of variation (Fig 4.1e).

The heterogenous specificity of labeling in publicly available data and the apparent tissue-specificity of gene coexpression slopes inspired our development of the k-spaces approach in other work (Markarian, Engelhardt, et al., 2025). We sought to apply it here to gene expression data of pathways to characterize the diversity and

similarity of gene coexpression across samples from different tissues. For subsequent analyses, we used data from the GTEx consortium, as it was generated using consistent sequencing technology and standardized sample collection procedures, and because it had consistently documented metadata.

Tissue-specific analyses suggest a pathway-wide module of correlated glycolysis genes with potential tissue-specific stoichiometry

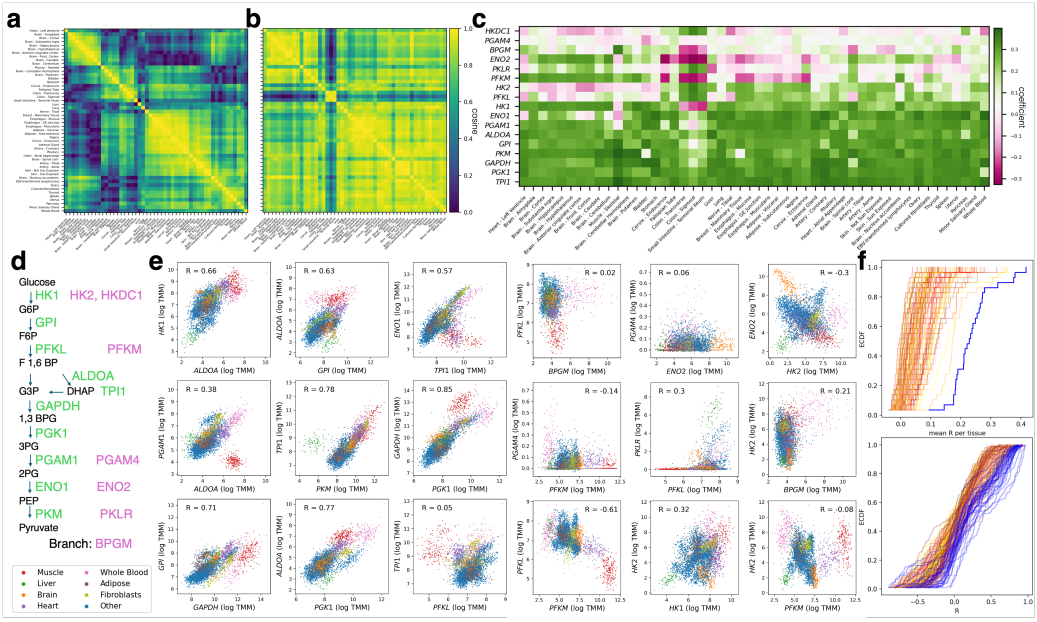


Figure 4.2: **The glycolysis pathway contains a pathway-complete module of genes consistently correlated across tissue subtypes but with tissue-specific stoichiometry. a,** Angle between principal eigenvectors of log tmm data for each tissue subtype, reported as the cosine. **b,** Angle between principal eigenvectors of z-transformed log tmm data for each tissue subtype. **c,** Coefficients of eigenvectors in (b). Sum of squared coefficients add up to 1 as eigenvectors are all length 1. **d,** Genes in glycolysis pathway sorted by pathway steps with "core" module genes selected for clustering highlighted. **e,** Selected pairwise scatter plots of glycolysis genes. Selected tissues are colored. Inset Pearson correlation coefficients are for linear regression of all data shown in the plot. **f,** Comparison of gene-gene correlation matrix to randomly selected groups of genes. Top: Each blue ECDF represents the Pearson correlation coefficients between glycolysis genes for a different tissue. Warm colored (brown, yellow, red, orange, or brick red) ECDFs represent the coefficients for a different tissue for one of 5 sets of randomly selected genes (see Figs 4.9, 4.10, 4.11, 4.12, 4.13). Bottom: Each ECDF displays the distribution of mean Pearson correlation coefficient of gene-gene correlation matrices over tissues (29 points per curve for 29 tissues) for a different set of genes. Blue represents the glycolysis genes. Warm colors represent 100 random sets of genes.

Before applying k-spaces to cluster samples, we first split the data by tissue subtype labels, and we fit a single latent variable model to each tissue separately to identify whether a signal of pathway gene covariation could exist and whether it differed by tissue. We analyzed the glycolysis pathway because it is ubiquitously and expressed across cell types and relatively well-understood.

We computed the top eigenvector of samples for each tissue subtype in log-transformed gene expression space, and we observed substantial similarity across tissue subtypes (Fig 4.2a). However, while the log transformation generally stabilizes variance, it does not produce equal variance among genes, meaning it was possible that the tissue-specific eigenvectors were largely driven by a single gene with wide expression range rather than a correlation structure between genes. Therefore, we additionally z-transformed the data for a scale-free comparison between tissues and visualized the eigenvectors' gene coefficients (Fig 4.2b,c). The tissue heatmap displayed strong similarity between nearly all tissue subtypes, and the gene coefficient heatmap (Fig 4.2c) suggested a consistent subset of correlated genes across tissue subtypes. We identified *PFKL, HK1, ENO1, PGAM1, ALDOA, GPI, PKM, GAPDH, PGK1,* and *TPI1* as a consistently correlated subset of genes in the glycolysis pathway. While chosen solely based on the consistent positive coefficients in the heatmap, these genes together span all steps of the glycolysis pathway, which has multiple isoforms for several steps (Fig 4.2d).

For all subsequent analyses, we did not apply a z-transformation and worked directly in log space, as applying the heuristic z-transformation to log-transformed data corresponds to exponentiating features by different degrees in gene expression space, and scaling data from each tissue separately before recombining the dataset is unphysical. Furthermore, z-transforming the entire dataset in the presence of clusters would not necessarily rescale features to equal variance within individual clusters. We also excluded *HK2* before clustering, which arguably could have been considered part of the correlated subset, for the practical reason that its expression varies across more orders of magnitude per tissue than other genes, meaning it would be the primary signal modeled by k-spaces.

Principal components alone only indicate the direction of maximum variance in the data, and a high percentage of variance explained by PC 1 may be driven by separation of clusters rather than data points lying near a line in the high dimensional space. Therefore, we visualized the data across a series of pairwise gene expression scatter plots sorted by gene order from the coefficient heatmap (Figs 4.19, 4.2c,e)

to confirm that there was indeed a correlated module of genes. Coloring by tissue type revealed clusters in the aggregate data corresponding to tissue groups, as observed in the ARCHS4 data as well. As we intended to use TMM normalization (Robinson and Oshlack, 2010) to combine data across the tissues, we examined the difference in distribution of correlation coefficients within each tissue subtype for the glycolysis pathway versus random sets of genes (Fig 4.2f). The substantially greater correlations within the glycolysis pathway than between random genes suggested our preprocessing was suitable and that the covariation in the data was potentially a meaningful signal rather than a technical or generic biological artifact. Taken together, these results suggest that there is correlated variation in gene expression for the glycolysis pathway across tissues for a common, pathway-complete set of genes but with shifting coexpression stoichiometry, and we set out to characterize this diversity with k-spaces.

k-spaces clustering reveals diversity in coexpression relationships of a consistently correlated set of glycolysis genes

k-spaces is a probabilistic generalization of k-means to cluster onto lines, planes, or higher dimensional affine subspaces. In this context, it generalizes the idea of linear regression between two genes to a best fit line between many genes (the eigengene) and further generalizes this to a probabilistic mixture model of best fit lines, each representing different context-specific 'modes' of gene expression scaling. In short, it fits a mixture of linear latent variable models to the data, determining subspaces of the feature space along which gene expression is scaled, fitting latent variables to represent the scaling, and determining which samples are found near which subspace. Outputs are affine subspaces and corresponding clusters in the data assigned to those subspaces. Post-clustering analysis could reveal similarity in subspace usage across tissues and subspace-specificity. This could alternatively be done by splitting data by the most fine-grained labels available to compute best fit lines with linear regression and heuristically comparing them to one another; however, this loses statistical power due to smaller sample sizes, and it remains possible that there are further unlabeled subpopulations that would confound the analysis.

We iteratively clustered from $k = 1 \ldots 49$ with subspace dimension $d = 3$ and found there were 24 subspaces for gene expression. Based on single-tissue estimates, we chose $d = 3$ to allow two additional free parameters for explicit modeling of some degree of uncorrelated variation or additional meaningful but minor axes of covariation (Fig 4.14 a-c). While $d = 1$ is most interpretable — the basis vector

corresponds to the principal component of the assigned data for a subspace, in the clustering setting, choosing too low of a value for d could lead to a piecewise model representation of an actually higher-dimensional subspace and a higher value of k: for example, representing a single planar ($k = 1$, $d = 2$) cluster with a two or three shifted linear clusters ($k = 2$, $d = 1$). We did not explore mixed dimensions of subspaces within the same model in this study. To increase the chances of finding the global maximum likelihood solution with our EM approach, we used deterministic annealing and successively built up our clustering solutions for each value of k using probability density estimates from lower-k models. We used the k^{th} solution to produce k informed initializations for the $k + 1^{th}$ EM runs along with 50 randomized initializations. Tracking ICL and the natural emergence of sparse clustering solutions revealed an effective value of k, akin to variational Bayesian GMMs in scikit-learn, with $k_{eff} = 24$ subspaces (Fig 4.14 e,f). The top basis vector (of the three) of each subspace correlated strongly with expression of the 'core' genes and even with some of the held out genes (Fig 4.3a). The resulting coefficient plot was qualitatively similar to the result obtained by splitting data with tissue subtype labels instead of clustering (Fig 4.2c), suggesting the k-spaces model was likely a meaningful representation.

We then assessed whether this produced coherent tissue groups, as the analysis prior to clustering implied that different coexpression scaling relationships could correspond to different types of tissues (Fig 4.2e). Liver, skeletal muscle, heart, lymphocytes, pancreas, adrenal gland, whole blood, tibial nerve, pituitary gland, thyroid tissues were assigned to their own subspaces, suggesting unique modes of gene coexpression, while other subspaces were shared by several tissues (Fig 4.3b-d). Generally, tissues were mostly found in one subspace, with exceptions of breast mammary gland and fibroblast samples. Breast samples were found with either adipose tissue or the minor salivary gland, and fibroblasts were located in two different subspaces, which could correspond to different cell states (Fig 4.3b). A heatmap depicting the proportions of samples of each tissue assigned to each subspace was sparse, and, in comparison, Leiden clustering of the same data with default parameters did not yield such a partitioning of tissues (Fig 4.3e), suggesting the tissue-subspace specificity in the subspace representation was non-trivial and a result of the model specification.

Hierarchical clustering on the tissue-cluster assignment proportions revealed a coherent grouping of tissue subtypes that could be explained by underlying cell type

composition (Fig 4.3d). Different subtypes of the same tissue (e.g., subcutaneous and visceral adipose tissue, sun-exposed and non sun-exposed skin, or various brain regions) were grouped together, consistent with expectations, but groupings across tissues and differences between subtypes for some tissues were meaningful as well and appeared to be due to cell type composition. For example, GTEx esophagus samples were taken from the muscularis, the gastroesophageal junction, or the mucosa. In our subspace clustering, the first two were paired with the sigmoid colon in a 'gastrointestinal smooth muscle' tissue group, while the latter, containing epithelial tissue, was paired with the ectocervix, vagina, and skin subtypes, all containing stratified squamous epithelium. The endocervix, in contrast, was paired with other samples from the female reproductive system, and the endocervix and ectocervix differ through a cell transition from squamous to columnar epithelium (Prendiville and Sankaranarayanan, 2017). These results indicated that the simple latent variable model was a coherent and effective representation of gene expression covariation.

Although the coherence of tissue groupings was encouraging, we also used quantitative approaches to assess the quality and validity of the clustering and subspaces learned by k-spaces. The first was to confirm that the sample partitioning by tissue identity was due to separability in the 10-dimensional gene expression space rather than mere interpolation of the mean gene expression values for each tissue. Visualization with 3D PCA showed some obvious linear clusters in the data, but many of the k-spaces clusters were close together, making it difficult to tell whether they were really distinct. Therefore, we reran the same k-spaces procedure from $k = 1 \ldots 20$ but on the data's projection into 3D PCA space. k-spaces did not recover a sparse tissue-cluster assignment matrix, confirming that the tissue groups were separable in the 10-D space using information not represented by the 3-D plot (Fig 4.15).

We then compared the compactness of the k-spaces model versus three null hypotheses: 1) no clusters exist and no genes are correlated (bootstrapping features independently), 2) no clusters exist but genes are correlated (fitting a single multivariate Gaussian to the whole data), 3) clusters exist, but within each cluster, features are uncorrelated (bootstrapping features independently within each tissue subtype). Both with data simulated from the null hypotheses and k-spaces fit to the null hypotheses' synthetic data, as k-spaces itself is a generative model, we found that the actual GTEx data was much more compact in its empirical density, and that feature was approximated well by k-spaces (Fig 4.3f). The compact density of the actual data meant there were true clusters or that the data lay near a low dimensional

manifold, and the close fit of the *k*-spaces model suggested that both were true: gene expression did not freely vary throughout gene expression space but was restricted to a set of low-dimensional linear subspaces.

We finally compared the *k*-spaces subspaces to the eigenvectors obtained using the tissue subtype labels instead of clustering to show that we had learned not only a compact representation but also one that accurately reflected the primary axis of gene expression variation in tissues. As we had clustered in logarithmic rather than linear space, both angular changes and translation are meaningful: parallel lines in log-log plots correspond to univariate polynomials (i.e., $y = ax^n$) with the same exponent but different slopes (e.g., $y = x$ and $y = 3x$ are parallel in log-log plots with a slope of 1 and shifted by $\log(3)$). Therefore, we compared both the angle of *k*-spaces basis vectors and the distance of samples' projections from their actual positions in gene expression space to the tissue subtype eigenvectors. We found that in comparison to representing the entire dataset with one eigenvector, null hypothesis 2, the data could be more accurately and also meaningfully represented in low dimension with 24 subspaces reflecting different modes of glycolysis 'core' gene coexpression among the 49 tissue subtypes.

Subspaces of tricarboxylic acid cycle gene expression display similar tissue groupings as in the glycolysis pathway

We subsequently applied the same methodology to study the genes in the tricarboxylic acid (TCA) cycle, as it was also a ubiquitously expressed metabolic pathway (Fig 4.4a). We found similarly strong correlation between genes and the *k*-spaces latent variables (Fig 4.4b), except for *PDHA2*, a testis-specific gene that catalyzes the same pathway step as *PDHA1*. Tissues were again partitioned into individual clusters (Fig 4.4c,d) with similar groupings of tissues as the glycolysis pathway. Notable exceptions were the separation of the two skin subtypes from the other epithelial tissues and the cerebellum from the other brain regions.

Pairwise scatter plots again showed confounding of Pearson correlation coefficients by distinct gene-gene relationships in certain tissues (Fig 4.4e). Comparison to the tissue-specific eigenvectors also showed that *k*-spaces better represented the primary axis of inter-sample gene expression variation for this pathway, with the exception of whole blood samples. However, whole blood samples are unique in the GTEx dataset as some samples were collected premortem, and earlier work has shown changes in gene expression after death by comparison of pre and postmortem

samples. *k*-spaces partitioned whole blood samples by pre- and postmortem status, representing whole blood with two subspaces. In contrast, the top eigenvector of all whole blood samples' TCA cycle gene expression is driven by postmoretem changes, consistent with findings reported elsewhere (Ferreira et al., 2018) and not representative of gene expression variation within either subgroup. It is another case of Simpson's paradox (Fig 4.16).

Gene expression subspaces of the cholesterol biosynthesis pathway reflect active versus inactive tissues

We chose cholesterol synthesis for our third metabolic pathway, because it is expressed in all cells for basic cell membrane maintenance (Groen et al., 2014), is non-mitochondrial (unlike TCA, which uses nuclear genes but could be driven by mitochondrial content variation), and does not share genes with glycolysis or TCA. Furthermore, it is presumably highly differentially active since some tissues use it in greater quantities for specific purposes (e.g., steroid hormone synthesis), and it is produced for the rest of the body by the liver (Simonen et al., 2011; Groen et al., 2014). Conveniently, it is also known to be transcriptionally regulated in addition to post-translational feedback inhibition (Goldstein and M. S. Brown, 1990; Bhattarai et al., 2021). We selected *HMGCS1, HMGCR, MVD, MVK, FDPS, FDFT1, SQLE,* and *LSS* for clustering with $d = 3$ (Figs 4.21, 4.17). In contrast to its results on the glycolysis and TCA pathways, *k*-spaces clustering produced in $k_{eff} = 17$, with many diverse tissues in one subspace, Subspace 0 (Fig 4.5 a,b).

The subspace containing many tissues was planar rather than linear— the latent variables representing the first two basis vectors had similar variance, 0.61 and 0.43, while the third trailed at 0.16. Basis 1 of Subspace 0, its top eigenvector, was not aligned with the tissue subtype eigenvectors and only correlated with *HMGCR, SQLE,* and *HGMCS1*, the first two of which are known points of negative post-translational feedback regulation (Goldstein and M. S. Brown, 1990; M. S. Brown and Goldstein, 1997; Kraemer, 2007; Foresti et al., 2013; Bhattarai et al., 2021). The second basis correlated with the remaining genes except *MVD*, which was represented by the third basis vector (Fig 4.5 c-f). This may be significant, but without corresponding data for cholesterol synthesis activity, it seems unwise to speculate about the meaning of shifts along the axes of this latent space.

Comparison to other subspaces, however, was more clearly interpretable (Fig 4.5g; PCA of all data and subspaces together is shown in Fig 4.18). Liver tissue produces

the majority of the body's cholesterol outside of dietary intake (Groen et al., 2014), while the adrenal gland requires cholesterol to synthesize steroid hormones for the body and can synthesize to meet those demands (Kraemer, 2007; J. Hu et al., 2010; C. E. Gomez-Sanchez and E. P. Gomez-Sanchez, 2024). Liver samples were found in Subspaces 5 and 11, and adrenal gland samples were found in Subspaces 11 and 15. All three subspaces correspond to markedly higher expression levels of the cholesterol synthesis genes than the non-tissue-specific Subspace 0. The shared subspace, 11, representing the highest levels, possibly corresponds to an active state for liver and adrenal gland tissues given that studies have shown upregulation of several to all cholesterol synthesis genes in response to physiological need (M. S. Brown and Goldstein, 1997; Sakakura et al., 2001). Some brain frontal cortex samples were also found in Subspace 5, the lower expression state for the liver, which is consistent with the high cholesterol requirements for myelin synthesis and its de novo synthesis in oligodendrocytes rather than import from circulation (Pierre Morell and Helga Jurevics, 1996; Saher et al., 2005; Khandker et al., 2022).

Discussion

Diversity of gene coexpression relationships and future single-cell or bulk analyses

Our work showed that gene expression in the pathways studied was restricted to low dimesional subspaces but also revealed more diversity across tissues than expected. Hierarchical clustering of tissue-subspace assignments suggested that the distinct underlying cell type composition of different tissues played a role in determining the gene expression subspace for each tissue or tissue group. The finer resolution of single-cell sequencing could paradoxically simplify analysis if a smaller number of underlying prototypical cell-type-specific eigenvectors were discovered. The large diversity in tissue-level signatures could be the result of the different combinations of a few possible cell-type signatures for gene expression scaling based on the composition of each tissue.

The interpretation of relative location within our learned subspaces as regulatory state is limited by the unknown variation in underlying cell composition. Single-cell analyses could avoid the problem of tissue composition heterogeneity, while at the bulk level, cell-type proportions can vary within samples of the same tissue type (Kim-Hellmuth et al., 2020). We likely saw this among mammary gland samples in the breast, which were found in equal shares in the subspaces otherwise representing adipose tissue and the minor salivary gland. Significant variation in

cell-type proportions could mean that variation within the latent space (and therefore the gene expression signal of the tissues represented by the subspace) could be primarily driven by this variation rather than pathway regulation, which presents a challenge for future analyses of pathway regulation.

Accurately inferring cell-type composition of bulk samples to correct for this in bulk analyses is difficult and could confuse changes in gene expression state with cell composition variation. The GTEx consortium provided inferred cell-type enrichment scores using Xcell (Kim-Hellmuth et al., 2020; Aran, Z. Hu, and Butte, 2017), but it could produce mistakes. For example, liver samples were reportedly confounded by cell-type composition variation (see Fig S1 of (Oliva et al., 2020); (Kim-Hellmuth et al., 2020)), and adipocyte enrichment was reported to be correlated with steatosis (see Fig S7B of (Kim-Hellmuth et al., 2020)). However, hepatic steatosis is not a proliferation of adipocytes in the liver but rather lipid accumulation within hepatocytes themselves (Nassir et al., 2015) — adipocytes are not found in the liver, and single-cell studies of the liver in hepatic steatosis and steatitis did not report any adipocytes (Perry et al., 2024; Fred et al., 2022). Given the GTEx sample collection instructions (“4.3.6.15 Liver” n.d.) to collect liver samples from the "central right lobe, 1 cm below capsule," it seems plausible that hepatocytes display an adipocyte-like expression state either as a result of or causing lipid accumulation and that this expression signature confounded Xcell’s results, although it is possible that the reality of sample collection meant contamination from outside the liver. This could be an isolated case, but it is among the reasons we did not explicitly model cell composition variation in the data and instead sought to simply characterize inter-sample variation. Single-cell analyses have the potential to provide clearer interpretations if the biological (Gorin et al., 2022) and technical stochasticity in measurements could be accounted for.

However, we believe bulk data is still worth working with for several reasons. The first is technical — at the single-cell level, measurements are stochastic and of low sensitivity, while at the bulk level, gene expression measurements start with raw data of tens of millions of counts. The second is biological — at the single-cell level, biological stochasticity (Gorin et al., 2022) comes into play and fluctuations in copy numbers of transcriptional machinery and transient cell states would be additional, unwanted latent variables driving expression covariation in the data. These two could be addressed by pseudobulking approaches, but deciding which cells to group together is itself a clustering problem that would be influenced by that noise. The

third is both biological and practical. Long-term, we may be interested in comparing patient samples to pathway gene expression profiles as references. Single-cell data will not be an appropriate biological comparison and technical batch effects will undoubtedly pose a challenge for direct comparison. Furthermore, the now low cost of bulk sequencing and its comparative experimental ease — integrity of samples obtained in surgeries is lower priority than patient health — makes it a more likely source for widespread high quality samples for study than single-cell sequencing, so we believe it is worth continuing to develop methods for and analyzing bulk data.

Pathway-wide coregulation of glycolysis gene expression

It was somewhat surprising that genes across all steps in the glycolysis pathway displayed significant covariation across samples given the biochemical evidence of extensive post-translational feedback regulation and known key rate-limiting steps in the pathway: hexokinase, phosphofructokinase, and pyruvate kinase. It is not immediately obvious why the remaining, non-rate-limiting steps should display significant correlation across samples. *HK2* is well-known to be transcriptionally regulated in insulin-responsive tissues and can control glucose flux into the pathway as the first step, multiple genes have been shown in various tissues to be transcriptionally regulated (Wu, Khan, and Lange, 2005), and transcriptional work in a cell line identified a large subset of the pathway that appears to be upregulated together in response to *mTORC1* activation (Düvel et al., 2010). Our observation is generally similar to that study, but our broader results showing some degree of tissue-specificity in the correlation structure highlights the need to perform comprehensive studies across tissues. A key difference, however, is that our approach is observational, a limitation of a study at this scale. While lacking the detail of experimental approaches that perturb a system to identify specific regulatory actors and isolate their effects, the potential benefit of abstracting regulation is the ability to study the coordinated action of those regulatory elements at larger scales reflected in an apparent latent variable describing the effective axes of covariation. Thus, large-scale observational approaches and focused experimental approaches can be complementary. Regardless, we see that correlation structure is generally similar across a pathway-complete, core subset of genes but still displays some heterogeneity and that the additional isoforms can display tissue group-specific coupling in expression scaling, although this does not explain why all steps in the pathway appear to be coupled at the transcriptional level.

Recent work involving biophysical modeling, however, may provide an explanation. After rate-limiting steps of glycolysis were discovered, it was shown in yeast that

despite phosphofructokinase acting as a rate-limiting step, overexpression of hexokinase and subsequent buildup of intermediates could arrest growth by uncoupling the glycolysis pathway through the reaction of Harden and Young (Arthur Harden and William John Young, 1906; Otto Meyerhof, 1945; Choe et al., 2025). With the first whole-pathway biophysical model of glycolysis using estimates of protein levels from literature, Choe et al. showed that it was essential to maintain concentrations of non-rate-limiting enzymes in vast excess of rate-limiting steps to prevent the buildup of such intermediates and harmful side reactions that would disrupt ATP-production. This is consistent with our finding of pathway-wide coupling of gene expression, as coregulation of non-rate-limiting steps to maintain their ratio to rate-limiting steps is a simple way to ensure their expression is in excess.

More generally, this result suggests studies of gene expression covariation in established pathways are worthwhile even when post-translational regulation is well-understood, and it provides motivation to scale analysis to the collection of curated pathways (Kanehisa and Goto, 1999; Milacic et al., 2024; Thomas, Hill, et al., 2019; Thomas, Campbell, et al., 2003). Meaningful biological signals could be recovered that increase our understanding of how these pathways are regulated or confirm what careful experiments and biophysical modeling have discovered. In turn, characterizing the subspaces where gene expression has been observed, rather than treating each gene independently in pathway expression profiles, could guide biophysical modeling or inform perturbation experiments regarding which regions of expression space to explore. Even simple observational studies determining whether samples taken in disease states lie outside of the subspaces of healthy samples would be useful.

Strategies for scaling

In the course of this work, we identified several challenges for unsupervised, automated scaling. The first is choosing the subset of genes to cluster on. Uncorrelated genes do not necessarily matter in the context of single k-spaces subspaces or probabilistic PCA fit to homogeneous samples originating from a single tissue subtype without clusters, but uncorrelated variable genes can create additional clusters and complicate interpretation when tissues are combined. The idea behind k-spaces and probabilistic PCA is that variance in the lower principal components is modeled uniformly with isotropic noise, and uncorrelated genes will be absorbed into this component of the model and modeled implicitly. Uncorrelated genes with similar or smaller variance than the correlated genes in the pathway that are consistently expressed at similar levels across tissues can be effectively ignored by the

k-spaces model in this manner. However, differential expression of uncorrelated genes between tissue subtypes produces distinct clusters in the whole-pathway gene expression space even with all else held equal, but it is not meaningful for analyses characterizing the different coexpression relationships between the correlated subset of genes. Consider a hypothetical pathway of five genes, four of which are very tightly correlated with the same coexpresion relationship in all tissues, while the last gene is expressed distinct levels in each tissue. Each tissue will trivially be in a different subspace of the pathway-wide gene expression space.

In this study, we addressed that issue by pre-selecting genes based on their coefficients in z-transformed data prior to performing clustering (e.g., Fig 4.3c) as well as visually examining scatter plots (e.g., Fig 4.19, 4.20, 4.21) sorted by their average coefficients across tissue subtypes in the z-transformed data. This manual approach is not necessarily scalable. However, already developed methods for gene network detection (Langfelder and Horvath, 2008; C. Gao et al., 2016; Lu et al., 2021; Kunes et al., 2024) could be applied to the pathway models rather than genome wide as typically done. They could be applied to individual tissues before merging results, or biclustering methods such as BicMix (C. Gao et al., 2016) that seek to identify context-dependent gene modules could identify the subsets of genes could be used. Then, *k*-spaces could to characterize those genes' coexpression similarity and diversity across tissues.

The second challenge is automated cluster verification. The GTEx data had tissue and tissue subtype labels that could loosely serve as a ground truth in the sense that tissues could be expected to be found only in one or a few states as opposed to spread out across many *k*-spaces subspaces. However, publicly available datasets from multiple sources, such as ARCHS4, would not necessarily have consistent labeling, and even in the GTEx data with tissue subtype labels, there is no ground truth for which subtypes can exist in multiple states and for which pathways and contexts. Furthermore, we used 3D visualization software and PCA to iteratively confirm clusters that were visibly separated from other data and then recomputed PCA without those clusters to confirm the next clusters, but this approach only works while at least one cluster is obviously separable. With the glycolysis pathway, we were able to visually confirm 18 subspaces this way, but the remaining 5 were not obviously separable in the 3 dimensions of their joint PCA.

In scenarios where several clusters visualized together are not obviously separable and distinct, it is tempting to view the clusters in pairs or to consider them in

isolation and use Scree plot-style analyses to confirm their linearity. However, the clusters were jointly determined and their boundaries and basis vectors are interdependent, making this approach misleading. This problem is not unique; in single-cell clustering, tests for the differences in location between clusters found with *k*-means suffer from miscalibrated p-values, which was addressed by recent selective inference methods developed by Gao, Bien, and Witten (L. L. Gao, Bien, and Witten, 2022). Given the close ties between *k*-means and *k*-spaces, it seems reasonable that selective inference methods could be developed to yield appropriate p-values for whether given *k*-spaces subspaces are in fact distinct, and subspaces could be merged appropriately post-clustering with a corresponding increase in the new cluster's dimension to denote the additional degree of freedom. If so, *k*-spaces could be deployed on entire pathway databases and results could be made available to the broader community to aid in interpreting RNA-sequencing results.

Lastly, it would be useful to develop methods for meta-analysis of the subspaces, as we showed with the apparent active versus basal or inactive cholesterol synthesis gene expression states between the liver, adrenal glands, and our mixed tissue subspace. Concise ways to summarize and interpret subspace relationships would be immensely useful, as prior knowledge about the types of samples found in each space could reveal the biological meanings of these subspaces. Similarly, relative position within the subspace would be useful as well to analyze using metadata, particularly in data collected under experimental conditions. Reliable, interpretable integration of such data paired with *k*-spaces representations could enable us to leverage over a decade of sequencing data and publicly-funded research to collectively make sense of gene expression covariation and its consequences for pathway regulation.

Data and code availability

We retrieved the ARCHS4 dataset human_gene_v2.2 from `https://maayanlab.cloud/archs4/download.html` on October 31, 2023. Keyword search for "skeletal muscle," "smooth muscle," and "cardiac muscle" was done on the ARCHS4 browser on October 30, 2023, and resulting lists of samples are in our Jupyter notebook.

We used archs4py v0.2.18, anndata v0.9.2, numpy v1.23.5, pandas v2.1.4, python v3.10.9, v1.0scanpy v1.10.2, scipy v1.11.4, sklearn v1.3.0, bokeh 3.1.1, and iqplot 0.3.2. A pre-release version of *k*-spaces was used and is available at `https://github.com/pachterlab/MEPSP_2024/kspaces`.

Author contributions

NM, LP, and PS conceived the study. NM generated the results and drafted the manuscript.

Acknowledgments

We thank Sina Booeshaghi for initial help with retrieving the ARCHS4 data. This work was supported by NIH F30 1F30GM156092-01 to NM and by NIH-NHGRI grant U24HG010859 to PWS, a Bren Professor of Biology.

Methods

Initial analyses with the ARCHS4 dataset

Human bulk RNAseq data was obtained from ARCHS4 version 2.2 and split by tissue. The archs4py python package was used to extract muscle samples using the metadata search function with keywords "skeletal muscle," "smooth muscle," and "cardiac muscle." We then filtered out samples with less than 20 million reads and single-cell samples, resulting in 2,246 samples. 1,361 samples were not in the list produced by searching "muscle" on the ARCHS4 website's download page, so the samples were further filtered for the intersection between the two, resulting in 885 samples (549, 332, and 4 samples of each muscle subtype respectively). Pathway gene lists for Fig 4.1 a-c were retrieved from 2022-07-01 release of Gene Ontology Causal Activity Models as described in (Markarian, Van Auken, et al., 2024). The human_correlation_archs4.f v1.0 correlation matrix across all samples was downloaded from the ARCHS4 website (Lachmann et al., 2018) and used quantile normalization (Bolstad et al., 2003). To produce mean correlation coefficients in muscle samples only, we quantile normalized the muscle data using the archs4py function 'normalize' with method = "quantile" before computing correlation coefficients. Complexes were retrieved from Reactome (Milacic et al., 2024) and pathways were retrieved from the Gene Ontology Causal Activity Model database (Thomas, Hill, et al., 2019; Good et al., 2021).

GTEx data preprocessing

The log-transformation, TMM, and RIN-regression steps were chosen a priori on the basis of the variance-stabilizing property of the log-transformation and recommendations in prior studies (Gallego Romero et al., 2014; Reiman et al., 2017; Robinson and Oshlack, 2010; S. Zhao, Ye, and Stanton, 2020; Y. Zhao et al., 2021), but comparisons to alternate strategies were performed as well in Figs 4.7, 4.9, 4.10, 4.11, 4.12, and 4.13. Preprocessing strategy was assessed using the glycolysis pathway prior to clustering (Figs 4.2, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13).

Per the recommendation of prior studies (Gallego Romero et al., 2014; Reiman et al., 2017), a cutoff for RIN was determined and then the effect RIN was regressed out, as RNA degradation has been shown to produce spurious correlations in gene expression levels due to similar degradation rates (Gallego Romero et al., 2014; Reiman et al., 2017). We chose a cutoff of RIN > 7 based on the substantial increase in variance of glycolysis gene log TPM expression values below this level. Representative examples of tissue-level RIN versus gene expression plots are shown

in Fig 4.6.

Following this, we depth normalized counts by TMM as it has been shown to be appropriate for comparison across samples from different tissues (Robinson and Oshlack, 2010; S. Zhao, Ye, and Stanton, 2020; Y. Zhao et al., 2021). We then log transformed the data, and separately for each tissue, we regressed out the effect of the deviation in RIN from 8.0 on each gene with a linear model using scanpy. We chose RIN = 8.0 as the RIN of most samples of most tissues were in between 7.0 and 9.0, while few tissues had substantial numbers of samples with RIN > 9.0. We regressed to a value of 8.0 rather than regressing out RIN and keeping the residuals, as that would not preserve shifts in mean gene expression between tissues. Each tissue was regressed separately, as it was evident from Fig 4.6 that the specific relationship between RIN and gene expression, or lack of correlation, was tissue-dependent. Kidney, testis, and prostate tissue samples were not regressed due to insufficient numbers of samples in this range and were excluded from analyses.

To assess whether preprocessing was appropriate and to compare to alternate preprocessing strategies, the empirical distribution of Pearson correlation coefficients in the glycolysis pathway was computed for each tissue and compared to the empirical distribution of correlation coefficients for 5 sets of 15 genes selected uniformly at random that were expressed at least 10 TPM on average per tissue across at least all but one tissues. A similar analysis was performed for 100 random sets of genes meeting the same criteria, but with only a single ECDF displaying the distribution of median absolute pearson correlation coefficients per tissue. This was performed for all combinations of TMM versus TPM, log-transforming versus no transformation, and regression of RIN versus no regression, and for the same 5 random sets of genes in Fig 4.2f, eigenvector similarity heatmaps and heatmaps of z-transformed eigenvector coefficients were produced with each preprocessing combination (Figs 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13). Based on the high abundance of positive correlations between random genes with TPM normalization and the heatmaps qualitatively resembling those of glycolysis, a genuine pathway gene set, we determined TPM was not appropriate, even for intra-tissue analyses. However, ECDFs showed substantial separation with TMM regardless of whether the log-transformation and regression were applied, and these preprocessing decisions did not appear to influence the heatmaps substantially. We proceeded with TMM normalization, log-transformation, and RIN-regression as described above, as it displayed intra-pathway gene correlation and a relative lack of correlation between

randomly selected genes.

k-spaces clustering

Subspace dimension for *k*-spaces was selected on a per-pathway basis by applying the kneedle algorithm (Ville Satopaa et al., 2011) to the data split by tissue, as shown and described in Fig 4.17. Glycolysis and cholesterol synthesis were clustered with $d = 3$, while the TCA cycle was clustered with $d = 2$.

PANTHER Pathways (Thomas, Campbell, et al., 2003) gene lists retrieved from Harmonizome (Diamant et al., 2025) were as follows. "Glycolysis:" *HK1, HK2, GPI, PFKM, PFKL, ALDOA, TPI1, BPGM, GAPDH, PGK1, PGAM1, PGAM4, ENO1, ENO2, PKM, PKLR* was retrieved from `https://maayanlab.cloud/Harmonizome/gene_set/Glycolysis/PANTHER+Pathways`. *PGAM2* was excluded as neither 'PGAM2' nor its ensembl ID retrieved from GeneCards, 'ENSG00000164708,' were found in the downloaded GTEx dataset. "TCA Cycle:" *ACO2, CS, FH, MDH1, OGDH, PDHA1, PDHA2, SDHC, SUCLG1* was retrieved from `https://maayanlab.cloud/Harmonizome/gene_set/TCA+cycle/PANTHER+Pathways`. "Cholesterol Biosynthesis:" *FDFT1, FDPS, GGPS1, HMGCR, HMGCS1, IDI2, LSS, MVD, MVK, PMVK, SQLE* was retrieved from `https://maayanlab.cloud/Harmonizome/gene_set/Cholesterol+biosynthesis/PANTHER+Pathways`.

All genes were retained for clustering in the TCA cycle, while *HK1, GPI, PFKL, ALDOA, TPI1, GAPDH, PGK1, PGAM1, ENO1, ENO2,* and *PKM* were used for glycolysis and *FDFT1, FDPS, HMGCR, HMGCS1, LSS, MVD, MVK,* and *SQLE* were used for cholesterol synthesis, as described in the main text.

For *k*-spaces clustering analysis, a class-balanced dataset was created by randomly selecting 200 samples from each tissue type with > 200 samples. The same class-balanced dataset was used for all three pathway analyses.

A k_{eff} was determined by iteratively increasing k from 1 to 49 while using deterministic annealing and clustering solutions from the $k - 1^{\text{th}}$ run as initializations along with 50 random initializations to maximize the chances of finding the globally optimal solution while tracking the ICL metric at each value of k. Deterministic annealing (Ueda and Nakano, 1998) was added to the *k*-spaces package, and beta = 0.3 with a 1.2x increase at each successive annealing step was used. The nonrandom initializations for the k^{th} *k*-spaces model were constructed as follows. For each of the $k - 1$ subspaces in the previous model, the cluster was split further by two subspaces

determined by running k-spaces on the points previously assigned to that subspace. These two new subspaces were appended to the list of other subspaces to form a set of spaces to initialize the k^{th} model from. Repeating this for each subspace provided $k-1$ nonrandom starting initializations along with the 50 randomized initializations. The ICL metric only stopped clustering early in one case. Instead, we found that the ICL metric appeared to asymptotically straighten past $k > 20$ to 25 but with a low but nonzero slope. We correspondingly found that many of the subspaces in clustering solutions with $k > 20 - 25$ had the minimal number of allowed samples ($d + 1$ samples for a d-dimensional subspace) and were practically empty. The k_{eff} non-empty subspaces from the final $k = 49$ clustering solution were retained for analyses, and samples assigned to discarded subspaces were reassigned to the retained subspaces with an additional E-step of the EM algorithm.

Leiden clustering in Fig 4.2d

The Leiden algorithm was applied to the same glycolysis gene expression data using default parameters and scanpy. The neighbor graph was constructed with scanpy.pp.neighbors with method = "gauss". scanpy.tl.leiden was run with the default settings and yielded 23 clusters.

Eigenvector analyses

The first k-spaces basis vector, equivalent to the top eigenvector or principal component of the data assigned to that space, was used in all analyses except where noted in Fig 4.5e. In 3D PCA plots, it was projected into the 3D space. Tissue subtype-specific eigenvectors were computed using the entire dataset, in contrast to the k-spaces basis vectors that resulted from clustering on the class-balanced data.

To produce the comparisons in Figs 4.3g, 4.4f, and 4.5c, reported results at the tissue level were the mean of tissue subtypes, weighted by the proportion of samples from each subtype. Δ-degree was the cosine of the tissue subtype-specific eigenvector and the representative k-spaces subspace's basis vector 1, converted to degrees by the inverse cosine function. Δ-RMS was the difference in RMS error of representing samples by their projection onto their assigned subspaces and the RMS error of representing samples by their projection onto their tissue-specific eigenvector. These were compared to the null hypothesis of no clusters, in which the dataset-wide top eigenvector was used instead of k-spaces basis vectors.

Tissue hierarchical clustering

Hierarchical clustering with scipy.cluster.hierarchy.linkage(y, method="average", metric="euclidean", optimal_ordering=True) was applied to the vectors representing the proportion of samples assigned to each subspace for each tissue subtype. For example, the input for the hierachical clustering in Fig 4.3e was the matrix represented as a heatmap in Fig 4.3c.

Comparison of probability density compactness

For the real data, synthetic data from a null hypothesis, or synthetic data from k-spaces fit to either the real data or the null hypotheses, the compactness of the data was assessed as follows. Expression space was partitioned into a grid, with unit lengths varied from 0.1 to 4.0 log2p1 TMM, as indicated on the x-axis. Grid boundaries were determined by the maximum and minimum values in each gene, padded by 2.0 log2p1 TMM. The number of grid units containing one or more data points was counted and displayed as a function of grid size. 4,468 data points were simulated from each k-spaces model and each null hypothesis to match the 4,468 data points in the class-balanced real dataset. This was repeated 20 times and averaged.

References

"4.3.6.15 Liver" (n.d.). In: *GTEx Tissue Harvesting Work Instruction*. Ver 03.05, p. 11. URL: https://dctd.cancer.gov/data-tools-biospecimens/biospecimens-biobanks/resources/sops/gtex/bbrb-pr-0004-w1.pdf.

Abu Rmaileh, Areej et al. (Apr. 2020). "Large-Scale Differential Gene Expression Transcriptomic Analysis Identifies a Metabolic Signature Shared by All Cancer Cells". en. In: *Biomolecules* 10.5, p. 701. ISSN: 2218-273X. DOI: 10.3390/biom10050701.

Alter, Orly, Patrick O. Brown, and David Botstein (Aug. 2000). "Singular value decomposition for genome-wide expression data processing and modeling". en. In: *Proceedings of the National Academy of Sciences* 97.18, pp. 10101–10106. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.97.18.10101.

Aran, Dvir, Zicheng Hu, and Atul J. Butte (Dec. 2017). "xCell: digitally portraying the tissue cellular heterogeneity landscape". en. In: *Genome Biology* 18.1, p. 220. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1349-1.

Arthur Harden and William John Young (Apr. 1906). "The alcoholic ferment of yeast-juice". In: *Proceedings of the Royal Society B* 77.519, pp. 405–420.

Bhattarai, Asmita et al. (Sept. 2021). “Regulation of cholesterol biosynthesis and lipid metabolism: A microRNA management perspective”. en. In: *Steroids* 173, p. 108878. ISSN: 0039128X. DOI: 10.1016/j.steroids.2021.108878.

Bolstad, B.M. et al. (Jan. 2003). “A comparison of normalization methods for high densityoligonucleotide array data based on variance and bias”. en. In: *Bioinformatics* 19.2, pp. 185–193. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/19.2.185.

Brown, Michael S and Joseph L Goldstein (May 1997). “The SREBP Pathway: Regulation of Cholesterol Metabolism by Proteolysis of a Membrane-Bound Transcription Factor”. en. In: *Cell* 89.3, pp. 331–340. ISSN: 00928674. DOI: 10.1016/S0092-8674(00)80213-5.

Choe, Mangyu et al. (May 2025). “Glycolysis model shows that allostery maintains high ATP and limits accumulation of intermediates”. en. In: *Biophysical Journal* 124.10, pp. 1562–1586. ISSN: 00063495. DOI: 10.1016/j.bpj.2025.03.037.

Diamant, Ido et al. (Jan. 2025). “Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources”. en. In: *Nucleic Acids Research* 53.D1, pp. D1016–D1028. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkae1080.

Düvel, Katrin et al. (July 2010). “Activation of a Metabolic Gene Regulatory Network Downstream of mTOR Complex 1”. en. In: *Molecular Cell* 39.2, pp. 171–183. ISSN: 10972765. DOI: 10.1016/j.molcel.2010.06.022.

Ferreira, Pedro G. et al. (Feb. 2018). “The effects of death and post-mortem cold ischemia on human tissue transcriptomes”. en. In: *Nature Communications* 9.1, p. 490. ISSN: 2041-1723. DOI: 10.1038/s41467-017-02772-x.

Foresti, Ombretta et al. (July 2013). “Sterol homeostasis requires regulated degradation of squalene monooxygenase by the ubiquitin ligase Doa10/Teb4”. en. In: *eLife* 2, e00953. ISSN: 2050-084X. DOI: 10.7554/eLife.00953.

Fred, Rikard G. et al. (Aug. 2022). “Single-cell transcriptome and cell type-specific molecular pathways of human non-alcoholic steatohepatitis”. en. In: *Scientific Reports* 12.1, p. 13484. ISSN: 2045-2322. DOI: 10.1038/s41598-022-16754-7.

Frost, H. Robert (June 2021). “Analyzing cancer gene expression data through the lens of normal tissue-specificity”. en. In: *PLOS Computational Biology* 17.6. Ed. by Shuangge Ma, e1009085. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1009085.

Gallego Romero, Irene et al. (Dec. 2014). “RNA-seq: impact of RNA degradation on transcript quantification”. en. In: *BMC Biology* 12.1, p. 42. ISSN: 1741-7007. DOI: 10.1186/1741-7007-12-42.

Gao, Chuan et al. (July 2016). "Context Specific and Differential Gene Co-expression Networks via Bayesian Biclustering". en. In: *PLOS Computational Biology* 12.7. Ed. by Xianghong Jasmine Zhou, e1004791. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004791.

Gao, Lucy L., Jacob Bien, and Daniela Witten (Oct. 2022). *Selective Inference for Hierarchical Clustering*. en. arXiv:2012.02936 [stat]. DOI: 10.48550/arXiv.2012.02936.

Goldstein, Joseph L and Michael S Brown (Feb. 1990). "Regulation of the mevalonate pathway". In: *Nature* 343, pp. 425–430.

Gomez-Sanchez, Celso E and Elise P Gomez-Sanchez (Feb. 2024). "Cholesterol Availability and Adrenal Steroidogenesis". en. In: *Endocrinology* 165.4, bqae032. ISSN: 1945-7170. DOI: 10.1210/endocr/bqae032.

Good, Benjamin M et al. (Oct. 2021). "Reactome and the Gene Ontology: digital convergence of data resources". en. In: *Bioinformatics* 37.19. Ed. by Zhiyong Lu, pp. 3343–3348. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btab325.

Gorin, Gennady et al. (Dec. 2022). "Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments". en. In: *Nature Communications* 13.1, p. 7620. ISSN: 2041-1723. DOI: 10.1038/s41467-022-34857-7.

Granados, Alejandro A., Nivedita Kanrar, and Michael B. Elowitz (Jan. 2024). "Combinatorial expression motifs in signaling pathways". en. In: *Cell Genomics* 4.1, p. 100463. ISSN: 2666979X. DOI: 10.1016/j.xgen.2023.100463.

Groen, Albert K. et al. (June 2014). "Cross-talk between liver and intestine in control of cholesterol and energy homeostasis". en. In: *Molecular Aspects of Medicine* 37, pp. 77–88. ISSN: 00982997. DOI: 10.1016/j.mam.2014.02.001.

GTEx Consortium (Oct. 2017). "Genetic effects on gene expression across human tissues". en. In: *Nature* 550.7675, pp. 204–213. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature24277.

Hu, Jie et al. (2010). "Cellular cholesterol delivery, intracellular processing and utilization for biosynthesis of steroid hormones". en. In.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki (Jan. 2009). "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". en. In: *Nucleic Acids Research* 37.1, pp. 1–13. ISSN: 1362-4962, 0305-1048. DOI: 10.1093/nar/gkn923.

Kanehisa, Minoru and Susumu Goto (1999). "KEGG: Kyoto Encyclopedia of Genes and Genomes". en. In: *Nucleic Acids Research* 28.1.

Khandker, Luipa et al. (Mar. 2022). "Cholesterol biosynthesis defines oligodendrocyte precursor heterogeneity between brain and spinal cord". en. In: *Cell Reports* 38.9, p. 110423. ISSN: 22111247. DOI: 10.1016/j.celrep.2022.110423.

Kim-Hellmuth, Sarah et al. (Sept. 2020). "Cell type–specific genetic regulation of gene expression across human tissues". en. In: *Science* 369.6509, eaaz8528. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aaz8528.

Kraemer, Fredric B. (Feb. 2007). "Adrenal cholesterol utilization". en. In: *Molecular and Cellular Endocrinology* 265-266, pp. 42–45. ISSN: 03037207. DOI: 10.1016/j.mce.2006.12.001.

Kunes, Russell Z. et al. (July 2024). "Supervised discovery of interpretable gene programs from single-cell data". en. In: *Nature Biotechnology* 42.7, pp. 1084–1095. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-023-01940-3.

Lachmann, Alexander et al. (Apr. 2018). "Massive mining of publicly available RNA-seq data from human and mouse". en. In: *Nature Communications* 9.1, p. 1366. ISSN: 2041-1723. DOI: 10.1038/s41467-018-03751-6.

Langfelder, Peter and Steve Horvath (Dec. 2008). "WGCNA: an R package for weighted correlation network analysis". en. In: *BMC Bioinformatics* 9.1, p. 559. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-559.

Lu, Jonathan et al. (Jan. 2021). "Causal network inference from gene transcriptional time-series response to glucocorticoids". en. In: *PLOS Computational Biology* 17.1. Ed. by Christina S. Leslie, e1008223. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1008223.

Markarian, Nicholas, Barbara E Engelhardt, et al. (Nov. 2025). "k-spaces: Mixtures of Gaussian latent variable models". en. In: *bioArxiv*. DOI: doi:https://doi.org/10.1101/2025.11.24.690254.

Markarian, Nicholas, Kimberly M. Van Auken, et al. (Mar. 2024). "Enrichment on steps, not genes, improves inference of differentially expressed pathways". en. In: *PLOS Computational Biology* 20.3. Ed. by Marc Robinson-Rechavi, e1011968. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011968.

Milacic, Marija et al. (Jan. 2024). "The Reactome Pathway Knowledgebase 2024". en. In: *Nucleic Acids Research* 52.D1, pp. D672–D678. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkad1025.

Nassir, Fatiha et al. (Mar. 2015). "Pathogenesis and Prevention of Hepatic Steatosis". en. In: *Gastroenterology & Hepatology* 11.3, pp. 167–175.

Oliva, Meritxell et al. (Sept. 2020). "The impact of sex on gene expression across human tissues". en. In: *Science* 369.6509, eaba3066. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aba3066.

Otto Meyerhof (Jan. 1945). "The Origin of the Reaction of Harden and Young in Cell-free Alcoholic Fermentation". In: *Journal of Biological Chemistry* 157.1, pp. 105–119. DOI: https://doi.org/10.1016/S0021-9258(17)41631-0.

Papatheodorou, Irene et al. (Jan. 2018). "Expression Atlas: gene and protein expression across multiple studies and organisms". en. In: *Nucleic Acids Research* 46.D1, pp. D246–D251. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkx1158.

Perry, Andrew S. et al. (Dec. 2024). "A prognostic molecular signature of hepatic steatosis is spatially heterogeneous and dynamic in human liver". en. In: *Cell Reports Medicine* 5.12, p. 101871. ISSN: 26663791. DOI: 10.1016/j.xcrm.2024.101871.

Pierre Morell and Helga Jurevics (Apr. 1996). "Origin of Cholesterol in Myelin". In: *Neurochemical Research* 21.4, pp. 463–470.

Pierson, Emma et al. (May 2015). "Sharing and Specificity of Co-expression Networks across 35 Human Tissues". en. In: *PLOS Computational Biology* 11.5. Ed. by Isidore Rigoutsos, e1004220. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004220.

Piro, Rosario Michael et al. (Nov. 2011). "An atlas of tissue-specific conserved coexpression for functional annotation and disease gene prediction". en. In: *European Journal of Human Genetics* 19.11, pp. 1173–1180. ISSN: 1018-4813, 1476-5438. DOI: 10.1038/ejhg.2011.96.

Prendiville, Walter and Rengaswamy Sankaranarayanan (2017). "Chapter 2., Anatomy of the uterine cervix and the transformation zone". In: *Colposcopy and Treatment of Cervical Precancer*. IARC Technical Report 45. Lyon, France: International Agency for Research on Cancer.

Reiman, Mario et al. (Aug. 2017). "Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples". en. In: *The FASEB Journal* 31.8, pp. 3298–3308. ISSN: 0892-6638, 1530-6860. DOI: 10.1096/fj.201601031RR.

Robinson, Mark D and Alicia Oshlack (Mar. 2010). "A scaling normalization method for differential expression analysis of RNA-seq data". en. In: *Genome Biology* 11.3, R25. ISSN: 1474-760X. DOI: 10.1186/gb-2010-11-3-r25.

Saha, Ashis et al. (Nov. 2017). "Co-expression networks reveal the tissue-specific regulation of transcription and splicing". en. In: *Genome Research* 27.11, pp. 1843–1858. ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.216721.116.

Saher, Gesine et al. (Apr. 2005). "High cholesterol level is essential for myelin membrane growth". en. In: *Nature Neuroscience* 8.4, pp. 468–475. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn1426.

Sakakura, Yuki et al. (Aug. 2001). "Sterol Regulatory Element-Binding Proteins Induce an Entire Pathway of Cholesterol Synthesis". en. In: *Biochemical and Biophysical Research Communications* 286.1, pp. 176–183. ISSN: 0006291X. DOI: 10.1006/bbrc.2001.5375.

Simonen, Piia et al. (Jan. 2011). "Cholesterol synthesis is increased and absorption decreased in non-alcoholic fatty liver disease independent of obesity". en. In: *Journal of Hepatology* 54.1, pp. 153–159. ISSN: 01688278. DOI: 10.1016/j.jhep.2010.05.037.

Subramanian, Aravind et al. (Oct. 2005). “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”. en. In: *Proceedings of the National Academy of Sciences* 102.43, pp. 15545–15550. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0506580102.

The Gene Ontology Consortium et al. (May 2023). “The Gene Ontology knowledgebase in 2023”. en. In: *GENETICS* 224.1. Ed. by A Baryshnikova, iyad031. ISSN: 1943-2631. DOI: 10.1093/genetics/iyad031.

The GTEx Consortium et al. (May 2015). “The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans”. en. In: *Science* 348.6235, pp. 648–660. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1262110.

Thomas, Paul D., Michael J. Campbell, et al. (Sept. 2003). “PANTHER: A Library of Protein Families and Subfamilies Indexed by Function”. en. In: *Genome Research* 13.9, pp. 2129–2141. ISSN: 1088-9051. DOI: 10.1101/gr.772403.

Thomas, Paul D., David P. Hill, et al. (Oct. 2019). “Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems”. en. In: *Nature Genetics* 51.10, pp. 1429–1433. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-019-0500-1.

Ueda, Naonori and Ryohei Nakano (Mar. 1998). “Deterministic annealing EM algorithm”. en. In: *Neural Networks* 11.2, pp. 271–282. ISSN: 08936080. DOI: 10.1016/S0893-6080(97)00133-0.

Ville Satopaa et al. (June 2011). “Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior”. In: *ICDCSW ’11: Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops*. Minneapolis, MN, USA, pp. 166–171. DOI: 10.1109/ICDCSW.2011.20.

Wu, Chaodong, Salmaan A. Khan, and Alex J. Lange (Nov. 2005). “Regulation of glycolysis—role of insulin”. en. In: *Experimental Gerontology* 40.11, pp. 894–899. ISSN: 05315565. DOI: 10.1016/j.exger.2005.08.002.

Zhao, Shanrong, Zhan Ye, and Robert Stanton (Aug. 2020). “Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols”. en. In: *RNA* 26.8, pp. 903–909. ISSN: 1355-8382, 1469-9001. DOI: 10.1261/rna.074922.120.

Zhao, Yingdong et al. (Dec. 2021). “TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository”. en. In: *Journal of Translational Medicine* 19.1, p. 269. ISSN: 1479-5876. DOI: 10.1186/s12967-021-02936-w.

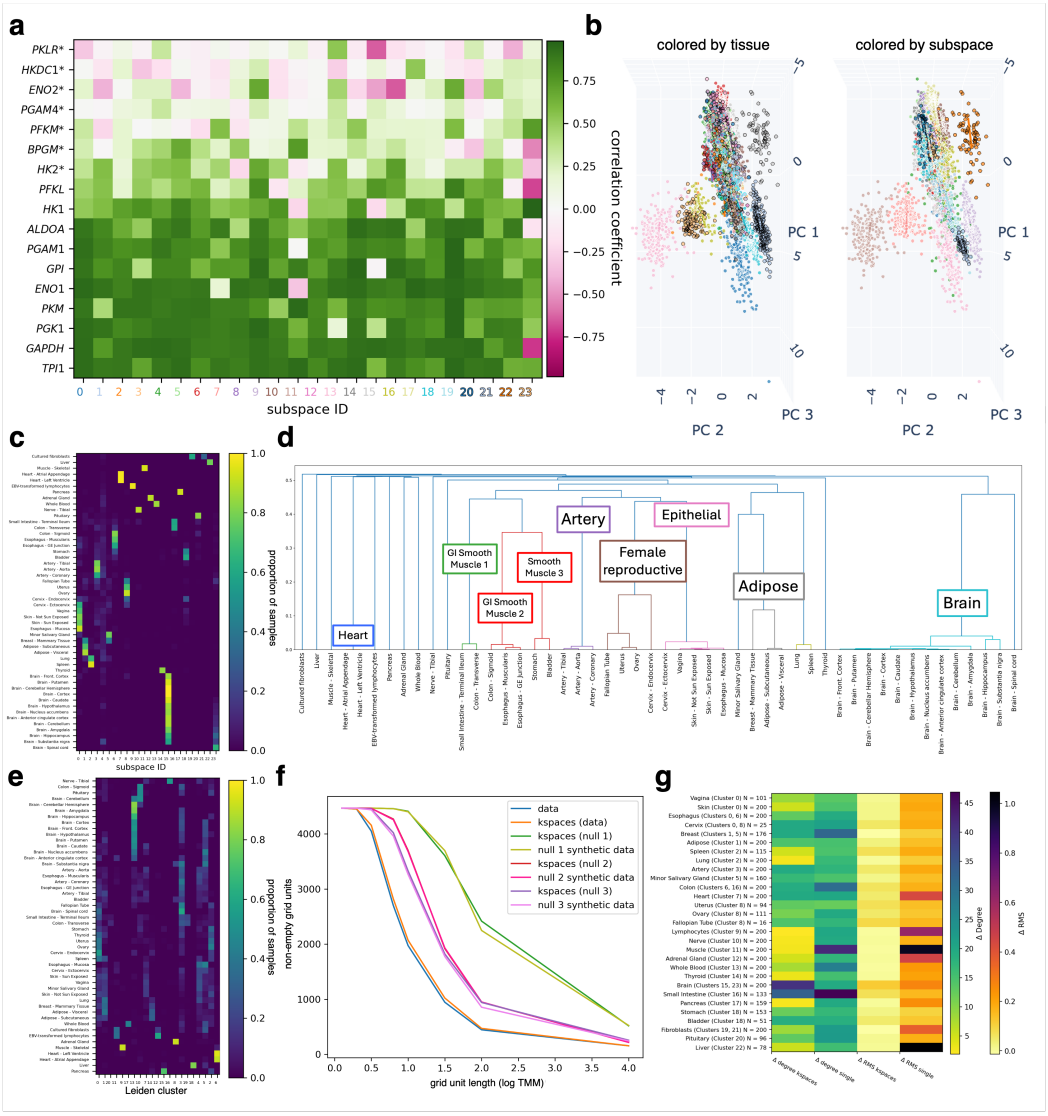


Figure 4.3: ***k*-spaces identifies tissue group-specific gene stoichiometry to characterize inter-individual expression scaling and scaling relationship diversity across tissues with 24 clusters at** $d = 3$**. a,** Correlation coefficients of the top latent variable (as shown in (b)) with expression values for each gene. Genes with an asterisk were not used in clustering. **b,** Left: tissue samples in PCA space colored by tissue subtype. Projection of the top eigenvector for each tissue subtype is shown with a line, extended 2 s.d. in either direction from the mean. Right: tissue samples in PCA space colored by *k*-spaces subspace assignment. Projection of the top basis vector for each subspace is shown with a line, extended 2 s.d. in either direction from the mean. While d=3 for the subspaces, only the basis vector with the greatest corresponding variance parameter is shown. Colors of subspace IDs on the x-axis in (a) serve as a legend for the subspace-cluster color scheme. **c,** Heatmap displaying the proportion of samples from each tissue assigned to each subspace. Color indicates the proportion of samples from a given tissue assigned to a subspace. **d,** Annotated dendrogram for tissue similarity using tissue-cluster assignments in (c).

Figure 4.3: **e,** Heatmap displaying the proportion of samples from each tissue assigned to each cluster by default Leiden clustering using scanpy. **f,** Compactness of actual data, simulated data produced by the k-space model fit to the data, and simulated data from three null hypotheses. **g,** Comparison for each tissue of the top eigenvector to the top basis vector of the representative k-spaces subspace and the single dataset-wide eigenvector from the null hypothesis of no clusters. As the data is log-transformed, vector translations correspond to angular differences in non log-transformed spaces, so comparison of both angle and translation (RMS difference) are shown.

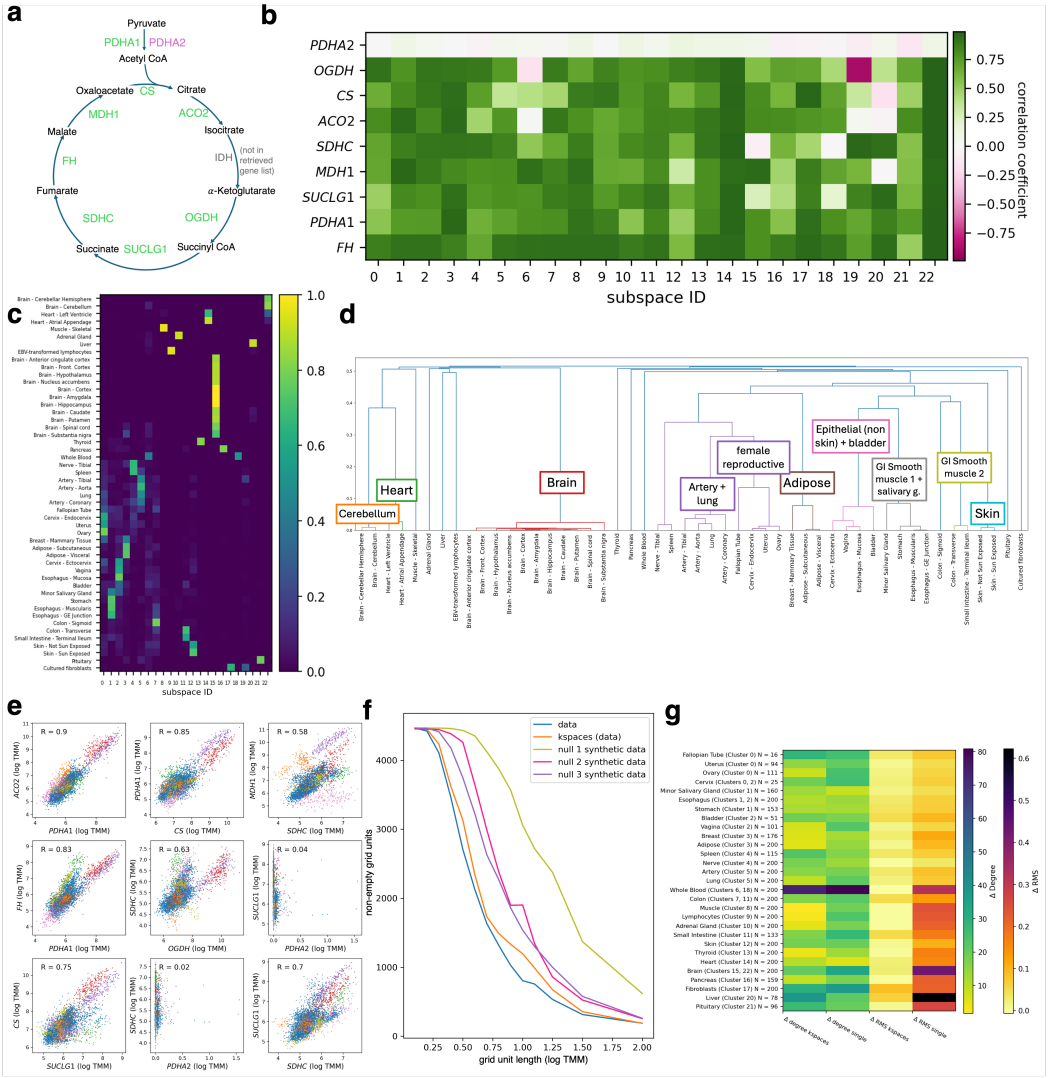


Figure 4.4: ***k*-spaces identifies inter-sample expression variation in the TCA cycle with 22 tissue groupings at $d = 2$ similar to those observed with the glycolysis pathway.**

Figure 4.4: **a,** TCA cycle pathway diagram. Genes colored green were correlated, but *PDHA2* was included in *k*-spaces clustering as well. *IDH* is part of the PANTHER Pathway but was not in the gene list retrieved from Harmonizome. **b,** Correlation coefficients of TCA genes with the top basis vector of each *k*-spaces subspace show that the learned latent variable can represent gene expression changes well for most genes. Clustering was performed with $d = 2$, meaning there are two basis vectors corresponding to variables, but only the first is shown here. **c,** Assignments of samples of each tissue subtype to each subspace. Color indicates the proportion of samples from a given tissue assigned to a subspace. **d,** Hierarchical clustering on the correlation of cluster assignments between tissues reproduces a similar grouping as for the glycolysis pathway. Exceptions include: separation of cerebellum from other brain regions, inclusion of the spinal cord with other brain regions, inclusion of the lung in the artery group, inclusion of bladder among the epithelial tissue group, separation of skin subtypes from other epithelial tissues, and inclusion the minor salivary gland with G.I. smooth muscle. Fibroblasts, liver, skeletal muscle, lymphocytes, pancreas, adrenal gland, whole blood, and pituitary gland samples are again assigned their own subspaces. **e,** Pairwise scatter plots of selected TCA genes, colored by tissue using the same color palette as in Fig 4.2e. Heart (purple) and skeletal muscle (red) generally have higher expression and tight correlations between genes. **f,** *k*-spaces represents the data with more compact probability density than competing null hypotheses, as in Fig 4.3f. **g,** Top eigenvectors of clusters learned by *k*-spaces provide a better representation of most tissues than the top eigenvector of all data together.

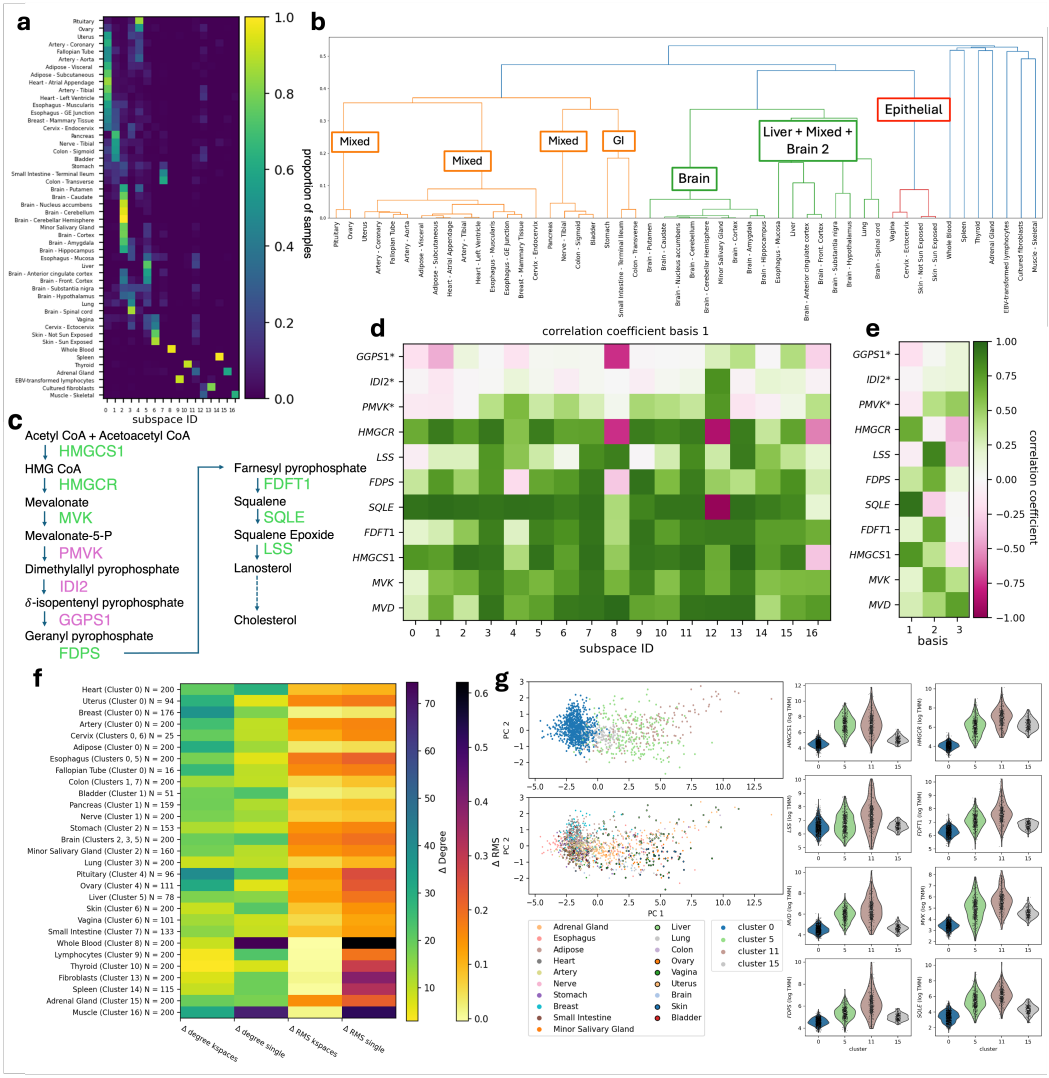


Figure 4.5: ***k*-spaces merges tissue groups together into a low expression state for the de novo cholesterol synthesis pathway, identifying 17 effective clusters with** $d = 3$**.** *k*-spaces does not recover distinct tissue profiles for cholesterol synthesis and merges many tissues into a basal expression state, with basis vectors correlating with known regulators. **a,** Heatmap displaying assignments of samples from each tissues to clusters. Color represents the proportion of samples from each tissue assigned to a cluster. Cluster 0 groups many tissues together that were not assigned to the same subspace for the glycolysis or TCA pathways. Similarly, liver samples are not uniquely and wholly assigned to their own subspace. Most liver samples are assigned to Subspace 5, which also represents over half of brain frontal cortex BA9 and brain anterior cingulate cortex BA24 (also located in the frontal cortex) samples. **b,** Hierarchical clustering of tissues on the basis of correlation between their subspace assignment proportions. *k*-spaces subspaces do not have the same tissue specificity as in the glycolysis pathway and TCA cycle, reflecting the heatmap in (a).

Figure 4.5: **c,** Cholesterol biosynthesis pathway diagram. Genes colored green were correlated, shown in Fig 4.17a. **d,** The top latent variable for each subspace generally correlates with most genes except *HMGCR* and *SQLE*, two genes with negative feedback regulation that are negatively correlated with the top latent variable in some subspaces. The primary latent variable in Subspace 0 only correlates with *HMGCS1*, *HMGCR*, and *SQLE*. Asterisks indicate genes held out from clustering. **e,** Correlation coefficients for all latent variables (basis vectors 1-3) for Subspace 0. Basis vector 2 correlates with most remaining genes, while basis 3 appears to represent expression of MVD. **f,** The top basis vector of k-spaces Space 0 is not reflective of the primary axis of variation within each tissue computed using tissue subtype labels. **g,** Data from Clusters 0, 5, 11, and 15 projected onto its first two principal components, colored by cluster above and tissue type below, with violin plots of gene expression. Cluster 5 includes liver and some brain regions, Cluster 15 is mostly adrenal gland samples, and Cluster 11 includes both liver and adrenal gland samples. All clusters are shown in Fig 4.18. Cluster 0 has low expression, particularly for *HMGCS1*, *HMGCR*, and *SQLE*. Clusters 5, 11, and 15, representing tissues with known higher de novo synthesis, have higher expression.

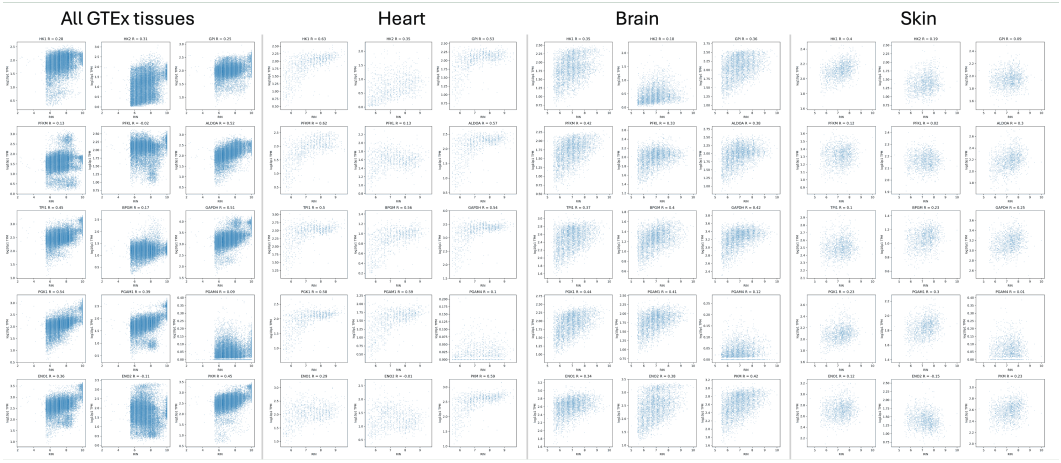


Figure 4.6: **Effect of RNA degradation on measured gene expression.** Sample RIN is plotted against log10p1 TPM expression of genes in the glycolysis pathway. Heart and brain are shown individually as representative examples of tissues with a strong interaction between RIN and expression, and skin is shown as an example of a tissue with weak interaction. The plots were used to determine a lower bound of RIN = 7.0 for filtering out samples before regression was performed to remove the effect of RIN variation. Combination of all GTEx tissues suggested gene expression changes with RIN may be tissue-specific. Heart and brain samples also exhibited greater variance with lower RIN, which sharply increased below RIN < 7 for heart samples. Skin and many other tissues did not show substantial effects of RIN on glycolysis gene expression.

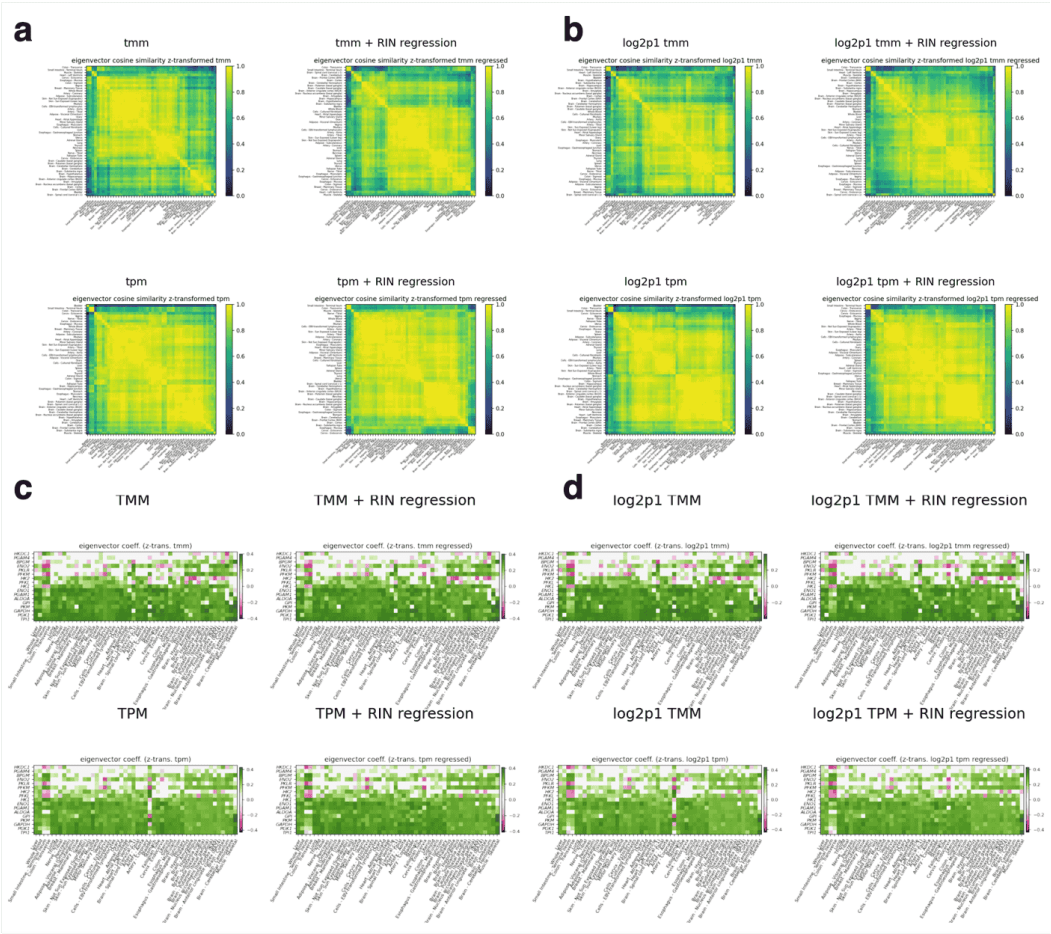


Figure 4.7: **Comparison of preprocessing strategies on glycolysis data.** TMM vs. TPM, the log2p1 transformation versus no variance-stabilizing transformation, and regressing out the effect of RNA degradation using the RIN as a proxy (Fig 4.6) versus no regression are compared. **a,** z-transformed eigenvector cosines (equivalent to the dot product because the vectors are scaled to a length of 1) as in Fig 4.2b for TMM and TPM, with and without regression to remove the effect of RIN. **b,** Heatmaps as in (a) but with log-transformed gene expression. Regression was performed after log-transformation. **c,** Gene coefficients in the eigenvectors in (a) as in Fig 4.2c for TMM, with and without regression of RIN. There appeared to be little qualitative difference resulting from the regression, but eigenvector cosines were slightly lower with TMM versus TPM, in contrast to random genes in Figs 4.9,4.10,4.11,4.12,4.13 where high cosine similarities were abundant with TPM but not TMM. Coefficients between -0.1 and 0.1 are displayed white (set to 0) for ease of visualization of insignificant coefficients given small plot size. **d,** Gene coefficients in the eigenvectors in (b) as in Fig 4.2c for log-transformed TMM, with and without regression of RIN. There again appeared to be little qualitative difference resulting from the regression. Plots looked qualitatively similar whether or not the

Figure 4.7: log transformation was used. There appeared to be little qualitative difference between TMM and TPM, in contrast to random genes where positive nonzero coefficients were abundant with TPM but not TMM. Coefficients between -0.1 and 0.1 are again displayed white.

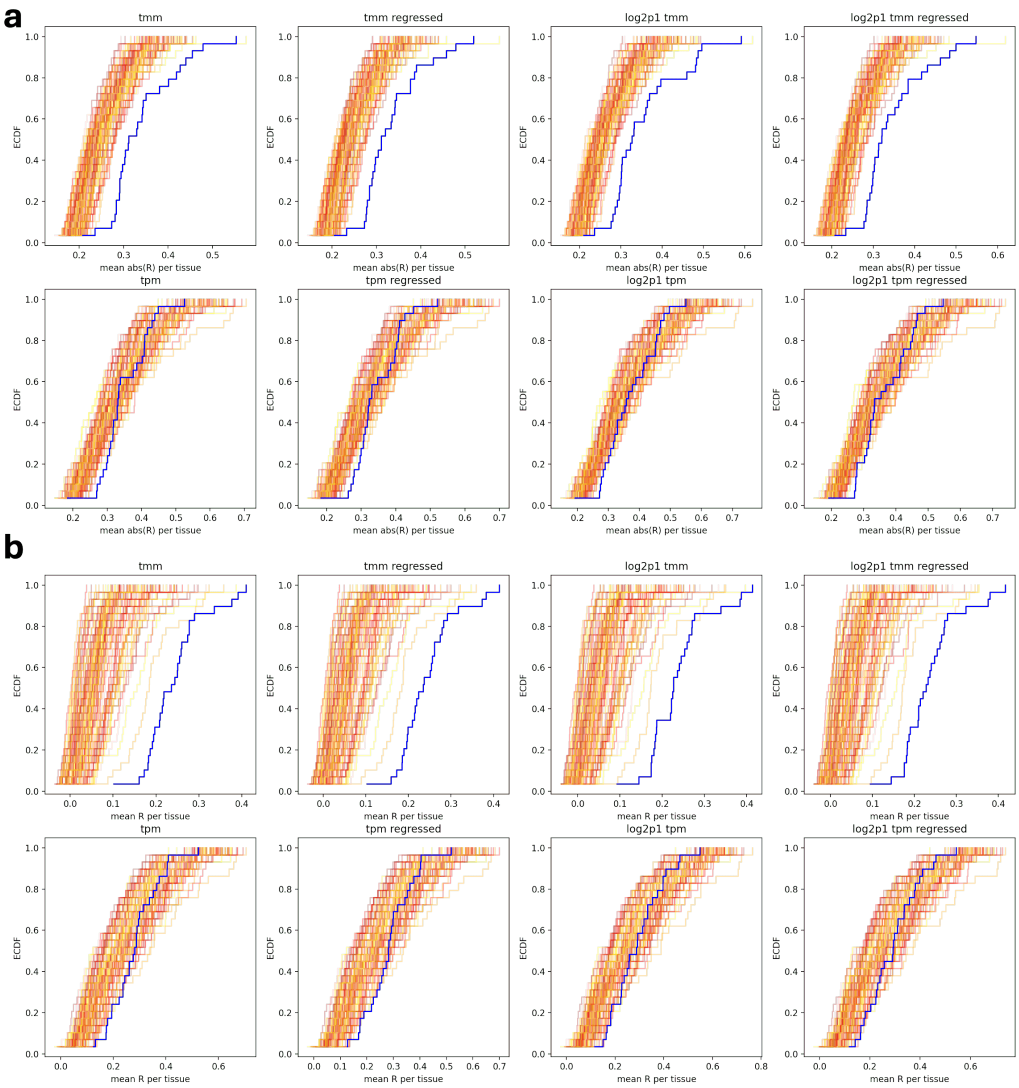


Figure 4.8: **ECDFs of mean correlation coefficients for different preprocessing strategies in glycolysis and randomly selected genes, as in Fig 4.2f.** Each ECDF displays the distribution of the mean Pearson correlation coefficient of gene-gene correlation matrices over tissues (29 points per curve for 29 tissues) for a different set of genes. Blue represents the glycolysis genes. Warm colors represent 100 random sets of genes. Plot titles indicate the preprocessing strategy. **a,** Mean correlation coefficient. **b,** Mean absolute correlation coefficient for the same gene sets.

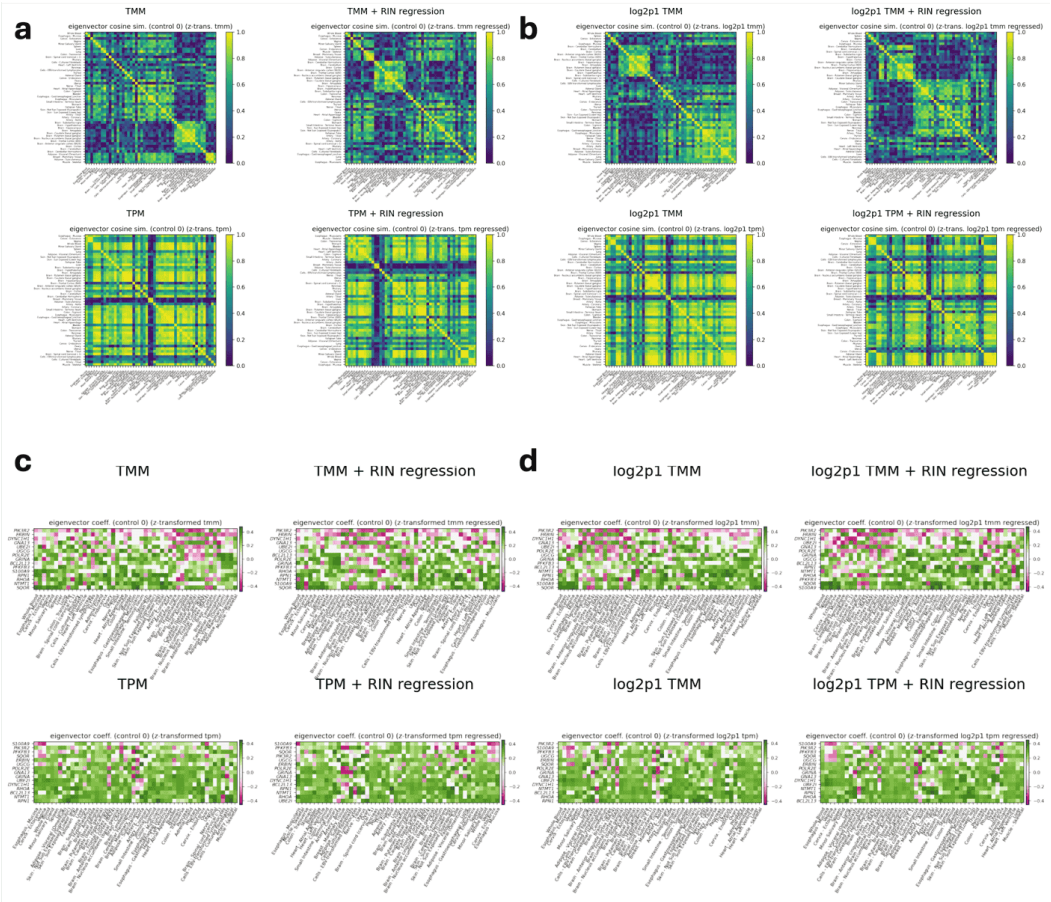


Figure 4.9: **Comparison of preprocessing strategies on a random set of 15 genes.** The genes are: *S100A9, PIK3R2, PFKFB3, SQOR, ERBIN, UGCG, POLR2E, GNA13, GRINA, UBE2I, DYNC1H1, RHOA, BCL2L13, NTMT1*, and *RPN1*. TMM vs. TPM, the log2p1 transformation versus no variance-stabilizing transformation, and regressing out the effect of RNA degradation using the RIN as a proxy (Fig 4.6) versus no regression are compared, as in Fig 4.7. **a,** z-transformed eigenvector cosines (equivalent to the dot product because the vectors are scaled to a length of 1) as in Fig 4.2b for TMM and TPM, with and without regression to remove the effect of RIN. **b,** Heatmaps as in (a) but with log-transformed gene expression. Regression was performed after log-transformation. **c,** Gene coefficients in the eigenvectors in (a) for TMM, with and without regression of RIN. Coefficients between -0.1 and 0.1 are displayed white (set to 0) for ease of visualization of insignificant coefficients given small plot size in (c) and (d). **d,** Gene coefficients in the eigenvectors in (b) for log-transformed TMM, with and without regression of RIN.

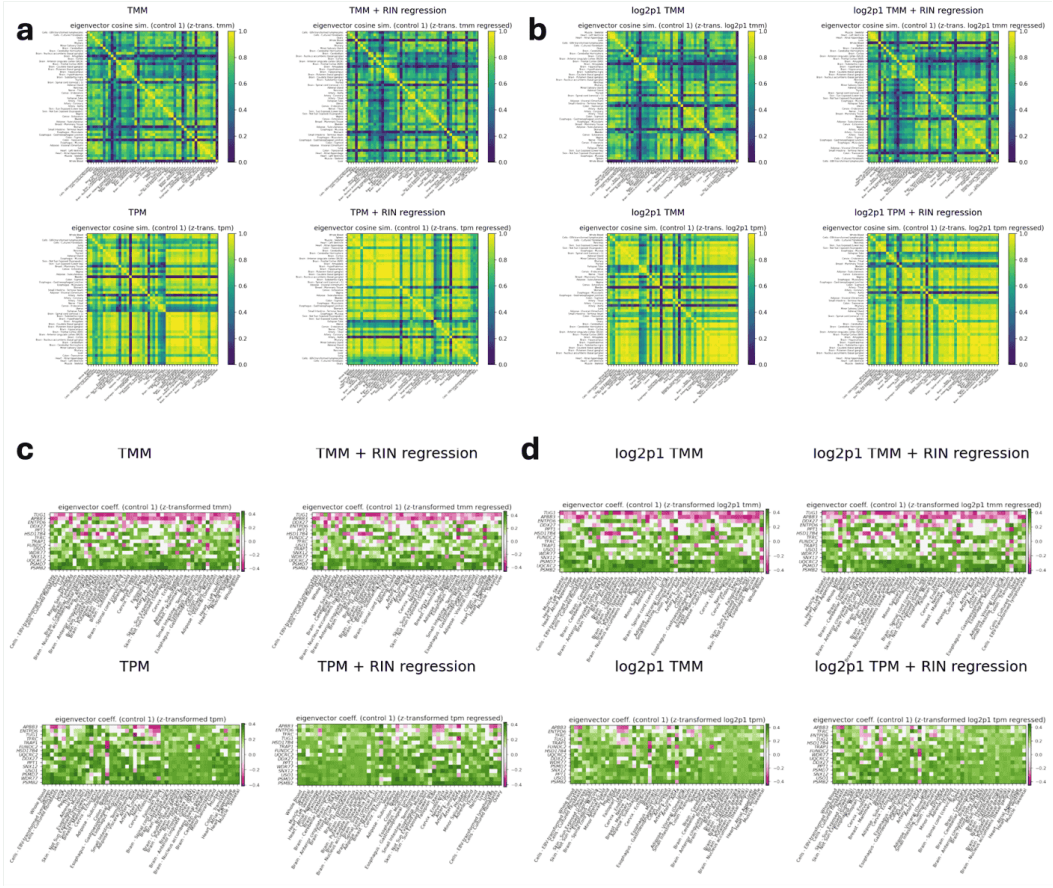


Figure 4.10: **Comparison of preprocessing strategies on a random set of 15 genes.** The genes are: *APBB3, ENTPD6, TUG1, TFRC, TRAP1, FUNDC2, HSD17B4, UQCRC2, HSD17B4, UQCRC2, DDX27, PPT1, SNX12, USO1, PSMD7, WDR77*, and *PSMB2*. TMM vs. TPM, the log2p1 transformation versus no variance-stabilizing transformation, and regressing out the effect of RNA degradation using the RIN as a proxy (Fig 4.6) versus no regression are compared, as in Fig 4.7. **a,** z-transformed eigenvector cosines (equivalent to the dot product because the vectors are scaled to a length of 1) as in Fig 4.2b for TMM and TPM, with and without regression to remove the effect of RIN. **b,** Heatmaps as in (a) but with log-transformed gene expression. Regression was performed after log-transformation. **c,** Gene coefficients in the eigenvectors in (a) for TMM, with and without regression of RIN. Coefficients between -0.1 and 0.1 are displayed white (set to 0) for ease of visualization of insignificant coefficients given small plot size in (c) and (d). **d,** Gene coefficients in the eigenvectors in (b) for log-transformed TMM, with and without regression of RIN.

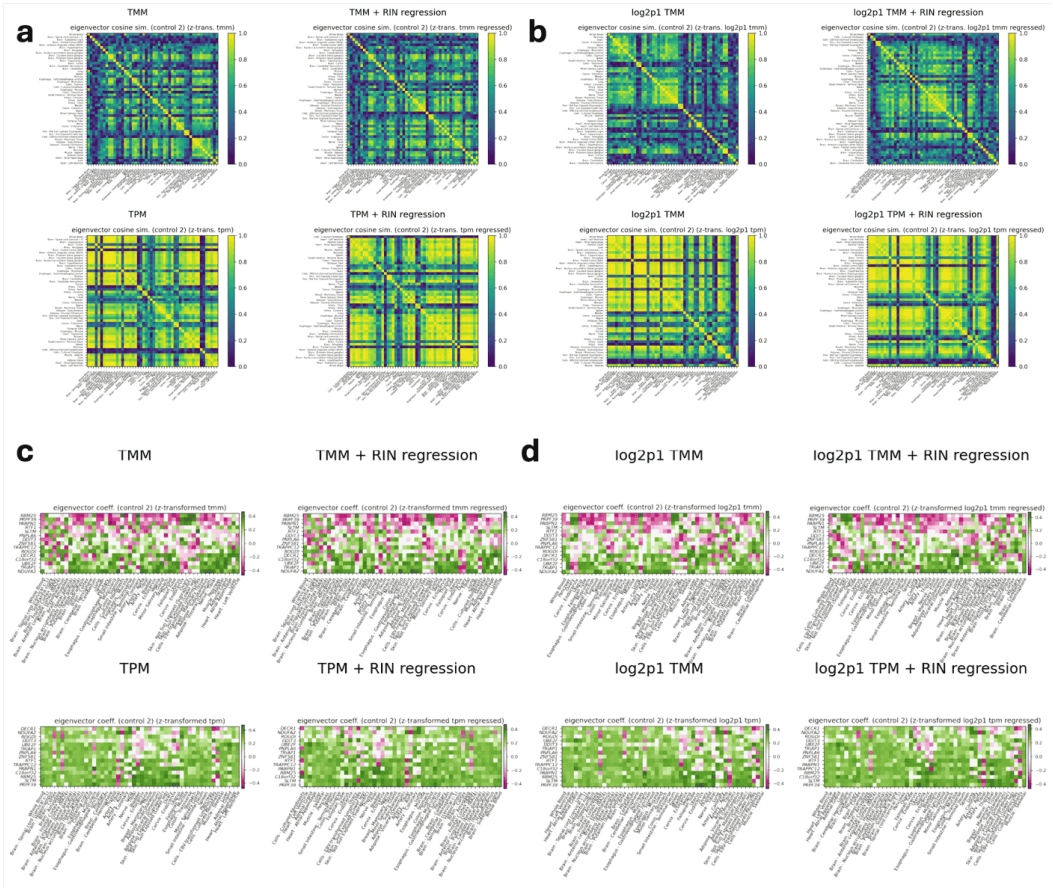


Figure 4.11: **Comparison of preprocessing strategies on a random set of 15 genes.** The genes are: *DECR1, NDUFA2, ROGDI, DDIT3, UBE2F, TRIAP1, PNPLA6, ZNF581, RTF1, TRAPPC12, PABPN1, C18orf32, RBM25, SLTM*, and *PRPF39*. TMM vs. TPM, the log2p1 transformation versus no variance-stabilizing transformation, and regressing out the effect of RNA degradation using the RIN as a proxy (Fig 4.6) versus no regression are compared, as in Fig 4.7. **a,** z-transformed eigenvector cosines (equivalent to the dot product because the vectors are scaled to a length of 1) as in Fig 4.2b for TMM and TPM, with and without regression to remove the effect of RIN. **b,** Heatmaps as in (a) but with log-transformed gene expression. Regression was performed after log-transformation. **c,** Gene coefficients in the eigenvectors in (a) for TMM, with and without regression of RIN. Coefficients between -0.1 and 0.1 are displayed white (set to 0) for ease of visualization of insignificant coefficients given small plot size in (c) and (d). **d,** Gene coefficients in the eigenvectors in (b) for log-transformed TMM, with and without regression of RIN.

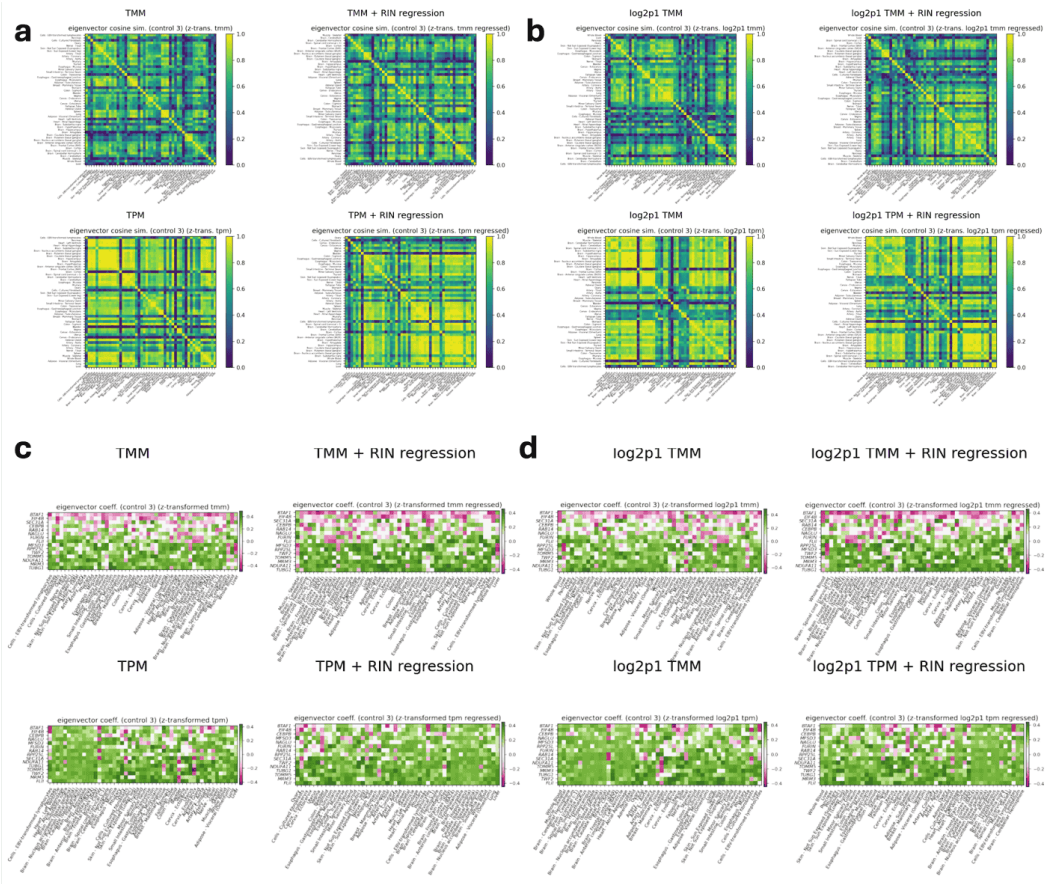


Figure 4.12: **Comparison of preprocessing strategies on a random set of 15 genes.** They are: *BTAF1, EIF4B, CEBPB, NAGLU, MFSD3, FURIN, RAB14, RPP25L, SEC31A, NDUFA11, TUBG1, TOMM5, TWF2, MRM3*, and *FLII*. TMM vs. TPM, the log2p1 transformation versus no variance-stabilizing transformation, and regressing out the effect of RNA degradation using the RIN as a proxy (Fig 4.6) versus no regression are compared, as in Fig 4.7. **a,** z-transformed eigenvector cosines (equivalent to the dot product because the vectors are scaled to a length of 1) as in Fig 4.2b for TMM and TPM, with and without regression to remove the effect of RIN. **b,** Heatmaps as in (a) but with log-transformed gene expression. Regression was performed after log-transformation. **c,** Gene coefficients in the eigenvectors in (a) for TMM, with and without regression of RIN. Coefficients between -0.1 and 0.1 are displayed white (set to 0) for ease of visualization of insignificant coefficients given small plot size in (c) and (d). **d,** Gene coefficients in the eigenvectors in (b) for log-transformed TMM, with and without regression of RIN.

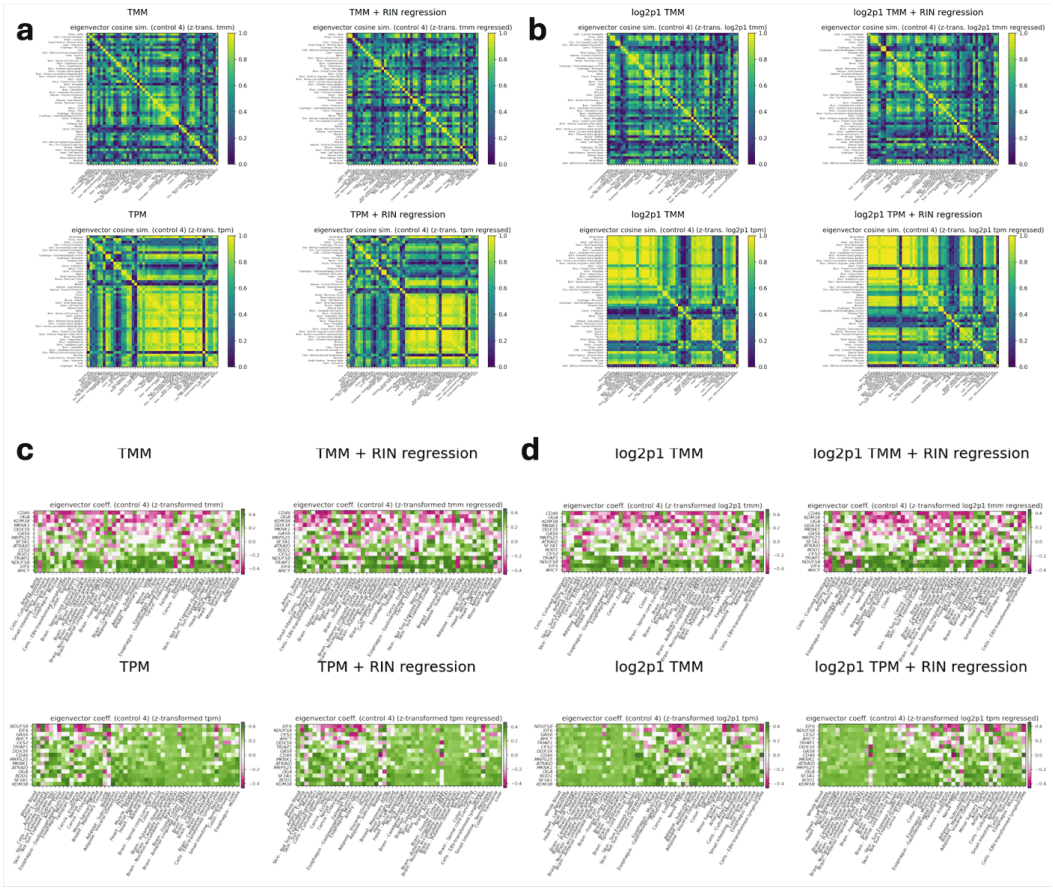


Figure 4.13: **Comparison of preprocessing strategies on a random set of 15 genes.** The genes are: *NDUFS8, EIF6, GAS6, AHCY, CES2, TRIAP1, DDX3X, CD46, MRPS25, MKNK1, ATRAID, OGA, BOD1, SF3A1*, and *KDM3B*. TMM vs. TPM, the log2p1 transformation versus no variance-stabilizing transformation, and regressing out the effect of RNA degradation using the RIN as a proxy (Fig 4.6) versus no regression are compared, as in Fig 4.7. **a,** z-transformed eigenvector cosines (equivalent to the dot product because the vectors are scaled to a length of 1) as in Fig 4.2b for TMM and TPM, with and without regression to remove the effect of RIN. **b,** Heatmaps as in (a) but with log-transformed gene expression. Regression was performed after log-transformation. **c,** Gene coefficients in the eigenvectors in (a) for TMM, with and without regression of RIN. Coefficients between -0.1 and 0.1 are displayed white (set to 0) for ease of visualization of insignificant coefficients given small plot size in (c) and (d). **d,** Gene coefficients in the eigenvectors in (b) for log-transformed TMM, with and without regression of RIN.

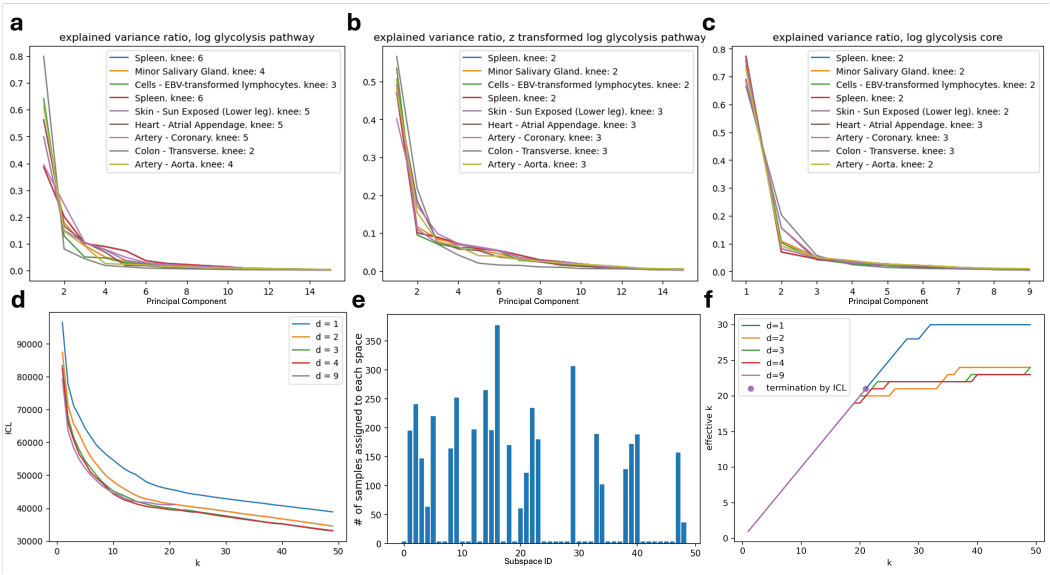


Figure 4.14: **Plots used in choosing *k*-spaces parameters *d* and *k*.** (a) - (c) were used to choose $d = 3$ based on 9 randomly chosen tissue subtypes (Brain subtypes were excluded as all subtypes were likely affected by cell composition variation), while (d) and (e) were used to choose k. **a,** Scree plot of principal component versus proportion of variance explained for log2p1-transformed gene expression in glycolysis pathway. Elbow detection performed with the kneedle algorithm (Ville Satopaa et al., 2011) is listed in the figure caption. Most elbows were at 4, 5, or 6 PCs, suggesting a higher dimensional PCA would be needed to represent the data. **b,** Scree plot as in (a) but with a z-tranformation (centering and scaling features to a variance of 1) applied to the log-transformed glycolysis expression data, a typical preprocessing step for PCA. Detected elbows are 2 or 3, suggesting a lower dimensional correlation structure between genes that can be well-represented with 2-3 latent variables. **c,** Scree plot of the same tissue subtypes but with log2p1-transformed gene expression of the selected "core" glycolysis genes (*HK1, GPI, PFKL, ALDOA, TPI1, GAPDH, PGK1, PGAM1, ENO1, PKM*). No z-transformation was applied, as applying such a transformation dataset wide would not necessarily rescale features to equal variance within individual clusters if distinct clusters are present. Elbow detection suggested that with the selected set of genes, 2-3 latent variables would be sufficient, so $d = 3$ was chosen for clustering in Fig 4.3. **d,** k vs Integrated completed likelihood (ICL) for clustering with various settings for d. d = 1,2,4, and 9 (a Gaussian Mixture Model) were included for comparison. Rather than flattening out as k increased, ICL decreased linearly for d = 1-4 for k > 20-25. **d,** Bar plot of the number of samples in each cluster shows sparsity of cluster usage at $k = 49$ for $d = 3$. Most clusters were "empty" and contained the minimum number of samples (4 = d +1), suggesting there was an "effective" value of k determined by the EM algorithm, similar to Bayesian Gaussian Mixture models in scikit-learn using an "effective" number of mixture components by setting the some of their weights close to 0. This was also observed for other values of k >

Figure 4.14: 20-25 and for other d (data not shown) but is quantified in (e). Subspace ID numbers do not correspond to those in the main text, as IDs were reassigned by the number of points per cluster and then "empty" subspaces were dropped. **f,** "Effective" k (as defined in (e)) versus k suggested that the choice of k was unimportant provided it was large enough. Transition points roughly corresponded to the k vs ICL curves straightening out in (d).

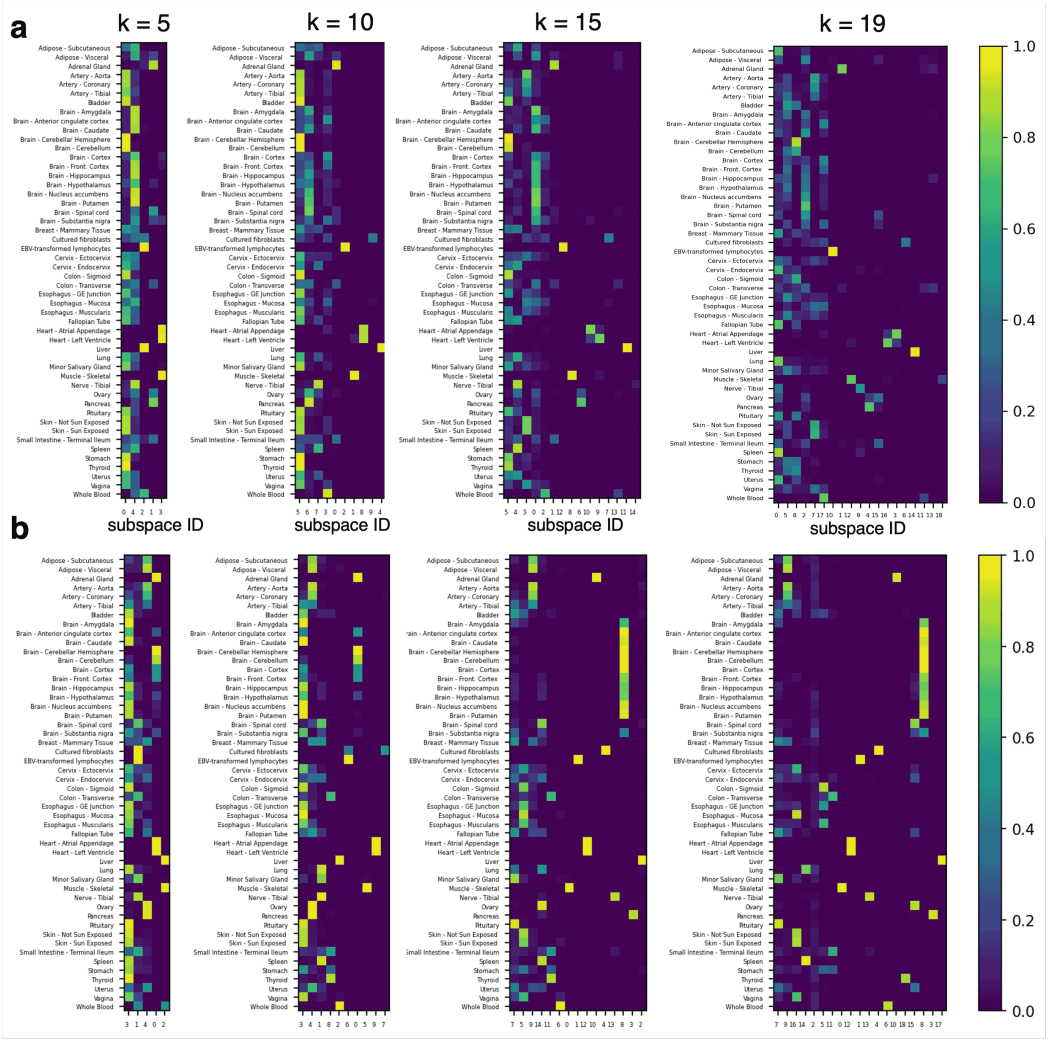


Figure 4.15: **Proportion of samples of each tissue subtype assigned to each cluster at various k when k-spaces was run on 3-D PCA data.** Rather than cluster in high dimensional space, clustering was run in the 3-D PCA space used to visualize samples to demonstrate that k-spaces was using information in the high dimensional space to separate samples of different tissue subtypes. For $k > 5$, k-spaces output on the high dimensional data clustered samples with significantly better alignment (sparser heatmap with brighter yellow indicating samples of the same tissue are more likely to be clustered together) with tissue subtype labels than k-spaces output on the 3-D PCA data. **a,** k-spaces output (d=1) on 3-D data of

Figure 4.15: 3-D PCA of the subset of 10 correlated glycolysis genes. **b,** k-spaces (d=3) output on the 10-D data of the glycolysis genes, as in the main text.

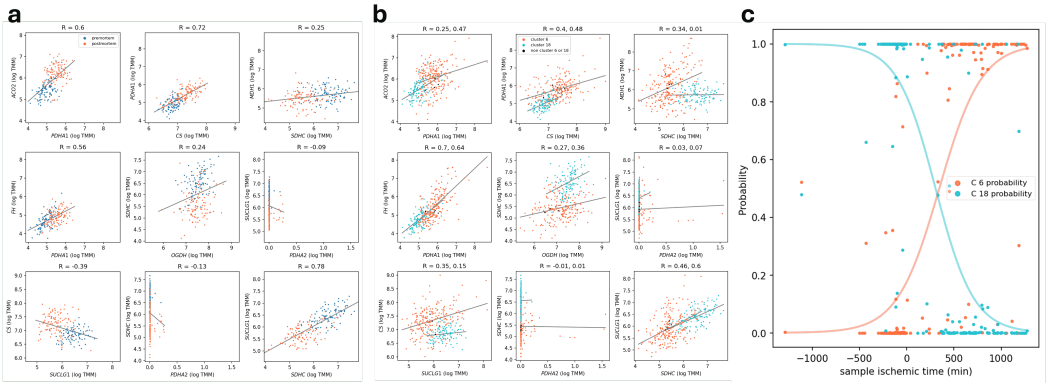


Figure 4.16: **Clustering separates premortem and postmortem whole blood samples. a,** Pairwise scatter plots of whole blood samples and randomly selected genes from the TCA cycle, colored by premortem or postmortem status. Each sample is from a different donor. Linear regressions and Pearson correlation coefficients for the gene pair over all whole blood samples is displayed. **b,** Pairwise scatter plots of the same genes and for all whole blood samples and any additional non-whole blood samples assigned to Clusters 6 or 18. Points are colored by cluster assignment. Each sample is from a different donor. Linear regressions and Pearson correlation coefficients for the gene pair over all samples in a cluster (regardless of tissue identity) are displayed. **c,** Ischemic time versus soft-assignment probabilities outputted by k-spaces for whole blood samples. Samples with negative ischemic times are premortem, and those with positive times are postmortem. Logistic regression curves on hard-cluster assignments are shown. They do not necessarily add to 1 as some whole blood samples are not assigned to Clusters 6 or 18.

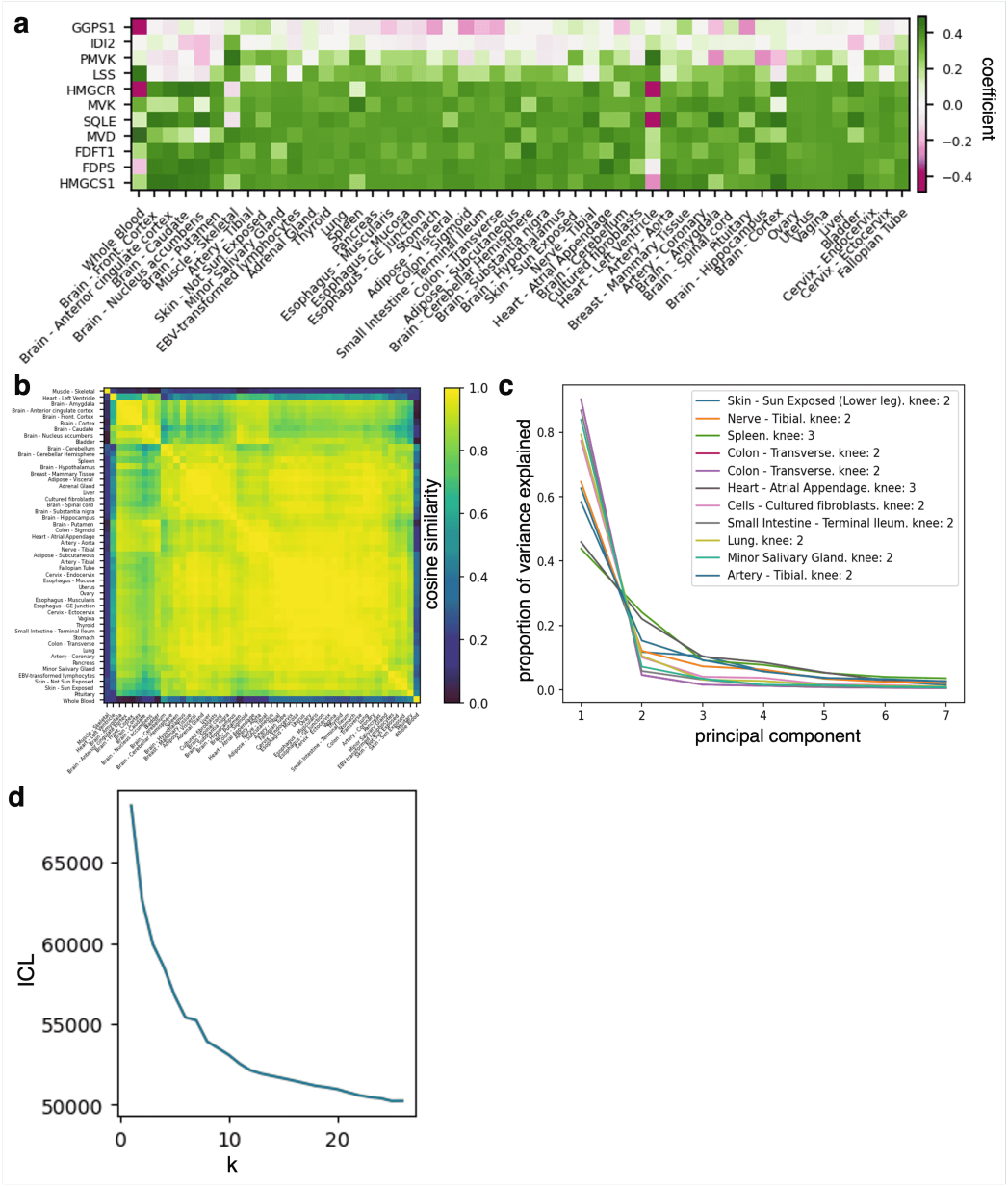


Figure 4.17: **Selection of genes from the de novo cholesterol synthesis pathway and selection of d and k for k-spaces clustering. a,** Coefficients of eigenvectors for each tissue subtype in its own z-transformed log space. *LSS, HMGCR, MVK, SQLE, MVD, FDFT1, FDPS,* and *HMGCS1* were selected for k-spaces clustering. **b,** Heatmap of cosine similarity of top log space eigenvectors for each tissue subtype showing very high similarity between tissues. **c,** Proportion of variance explained versus principal component for randomly selected tissue subtypes. Knees detected with the Kneedle algorithm are shown in caption. $d = 3$ was chosen as a result. **d,** Successive k-spaces clustering was halted at $k = 26$ because the ICL criterion increased.

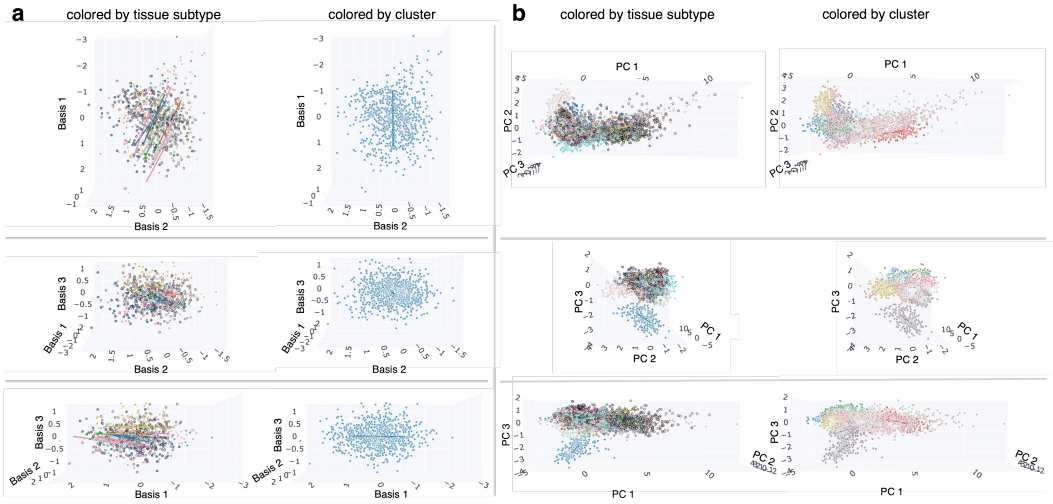


Figure 4.18: **Data colored by tissue type and cluster for the 8 selected genes in the de novo cholesterol synthesis pathway.** Colored by tissue subtype (left) and cluster (right), with the projections of the top eigenvectors of tissue subtypes or basis vector 1 of k-spaces subspaces included. **a,** Data from Cluster 0 projected onto the three basis vectors of Cluster 0. **b,** All samples projected onto the principal components of the data overall.

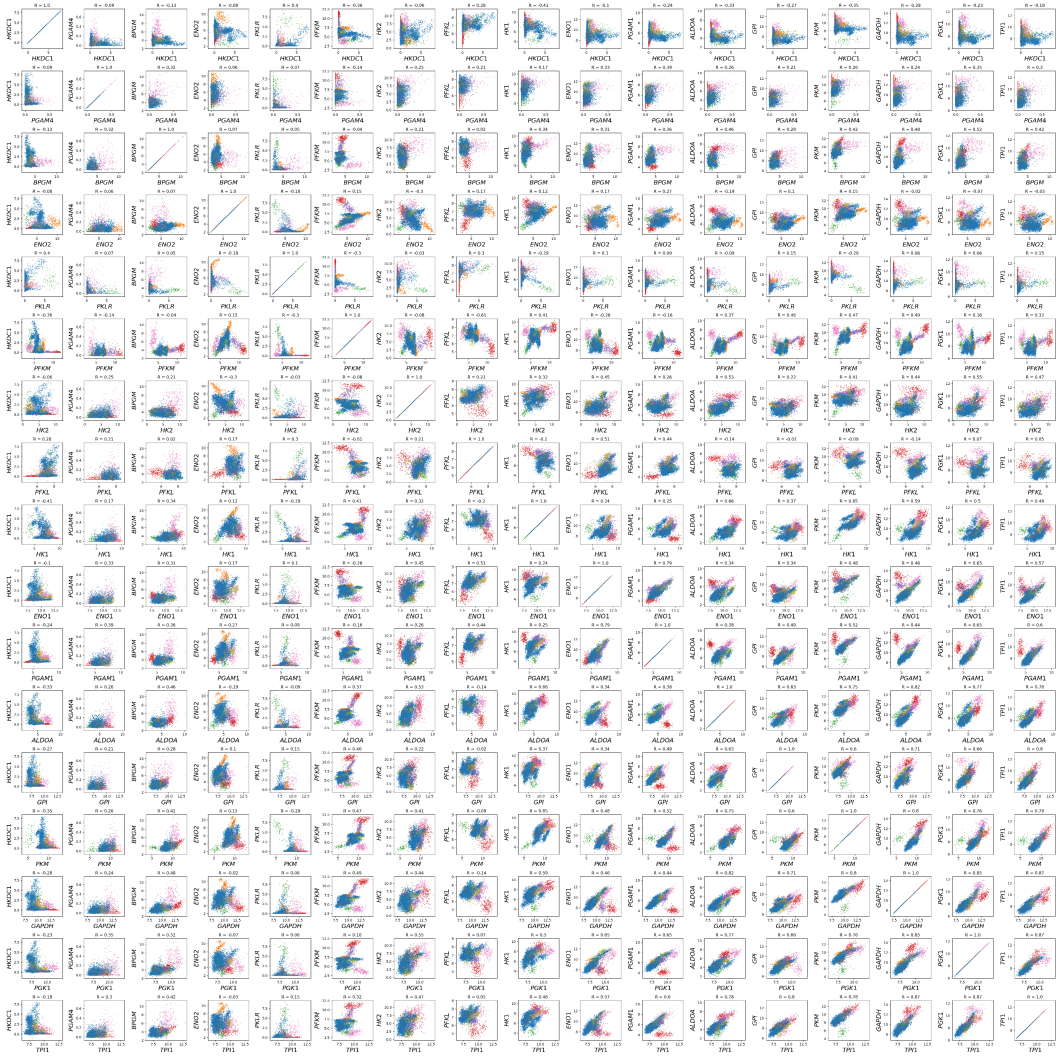


Figure 4.19: **Pairwise scatter plots of glycolysis genes.** Genes are ordered by mean coefficient in Fig 4.2c, and samples are colored by tissue type, as in Fig 4.2e: skeletal muscle - red, liver - green, brain - orange, heart - purple, whole blood - pink, adipose - brown, fibroblasts - olive green, other tissues - blue. Units of expression are log2p1 TMM-normalized counts. Titles denote overall Pearson correlation coefficients.

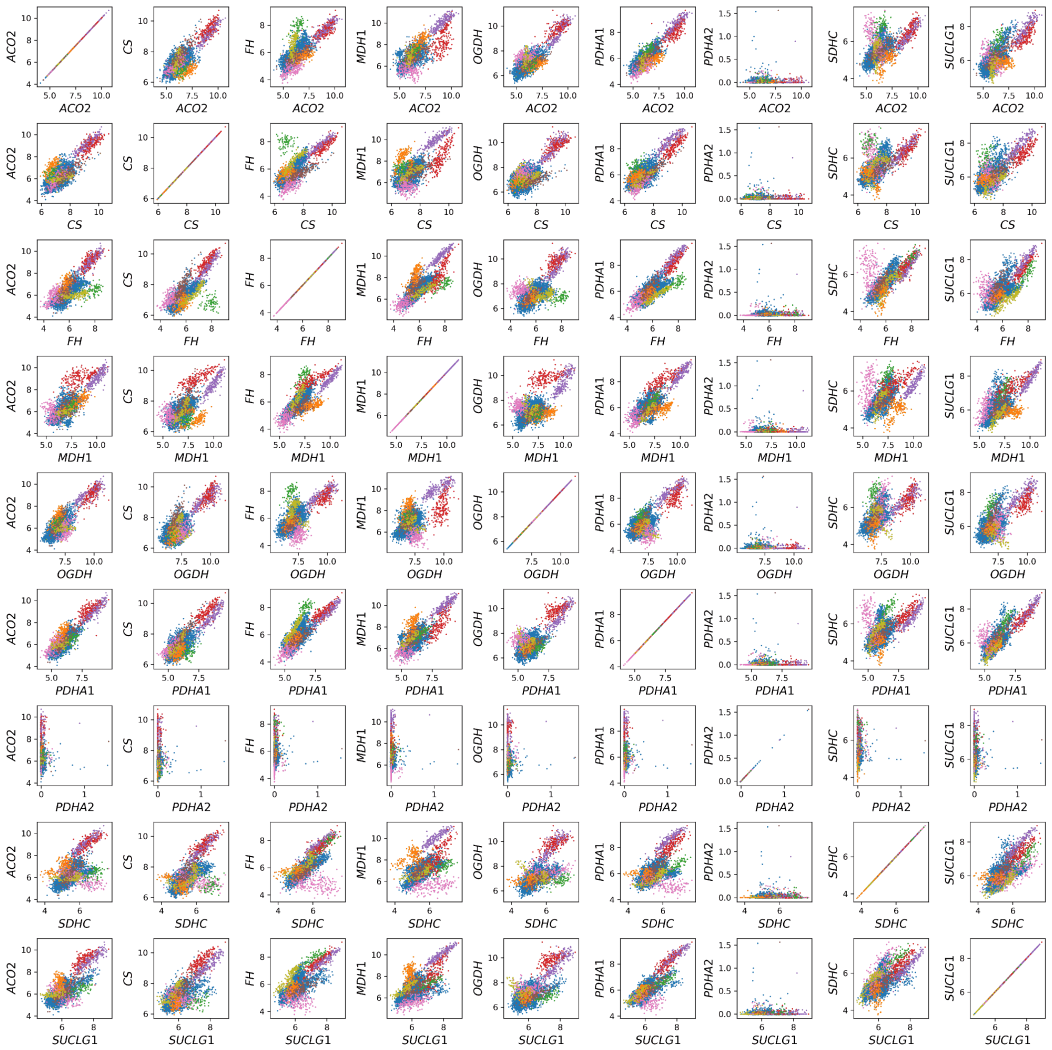


Figure 4.20: **Pairwise scatter plots of TCA cycle genes.** Samples are colored by tissue type, as in Fig 4.2e: skeletal muscle - red, liver - green, brain - orange, heart - purple, whole blood - pink, adipose - brown, fibroblasts - olive green, other tissues - blue. Units of expression are log2p1 TMM-normalized counts. Titles denote overall Pearson correlation coefficients.

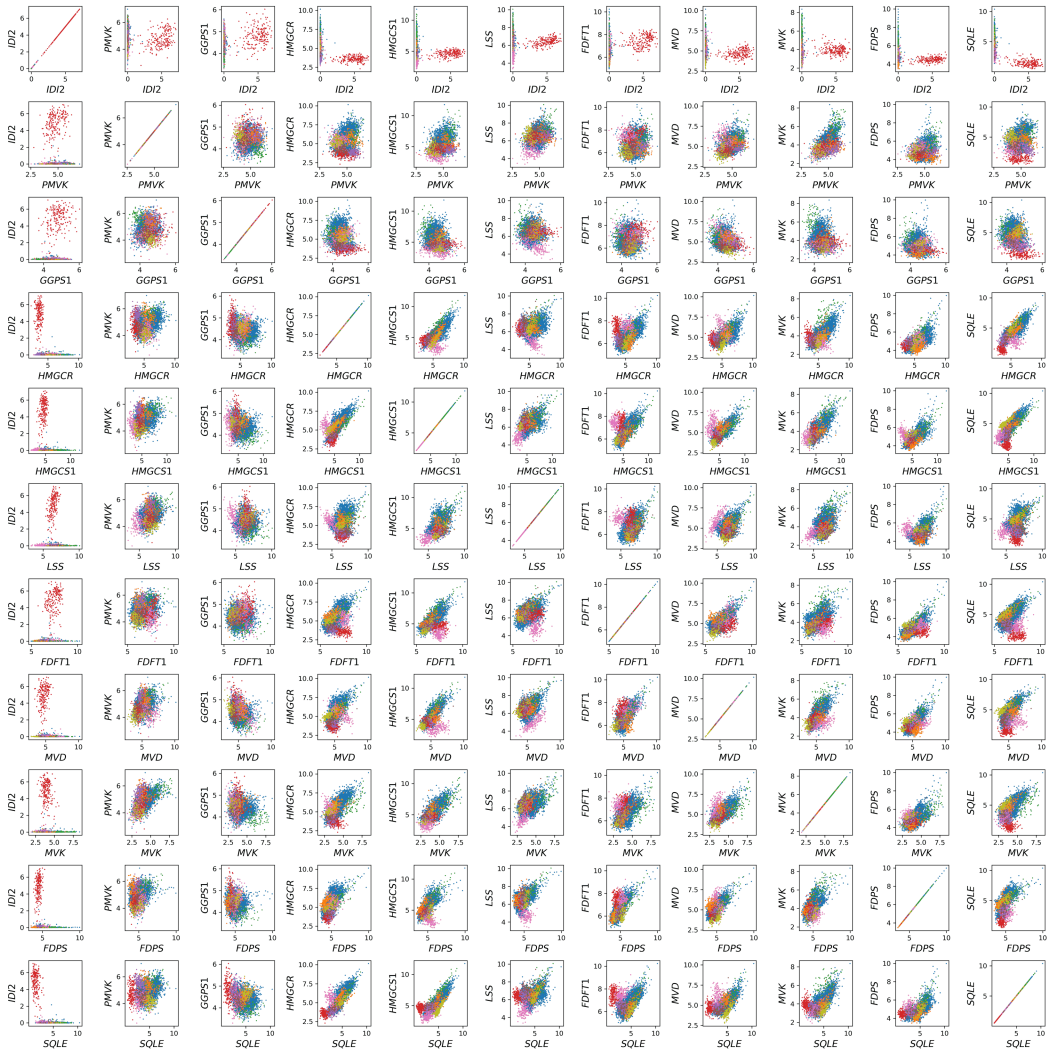


Figure 4.21: **Pairwise scatter plots of cholesterol synthesis genes.** Genes are ordered by mean coefficient in Fig 4.17a, and samples are colored by tissue type, as in Fig 4.2e: skeletal muscle - red, liver - green, brain - orange, heart - purple, whole blood - pink, adipose - brown, fibroblasts - olive green, other tissues - blue. Units of expression are log2p1 TMM-normalized counts. Titles denote overall Pearson correlation coefficients.

Chapter 5

CONCLUSION

At Cold Springs Harbor, I gave a talk that Dana Pe'er introduced as "Back to the basics." The conference was titled "Probabilistic Modeling in Genomics" but was almost entirely about population genetics, save for the systems biology section in which *k*-spaces said, "Hello World." I only discussed the manifold separation problem in epigenomics, which must have seemed somewhat out of place, but I am grateful to the organizers who selected my abstract for a talk, because they must have thought that the basics still had value. I think her introduction was fitting, not just for that talk or for that paper, but for this thesis as a whole.

k-spaces builds upon a foundation of simple, well-studied techniques — the EM algorithm, least squares, *k*-means, PCA, and the Gaussian or normal distribution — to define a model that is interpretable and tractable. When we have nonlinear manifold learning, machine learning embedding techniques, and neural network autoencoders today, *k*-spaces is indeed going back to the basics, but it adds just another layer or two to build up to a model capable of representing complex data. Some problems are currently intractable with classical techniques, but there are others, especially in biology, where there is reason to believe that systems can be explicitly modeled in an interpretable way. Looking ahead, the *k*-spaces framework can let us build the tools to do so.

Chapter Two, my work on enrichment analysis, is also another case of back to the basics. Paul asked me to program enrichment analysis for a new class of pathway models for my rotation, and because I was unaware of the application programmatic interface for getting gene lists from the database and was parsing the data structures myself, I ran into "sets," or groups of genes individually sufficient to perform a pathway step. I asked Paul what he wanted me to do with them, and he said the steps of the pathway matter, not the genes, and that became my first paper. "Sets" have been in the pathway databases for years but were probably never addressed because they were either not around or were uncommon at the time when the enrichment analysis field got started and the first tools were programmed. Compared to other enrichment analysis methods that built upon the original works, it is also very conceptually simple, but that small change is enough to make a big difference.

From there, I was inspired to start the work that became Chapter Four, because I realized in trying to interpret the results that both for the sets and for the remaining genes, the covariation structure was not in our pathway models. As I discuss in the introduction, neither RNA-sequencing nor sufficient scale of data was available when the field was thinking about basic things such as covariance structure within known pathways, but I think it is time to get back after that problem. Neither Gaussian Mixture Models, nor enrichment analysis, nor pathway models are particularly hot areas of research today, but there is great potential in revisiting these concepts with new perspectives. There is no doubt that accurate biophysical models of cells from first principles are the ultimate goal we should strive for if we want to say we truly understand biological processes, but they are a long ways away, which is why we try to bridge the gap with statistical and information-theoretic approaches. If we can produce generative models for gene expression in pathways and integrate publicly available data across sequencing platforms and data collection techniques, we can build up from those basics and work together to answer the question, "What are cells doing?"

As for where to go next, there are many paths to follow, but there is a key question upon which many of these will converge. As described in the conclusion of Chapter Four, better validation techniques will enable faster scaling for pathway studies, and I believe the answer to "how many clusters are there?" lies with selective inference. This question is also dependent, however, on the dimension of the clusters. This is one place where we cannot simply borrow the tools developed by others for k-means or PCA. Automated methods for dimension selection in PCA do not account for the presence of other mixture components, and techniques for selecting the number of clusters in k-means would ignore the fact that mismatched true cluster dimension and mixture component dimension can alter the number of mixture components allocated by the model. More elegant methods than exhaustive model fitting to the whole data would be beneficial, as a technique like selective inference could quantify uncertainty in certain mixture components rather than the whole model fit. Other future applications for k-spaces depend on generalization to other distributions, such as the negative binomial for single-cell data or some count-based distribution to study latent structure in genotype matrices in population genetics. The key to these projects and others to come is in cluster validation, and if I could get an answer to any question, that would be the one.