**Enter the matrix: Interpreting unsupervised feature learning with matrix decomposition to discover hidden knowledge in high-throughput omics data**

## Names/Affiliations

Genevieve L. Stein-O'Brien[1,2,3], Raman Arora[4], Aedin C. Culhane[5,6], Alexander V. Favorov[1,7], Casey S. Greene[8,9], Loyal A. Goff[2,3], Yifeng Li[10], Aloune Ngom[11], Michael F. Ochs[12], Yanun Xu[13], Elana J. Fertig[1]

1 Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins School of Medicine, Baltimore, MD
2 Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD
3 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD
4 Department of Computer Science, Institute for Data Intensive Engineering & Science, Johns Hopkins University, Baltimore, MD
5 Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA
6 Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA
7 Vavilov Institute of General Genetics, Moscow, Russia,
8 Department of Pharmacology, Perelman School of Medicine, University of Pennsylvania, PA
9 Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, PA
10 Digital Technologies Research Centre, National Research Council Canada
11 School of Computer Science, University of Windsor, Ontario, Canada
12 Department of Mathematics and Statistics, The College of New Jersey, Ewing Township, NJ,
13 Department of Applied Mathematics & Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD

**Summary**

High-dimensional data is currently standard for biological inquiry. Biological systems are comprised of interrelated gene regulatory mechanisms, gene-gene interactions, and cellular interactions. These interactions induce low-dimensional structure within the high-dimensional data. Matrix factorization, also known as compressed sensing, learns low-dimensional mathematical representations from high-dimensional data. These factorization techniques can embed assumptions about pleiotropy, epistasis, inter-relationships between complex traits, and context-dependent interactions. They have been applied to uncover new biological knowledge in a breadth of topics ranging from pathway discovery to time course analysis. These techniques have been applied to data from diverse high-throughput omics technologies, including bulk and single-cell data. There are numerous computational techniques within the class of matrix factorization, each of which provides a unique interpretation of the processes in high-dimensional data. We review the visualization and applications of matrix factorization to systems-level analyses, which are diverse and require standardization to enable biological interpretation. Codifying the techniques to decipher biologically relevant features with matrix factorization enables their broad application to discovery beyond the limits of current biological knowledge—answering questions from high-dimensional data that we have not yet thought to ask.

**Keywords:** systems biology, bioinformatics, genomics, unsupervised learning, matrix factorization, dimension reduction, compressed sensing, integrated analyses, visualization, single cell

## Introduction

Many of today's biological experiments involve intensive computation. *In vitro* or *in vivo* experimentation was once the only way to uncover new biological knowledge. High-throughput technologies have ushered in a data deluge era in biology (Bell et al., 2009; Sagoff, 2012) and empowered *in silico* experimentation. Specifically, new insights are often revealed through computation analysis of this data prior to experimental validation. The prominent database of public genomics data in the United States, GEO (Barrett et al., 2013) currently contains data from over 100,000 studies, and over 70,000 studies are stored in the comparable European database ArrayExpress (Kolesnikov et al., 2014). Each study contains tens to hundreds of samples. Single-cell sequencing (scRNA seq) technologies further increase the sample size of comparative bulk datasets by at least two orders of magnitude. 10X Genomics released a scRNA seq dataset of 1.3 million mouse brain cells, and plans are underway for the construction of an entire human cell atlas (Regev et al., 2017). The increasing size of these databases cause today's biological data to be fundamentally high-dimensional. Computational methods to learn the structure of these high-dimensional datasets are essential to leverage the full power of transcriptional, epigenetic, proteomic, and single-cell datasets.

High-throughput biology is poised to characterize complex biological processes (CBPs) resulting from pleiotropy, epistasis, trans-eQTLS, and complex traits. Each CBP involves distinct combinations of numerous molecular components. Multiple CBPs can occur concurrently in a single biological sample. Different CBPs can use the same molecular component. Although complicated, the resulting relationships between CBPs and similarities between samples effectively constrain the structure of high-dimensional datasets. These constraints can be leveraged in computational methods called "matrix factorization" to reduce the high-dimensional data into low-dimensional structures. Such dimension reduction of the biological data highlights perspectives and questions that investigators have not yet considered, and also enables tractable exploration of otherwise massive datasets.

Here, we review the applications of MF algorithms to high-throughput biology and how the results of these applications are interpreted to provide new biological knowledge. There are numerous reviews discussing the mathematical and technical details of MF algorithms (Abdi et al., 2013; Li et al., 2016; Meng et al., 2016c; Ochs and Fertig, 2012) and their applications to microarray data (Devarajan, 2008). In this review, we focus on the subject of biological applications of MF algorithms and the interpretation of their results for *in silico* experimentation for the first time since the advent of sequencing technologies. We describe a variety of MF algorithms applied for high-throughput data analysis and briefly detail their differences and similarities in biological inference. We then describe how the interpretation and visualization of the results of MF analyses are used for pressing biological questions. We focus on examples in pathway analysis, subtype and clonal identification, time course analysis, and single cell data. The examples are selected from limitless applications of MF to biology to highlight a wide range of analysis and visualization techniques to learn new biology from high-throughput data.

### Overview of matrix factorization in genomics

The natural representation of high-dimensional biological data is a matrix of the measured values (expression counts, methylation levels, protein concentrations, etc) with different samples represented in the columns of the matrix and with different molecular features (genes, proteins, etc.) represented in the rows (Fig 1A). If two samples exhibit a similar phenotype, the corresponding columns of the data will often reflect that relationship. CBPs occur simultaneously in multiple samples to different degrees, introducing additional structure to the columns of this matrix. At the same time, coordinated regulation of genes introduces correlation between distinct rows of the data.

Matrix Factorization (MF) is a class of diverse algorithms to reduce data dimensionality through a decomposition approach which learns one matrix that describes the structure between genes and another matrix that learns the biological processes for each sample from the original data matrix (Fig 1A). These two matrices are interrelated, with each set of learned coregulated genes associated with a single biological or technical process. The matrices are learned based upon the assumption that the number of biological processes occurring in the input dataset is smaller than either the number of rows or columns in the input data. This assumption is met today, now that whole-genome measurements of large numbers of samples datasets exhaustively measure biological systems and can, in principle, fully capture CBPs.

Since MF was first applied to microarray data analysis in the early 2000s (Alter et al., 2000; Fellenberg et al., 2001; Moloshok et al., 2002b), the breadth of MF algorithms for high-throughput biology has grown along with their broad applications. Correspondence analysis ((Abdi and Williams, 2010; Fellenberg et al., 2001)independent component analysis (ICA), (Hyvärinen et al., 2004; Kong et al., 2008; Lee and Batzoglou, 2003; Yao et al., 2012), factor analysis (Abdi et al., 2013; Wouters et al., 2003), canonical correlation analysis (CCA) (Lê Cao et al., 2009; Takane et al., 2008; Tenenhaus and Tenenhaus, 2014), Laplacian eigenmaps, and spectral maps (Wouters et al., 2003) are related to the mathematics mainstays of singular value decomposition (SVD) (Eckart and Young, 1936), and principal component analysis (PCA) (Hotelling, 1933; Ma and Dai, 2011; Pearson, 1901; Wall et al., 2003) based analysis. While, Non-negative Matrix Factorization (NMF) (Devarajan, 2008; Ochs and Fertig, 2012), manifold learning (Maindonald, 2009; Moon et al., 2017), and compressed sensing (Cleary et al., 2017; Vattikuti et al., 2014) draw on image processing methods (Lee and Seung, 1999; Ochs et al., 1999). With the rapidly accelerating rate of high-throughput biological data acquisition, these MF techniques enable expanding beyond simple single study designs to power systems-level analyses. Thus, the prescience of early applications of MF is becoming appreciated in the current data deluge era. Yet, the challenges of biological interpretation that were first addressed for microarrays in the review by Devarajan (Devarajan, 2008) remain, and have grown with the complexity of today's datasets (Li et al., 2016; Meng et al., 2016d).

## Moving from mathematical terminology to distinguishing sources of biological and technical variation in high-throughput data

Throughout this review, we first define the relevant terminology and then describe applications to explain the interoperability and comparability of MF approaches in biological systems. Historically, the independent rediscovery of MF in multiple fields created distinct terminologies that are used interchangeably (Abdi and Williams, 2010; Escoufier, 1987; Meng et al., 2016d). In the specific application of MF to genomics, terms including modules, features, eigengenes, metagenes, meta-pathways, and patterns are analytical orthologs. Specific terms represent relationships between molecular measurements (e.g., modules, meta-pathways, or signatures) and other terms represent relationships between samples (e.g., patterns (Moloshok et al., 2002b), metagenes (Brunet et al., 2004), or controlling factors (Kong et al., 2008)). Some terms, such as features or components, are used interchangeably to refer to relationships between molecular measurements or samples depending on the context of the input data and analysis. The word decomposition is often used interchangeably with factorization in referring to the broader class of algorithms. Moving forward, it is essential to have a unified formula-free language to describe the biological meaning of the terminology of MF algorithms.

Fundamentally, the goal of every MF analysis is to disentangle the biological, experimental, and technical factors that distinguish samples or groups of genes. Each factor results in a source of variation shared between multiple samples and/or molecular measurements in the data and is often unknown a priori. Some examples of biological sources of variation are cell types, inter-individual genetic variation, environmental conditions, pharmacological perturbations, and dynamic biological

processes (Hansen et al. 2011; Wagner et al. 2016). Experimental sources of variation arise from tissue age, lack of cell cycle synchronization, and tissue contamination. Technical sources of variation arise from biases in measurement technology, protocol, or sample batches (Leek et al., 2010). We refer to the collection of all of these features as CBPs.

In biological applications, MF analyzes a data matrix that has each sample as a column and each observed value (expression counts, methylation levels, protein concentrations, etc.) as a row. MF seeks to identify the sources of variation within that data matrix by decomposing that data into two related matrices (Fig 1). The rows of the first matrix contain the observations and the columns of the second contain the samples. Here, we adopt the notation of calling the first observation-level matrix the "amplitude" matrix and the second sample-level matrix the "pattern" matrix (Moloshok et al., 2002b). The amplitude matrix represents relationships between observations (typically genes) present in the data and the pattern matrix relationships between samples (typically cell-types/lines, patients, or experimental conditions). The corresponding profiles in the columns of the amplitude matrix or profiles in the rows of the sample matrix are both referred to as "features" or "components". Ideally, the relationships between samples reflect the comparative use of distinct sources of biological, experimental, and technical variation in each sample in the dataset.

The amplitude and the pattern matrices are interrelated. Each column of the amplitude matrix has a corresponding row in the pattern matrix that represent the same CBP. Specifically, the values of a column in the amplitude matrix are called gene weights and represent relative gene expression values when an inferred CBP is used. Similarly, the values in each row of the pattern matrix called sample weights and are proportional to the extent to which that CBP is used in a given sample. The so-called "inner dimension" (columns of the first matrix and rows of the second matrix), corresponds to the number of CBPs learned from the data and is often called the "rank" or "dimensionality" of the input data matrix. Determining the number of features to use in the factorization is critical to its interpretation. The appropriate selection depends upon the algorithm and is an active area of research, with analytical methods for dimensionality first introduced in the context of batch correction to remove technical and experimental sources of variation(Leek, 2011).

MF approaches are also distinguished by the algorithms used to solve for the amplitude and pattern matrices in the decomposition. The computational differences between these algorithms have been described in numerous reviews (Berry et al., 2007; Devarajan, 2008; Ochs and Fertig, 2012; Wang and Zhang, 2013). Perhaps the most prominent of methods is principal component analysis (PCA). This method searches for components among gene and samples that are uncorrelated. The first component explains the maximum variation in the data, the second the maximum variation after that first component is removed, and continuing so that each subsequent factor explains a smaller degree of variation in the data than the previous one. Plots of the weights of genes or samples in each component can be used to learn sources of separation in the data (Roden et al., 2006). Because PCA learns features that explain most of the variation in the data, the algorithm will necessarily combine signal from the multiple sources of biological and technical factors in the data (Ochs and Fertig, 2012). In this regard, PCA may be of use only in experimental paradigms in which the processes or conditions of interest represent the strongest sources of variation by design, and of less use as an exploratory approach in more unstructured experimental designs. This aspect of PCA can be exploited to learn features that explain most of the variance independent of known biological phenotypes to remove technical artifacts for supervised analysis (Leek and Storey, 2007) or to learn features that explain most of the variance independent of the technical artifacts from known batches to retain biological signal for unsupervised analyses (Parker et al., 2014). However, PCA conflates multiple biological processes into single components and, thus, is unable to learn the specific genes co-regulated by a specific CBP. The algorithm selected for analysis must consider both the data types and biological questions asked. The remainder of this review provides examples of alternative MF techniques suited to address critical questions.

## Data-driven gene sets from MF provide context-dependent coregulated gene module and pathway annotations

Scientists often rely on approaches centered around identifying coregulated networks corresponding to functionally related gene modules or pathways (Irizarry et al., 2009; Khatri et al., 2012). This often results in a two-step process where differential expression analysis is used to identify genes, and then interpretation occurs in the context of pathways gleaned from the literature by human curation (Bauer-Mehren et al., 2009; Tsui et al., 2007). A challenge with such pathway-level interpretations is that they often lack important contextual information(Khatri et al., 2012): a gene may be part of a pathway but it may not be expressed in certain cellular contexts. Furthermore, the pathway memberships for many genes are not fully elucidated. Interpretation may also be challenging when a gene plays different roles in two different cellular contexts or lineages. Patterns can be observed specifically associated with the sex of the individual contributing tissue samples (The GTEx Consortium, 2015) or the strain of microbial samples (Tan et al., 2017). As a result, different pathway annotations often have conflicting information about gene membership for the same pathway (Khatri et al., 2012; Liu et al., 2014). Moreover, analyses conditioned on curated gene sets are limited to pathways that have already been observed, limiting scientists ability to discover entirely new patterns.

MF can derive data-driven gene signatures which contain the context of the gene activity in the data matrix. The output of these algorithms are then an amplitude matrix with columns reflecting the correlation structure between genes. Each column can then be used as a pathway reflecting coregulation of the genes contained within (Fig 2, middle). In this way, MF techniques are able to account for gene reuse in multiple pathways. The values of most amplitude matrices are continuous weights describing relative contribution of a gene to a given pathway. However, as marker genes are often used to discretize biological phenomena for experimentation, gene set annotations are often binary (Fig 2, left). Many groups have developed thresholding techniques to select which genes belong to a pathway for binary membership, providing an output similar to clustering techniques (Fertig et al., 2012b; Kim et al., 2017). Other studies also integrate the literature-derived gene signatures in these thresholds to refine the context of pathway databases (Fertig et al., 2012a; Kossenkov et al., 2007). To learn context dependent genes sets, MF can be applied to datasets with well-defined experimental perturbations (Fertig et al., 2012b). Methods based on high-weight genes comprise signatures akin to many curated gene set resources (Tan et al., 2017), whereas genes that are most uniquely associated with a specific pattern may be biomarkers of the cell type or process associated with that pattern (G. L. Stein-O'Brien et al., 2017).

Downstream interpretation can consider these similar to the use of literature-based gene sets. Notably, they can be used as inputs to pathway analyses from differential expression statistics from independent datasets. These pathways are analogous to the hierarchical-clustering based gene modules (Segal et al., 2004) and gene expression signatures from public domain studies in the MSigDB gene set database and analysis tool(Subramanian et al., 2005). Although binary pathway models are substantially easier to interpret, continuous values from the original factorization are more reflective of the data. DeTomaso and Yosef found that weighted gene signatures that take advantage of these continuous values can be more robust to noise and missing values in the data (DeTomaso and Yosef, 2016). If a gene's expression level is poorly measured in a sample due to low sequencing depth, other genes in the same factor can imply the actual expression level of the gene in question. This is because factorization considers each gene in the context of all other genes in the factor, improving the robustness of findings. These continuous signatures can be associated directly with other datasets using projection methods (DeTomaso and Yosef, 2016; Fertig et al., 2013a) or profile correspondence methods (Irizarry et al., 2005).

The properties of the gene modules depend critically on the MF algorithm selected for analysis. PCA is ill-suited to pathway creation due to its mixture of multiple biological processes within a single

factor as described above. Instead of inferring uncorrelated components, Independent Component Analysis (ICA) learns factors that are statistically independent. This assumption of independence ensures that the patterns learned in the data are from distinct processes, thus providing more accurate gene sets than PCA (Engreitz et al., 2010; Lee and Batzoglou, 2003; Teschendorff et al., 2007). Comparison analyses in Rotival et al (Rotival et al., 2011) found that ICA could identify additional modules and modules with higher interpretability with known biological function than the network-based clustering method WGCNA (Weighted Correlation Network Analysis, (Langfelder and Horvath, 2008)).

Yet, due to the evolution of biological systems, genes too have evolved to have numerous functions depending upon the environmental context and which other genes are co-regulated. This coregulation may violate the assumptions of independence in ICA. Therefore, learning gene modules in a manner that can account for a gene's participation in multiple pathways is critical to accurate biological interpretation. Non-negative matrix factorization (NMF) methods instead constrain all elements of the amplitude and pattern matrices to be greater than zero (Lee and Seung, 1999; Ochs et al., 1999). This model has a twofold advantage. First, it models the non-negative nature of transcriptional data. Second, it enables algorithms to learn shared components in multiple components thereby modeling coregulation(Moloshok et al., 2002a). So-called sparsity can be encoded in all of these methods to promote more zero values for the amplitude matrix (Bouwmans and Zahzah, 2014; Fertig et al., 2010; Li et al., 2017; Tenenhaus et al., 2014; Witten et al., 2009). In the extreme case of "binary factorizations", all elements of the amplitude matrix are either zero or one and therefore require no thresholding to define gene modules (Zhang et al., 2010). Encoding less stringent sparsity constraints has been found to further improve pathway inference because it models biological parsimony (Gao and Church, 2005; Kim and Tidor, 2003; Kossenkov and Ochs, 2010, 2009; Ochs and Fertig, 2012).

MF learning of functional gene modules can also be applied to learn functional gene modules on non-transcriptional datasets. For example, Alexandrov et al applied NMF to the mutational landscape of tumors defined from a dataset that counts the number of nucleotides that differ from the reference genome in the context of the preceding and following nucleotides (Alexandrov et al., 2013). This analysis defined mutational signatures associated with distinct cancer driving processes (Alexandrov et al., 2016). Additional extensions have been applied to learn allele combinations in phenotypes (Favorov et al., 2005) and transcript-regulation of genes (Zakeri et al., 2017).

*Section figure: definition of gene signature from a matrix and illustration of continuous gene weights vs binary gene sets in subsequent gene set analyses*

## MF learns relationships between samples that define population stratification, tissue composition, cell types, disease subtypes, and clonality

MF analyses have been employed in large cohorts of samples to define the relationships between individuals or sample groups. Sample-level discovery analyses interpret the relative magnitude of sample weights in the pattern matrix. The output of sample-level discovery with MF is similar for gene signatures in the interpretation of the amplitude matrix. Namely, these values can be binary to perform clustering (Jiang et al., 2004; Yeung et al., 2001) or continuous to define more subtype relationships between samples (Abbas et al., 2009; Erkkilä et al., 2010; Venet et al., 2001). When the learned sample relationships reflect known biology, the corresponding gene signatures can be projected onto new datasets to learn similar sample co-relationships within new datasets. A common example of such an analysis is "computational microdissection" techniques, which often learn the cell type composition of the sample based upon learned gene signatures for each cell type (Hackl et al., 2016).

Just as gene signatures can be learned with a wide variety of MF algorithms, so to can sample-specific relationships. A compelling case for inference of genetic co-relationships with MF occurred by applying PCA to SNP data from 3,000 European individuals. This analysis found that vast majority of biological variation was explained by the longitude and latitude of the country of origin of these individuals (Novembre et al., 2008). Although population of origin is a dominant source of biological variation, common variants associated to disease risk or related traits are also present in the data motivating GWAS studies (McCarthy et al., 2008). These different sources of signal in GWAS data is an example of the principle of "multiple-views" of the data. Namely, there are multiple equally valid sample groupings that reflect the biology of the samples selected for analysis. Such multiple views may occur within a single factorization. For example, application of ICA to molecular subtype discovery in gene expression profiles of 198 bladder cancer (Biton et al., 2014) extracts multiple layers of signal from the data including components that are correlated with batch effects, GC content, biological processes, tumor cells, and cells in the microenvironment. Some components were associated with a subset of cancers. For example a gender component was enriched in lung, kidney and bladder cancers. Bladder and breast cancer both shared a basal-like component. By contrast clustering based approaches were unable to discover overlapping biology between cancers from different anatomical locations (Hoadley et al., 2014) in gene expression profiles of TCGA cancers. The (Biton et al., 2014) study is an elegant demonstration of the power of matrix decomposition to identify multiple "layers", or multiple "sub-groups" in the data. The first set of sample groupings were technical, but by peeling away these, the underlying biology and true biological subtype emerge.

Multiple views may also arise from equally valid views of the data learned from analysis of different subsets of the data or different algorithms. For example, an NMF-class algorithm separated tissue-specific patterns from postmortem samples in the Genotype-Tissue Expression (GTEx) project (Dey et al., 2017). This algorithm found a pattern that combined all samples from brain regions when applied to all tissue samples in GTEx, but separated the distinct brain regions when applied to only tissue samples from the brain. A different sparse NMF algorithm simultaneously separated these brain regions from GTEx with additional patterns that are associated with the individual who donated those samples (G. L. Stein-O'Brien et al., 2017). The dimensionality (number of factors) also impacts the views of the data learned from MF analysis. For example, the same sparse NMF algorithm applied to GTEx data was also applied to high-throughput data for a set of head and neck tumors and controls for a range of dimensionalities (Fertig et al., 2013b). This analysis revealed a hierarchical relationship between features learned with larger numbers of clusters. Notably, two patterns separated tumor and normal samples. When this MF was applied with five patterns, the tumor pattern split into the two dominant clinical subtypes of head and neck cancer. The hierarchical relationship between patterns has been used to assess the robustness of patterns to quantify the optimality of the factorization (Bidaut, 2010) and learn the optimal dimensionality of the factorization (Bidaut et al., 2006). Plotting the relative magnitude of the patterns learned in each sample in the head and neck cancer study demonstrates further relationships between gender and smoking that impact expression to a lesser degree than does tumor status (Fertig et al., 2013b). This observation highlights the ability of continuous features from MF to simultaneously learn the extent to which multiple variables impact biological variation at different degrees.

Learning sample-specific properties in tumors introduces a further degree of complexity. Specifically, tumors have significant intra- and inter-tumor heterogeneity that contributes to the biological variation in their molecular data. Studies often examine the "bulk" tissue without dissecting the tumor from the stroma and infiltrating immune cells, but a single tumor may also contain numerous subclones as the tumor evolves to acquire and accumulate different driver events (e.g., mutations). Previous MF studies assume that a finite number of subclones are contained in each tumor sample. These algorithms constrain the patterns to learn the proportion of the tumor that arises from each subclone by constraining the relative magnitude of the patterns for all the subclones in a single tumor sample to sum to 100% (Deshwar et al., 2015; Lee et al., 2016; Nik-Zainal et al., 2012; Roth

et al., 2014; Xu et al., 2015). Whereas the tissue-type contamination will contain similar cell types in each tumor, the molecular alterations in tumors may be unique to each tumor (Jiang et al., 2016). Assumptions about the evolutionary mechanisms of the accumulation of molecular alterations can be encoded in the factorization to model the resulting heterogeneity of these clones (Xie et al., 2017; Xu et al., 2015). These MF algorithms further utilize recent advances in machine learning to simultaneously learn the composition (pattern matrix) and total number (dimensionality) of the tumor subclones.

## From snapshots to moving pictures: simplifying time course analysis

Entwined in the challenge of decomposing cell types and subpopulations is the fact that all biological systems are highly dynamic. High-throughput, time course datasets are emerging in the literature to account for these dynamics. The central goal of time course analysis is to determine the extent to which specific molecules change over time in response to perturbations (e.g. developmental time, environmental factors, disease processes, or therapeutic treatments). Associating molecular alterations often relies on specialized bioinformatics techniques for time course analysis, such as those reviewed in (Bar-Joseph et al., 2012; Liang and Kelemen, 2017). However, as a MF problem complicated models and corrections for correlations are not required to analysis time course data. This is because the continuous weights for each sample in the pattern matrix can vary across samples collected across distinct time points. These weights can encode the timing of regulatory dynamics directly from the data (Figure 4A).

Both ICA and NMF were found to have signatures characterizing the yeast cell cycle and metabolism in early time-course microarray experiments (Liebermeister, 2002; Moloshok et al., 2002b). Sparse NMF techniques using Bayesian method to obtain the solution further had patterns that reflected the smooth dynamics of these phases (cite early Ochs and Kossenkov). This approach has been shown to simultaneously learn pathway inhibition and transitory response to chemical perturbation of cancer cells (Ochs et al., 2009). These transitory dynamics were associated with an unknown, transitory DNA damage response. Thus, using MF algorithms for time course analysis can learn unknown dynamics from the data without relying on *a priori* knowledge.

MF techniques can also analyze time course data from a wide range of experimental conditions, including those with multiple perturbations or from multiple individuals. In case of HPN-DREAM, MF analysis of phospho-proteomics data determine relative dynamics of pathway perturbations (Hill et al., 2016). The relative strength of inhibition and stimulation experiments was inferred by comparing the magnitude of the rows of the pattern matrix for all experimental conditions at each time point (Figure 4B).

Similar analysis of healthy brain tissues learned the dynamics of transcriptional alterations common to the ageing process from multiple individuals (Fertig et al., 2014). In this case, the pattern matrix is compared across individual donors at each time point to learn both the dynamics and heterogeneity of the molecular processes associated with aging (Figure 4C). The subclone analysis techniques described in the previous section have also been applied to repeat samples to learn the dynamics of cancer development, elucidating the molecular mechanisms that give rise to therapeutic resistance and metastasis. Even if the same number of biological features exist, the rate or timing of related features in different molecular modalities may be offset (G. Stein-O'Brien et al., 2017). These discrepancies by data modality suggest that different regulatory mechanisms may be responsible for initiating and stabilizing the malignant phenotype (G. Stein-O'Brien et al., 2017).

## Integrated analysis of multiple omics data and emerging tensor decomposition methods

Increasingly multi-omics data are generated on a biological system, or tissues in order to fully elucidate the molecular networks that govern the phenotypes. While the concept that no gene, protein, etc works in isolation isn't new, the ability to finally test systems biology hypothesis is only

now being realized. Much of this progress is the result of advances in DNA sequencing, high-throughput epigenetics and mass spectrometry (Bantscheff et al., 2012). MF has the advantage that each dataset is transformed and scaled into a common dimension. This enables the simplest integration of creating a new transformed dataset across multiple omics data types, permitting easy visualization of common trends in a standard MF analysis. For example, Culhane et al. (2003) applied an MF approach that discovers components that maximize covariance, or correlated information among data sets, to multiple gene expression studies of the same cell lines. This approach, coinertia analysis, combined information across platforms to discover which microarray platforms had the most informative set of genes. They later extended this analysis to include more diverse omics data and found that proteomics enables discovery of pathways that not observed in analysis of RNA-seq or microarray gene expression studies (Meng et al., 2014). Fertig et al. (2013) further encoded the repression of gene expression by promoter methylation and applied MF to a combined data matrix containing a set of rows for the expression values and a set of rows for the methylation values. Other algorithms have used joint modeling of features learned separately in different data types to power integrated inference (Meng et al., 2016a; Mo et al., 2013). Multi-feature-set matrix factorization, such as the Bayesian group factor analysis (Li et al., 2016), can also identify common and specific factors among different molecular levels.

Another class of MF modeling techniques that has been used for integrated data analysis is based on a three dimensional representation of the data with each dimension indicating a different data modality. These approaches, termed tensor decomposition, discover the linear relationships between three or more datasets, decomposing datasets into more than two composite matrices, reflecting the shared sources of variation between all pairs of matrices. As a result, these approaches extract informative components that share similarity within and between data sets (recently reviewed by (Meng et al., 2016d; Tenenhaus and Tenenhaus, 2014)). Similar to the simpler MF-based integration techniques, the relative weighting of different datasets is also critical in tensor decomposition. One can weight datasets by size, quality, expected amount of information, or one datasets can act as a reference for the others (Meng et al., 2016d). For example, to extract genotype-phenotype associations, clinical covariates could form a reference matrix to extract gene expression and copy number variants of interest (Carvalho et al., 2008; Nyamundanda et al., 2017). They also enable association of QTLs with transcriptional patterns in large GWAS studies. (Hore et al., 2016; Welsh et al., 2017).

Despite the particular method used for integration, the additional power of systems level analysis has been repeatedly demonstrated by the increased power provided by multiplatform approaches. Gene set analysis performed on integrated molecular profiles, those that integrate over several gene expression platforms (e.g. microarray and RNA-seq) or molecules (e.g. gene and proteins), outperform gene set analysis performed on gene expression data alone (Meng et al., 2016b). However, this style of post hoc integration has also highlighted some of the unique changes facing data integration. As the amount of information contained in different molecular classes varies compression can result in different numbers of features arising in the different modalities (Di Pierro et al., 2016; G. Stein-O'Brien et al., 2017) similar to the discrepancies between the features observed in distinct data modalities for time course analysis.

## MF enables unbiased exploration of single cell data for phenotypes and molecular processes

The rapid evolution and adoption of high-throughput single cell technologies has enabled investigation of the molecular composition, dynamics, regulatory mechanisms, and cellular states at the resolution of the fundamental units of life. The massive scale of these data and specific technological challenges unique to measurements at this resolution are generating new challenges for data analysis, algorithm development, and interpretation of results (Yuan et al., 2017). MF approaches are scalable and therefore are increasingly important in molecular profiling as the capacity of

molecular studies increases. Tens of thousands, or even millions of samples can be succinctly visualized on a plot when the dimensionality of single-cell datasets is reduced with MF analyses.

One common application of single cell expression profiling is the identification of discrete subpopulations or sub-'types' of cells in a complex, heterogeneous population such as a tissue sample. In the application of MF to single cell gene expression estimates, one can simultaneously learn patterns corresponding to subtype-specific gene expression signatures, and the relative weight of these patterns on individual cells will be indicative of cluster/subtype identity. Thus, MF approaches are a natural choice in scRNA-seq data analysis and are often additionally used for reducing dimensionality, identifying and removing batch effects, cell type annotation, and summarization of data (Fan et al., 2016-3; Moon et al., 2017; Trapnell et al., 2014; William Townes et al., 2017). In linear NMF, the factors are additive. Namely, in these approaches many biological pathways can be combined in a linear fashion to form a molecular phenotype. A further class of non-linear factorization techniques has been developed to enhance visualization of trajectory structures in scRNA-seq or single cell proteomic data (Maaten and Hinton, 2008; Pierson and Yau, 2015; Risso et al., 2017; Trapnell et al., 2014); however, most of these methods still rely on a linear form of MF, commonly PCA, for initialization. A further class of MF approaches applied to scRNA-seq summarize factors using gene sets and these include factorial single-cell latent variable model (f-scLVM) (Buettner et al., 2016), and over-dispersion gene sets analysis of scRNA-seq (PAGODA)(Fan et al., 2016-3).

## Discussion

Matrix factorization algorithms are versatile techniques with broad applications to unsupervised clustering, biological pattern discovery, component identification, and prediction (See summary table). They can integrate diverse data types and account for technical sources of variation in data. The resulting features in MF techniques directly uncover the nature of the data. As each algorithm aims to identify certain types of features, in some cases different algorithms will learn distinct features from the same dataset. Therefore, investigators may benefit from applying multiple techniques with different properties, or by carefully considering the dataset and question to select exactly the right technique for that question. The features that MF techniques can extract are constrained by the dataset used to train them. These algorithms cannot learn unmeasured features nor can they correct for complete overlap between technical artifacts and biological conditions. Thus, being mindful of experimental design when selecting data sets, and choosing those that are broad enough to cover the relevant sources of variability, is essential to powering systems level analyses with MF.

The origins of MF pre-date modern genomics (Eckart and Young, 1936; Hotelling, 1933; Pearson, 1901), and even the discovery of the structure of DNA, however these methods are ideal to find meaningful subtypes and features in large scale genomics data (Carvalho et al., 2008; De la Cruz and Holmes, 2011; Lee and Seung, 1999; Meng et al., 2016d; Ochs et al., 1999). The field of matrix decomposition is vast. Numerous extensions build upon this basic framework. Each of these have slightly different variations on the mathematical emphases. The selection of a matrix factorization algorithm depends on the nature of the specific biological question and the data modalities. The description of a variety of algorithms is well-reviewed in other studies (Berry et al., 2007; Li et al., 2016; Meng et al., 2016d; Ochs and Fertig, 2012; Wang and Zhang, 2013; Zhou et al., 2014) and is beyond the scope of this article. New analytic algorithms for MF to analyze emerging sequencing technologies and theoretical advances to big data analysis are an active area of research in both bioinformatics and traditional data science disciplines.

Because MF algorithms learn structures directly from unlabeled data, they are in the class of unsupervised learning algorithms. These tools are often most suitable to solve the major challenges of bioinformatics (Greene, 2016). However, there are three areas where we expect to see substantial technical improvement. (1) Currently most models encode a linear relationship between observed data and hidden variables, while more complex interplays might only be captured by deeper models. (2) Single-cell sequencing data are approaching the size of all bulk datasets to date. It will soon be

commonplace to have a single-view or multi-view single-cell data set with millions of samples. Matrix decomposition methods require the computation and manipulations of similarity matrix and fast computation approaches are required (Abraham et al., 2017) (3) A specific view in multi-omics data follows a unique distribution and only tells a complementary piece of a story. Thus, how to optimize the care of these information will be crucial to an integrative analysis. In this new chapter of artificial intelligence, the generalizations of matrix factorizations to deep learning will conquer these technical challenges.

Future biological studies will be more often coordinated in large consortia such as the TCGA, RoadMap, GTEx, FANTOM, and Human Cell Atlas projects. Communication and understanding between multidisciplinary scientists are essential to select the appropriate analytics algorithm (Ronan et al., 2016). This requires new strategies for collaboration and better cross-disciplinary training in both fields. This review provides details of the interpretation of unsupervised analysis with MF to better understand why one algorithm would be selected over another. Better techniques for visualization and interpretation of computational results will be post hoc analyses challenges. Altogether, we will be able to develop and interpret optimal unsupervised learning to solve the data deluge and direct new areas of experimentation and of precision medicine.

## Acknowledgements

We would like to thank J Brian Byrd, Irene Gallego Romero, Lillian Fritz-Laylin, Luciane Kagohara, Louise Klein, Daniela Witten, and members of New PI Slack for their insightful feedback.

This work was supported by National Institutes of Health [grant numbers NCI 2P30CA006516-52 and 2P50CA101942-11 to A.C.C., NCI R01CA177669 E.J.F., NLM R01LM011000 to M.F.O. and NCI P30 CA006973], the Johns Hopkins University Catalyst and Discovery Awards to E.J.F., the Johns Hopkins University IDIES Award to E.J.F. and RA, the Johns Hopkins School of Medicine Synergy award to E.J.F. and L.G., a grant from The Gordon and Betty Moore Foundation (GBMF 4552) to CSG, Alex's Lemonade Stand Foundation's Childhood Cancer Data Lab (CSG), the Department of Defense BCRP [award number BC140682P1 (ACC)], the National Science and Engineering Council of Canada [NSERC DG grant number RGPIN-2016-05017 (AN)], the Windsor-Essex County Cancer Centre Foundation [Seeds4Hope grant number 814221 (AN)], Hopkins inHealth and Booz Allen Hamilton (90056858) to Y.X., the Russian Foundation for Basic Research KOMFI 17-00-00208 an NIH: NCI P30 CA006973 to A.V.F and National Research Council Canada to Y.F.. Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense.

## References

Abbas, A.R., Wolslegel, K., Seshasayee, D., Modrusan, Z., Clark, H.F., 2009. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4, e6098.

Abdi, H., Williams, L.J., 2010. Principal component analysis. WIREs Comp Stat 2, 433–459.

Abdi, H., Williams, L.J., Valentin, D., 2013. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Comp Stat 5, 149–179.

Abraham, G., Qiu, Y., Inouye, M., 2017. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics 33, 2776–2778.

Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., Campbell, P.J., Vineis, P., Phillips, D.H., Stratton, M.R., 2016. Mutational signatures associated with tobacco

smoking in human cancer. Science 354, 618–622.
Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, Å., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjörd, J.E., Foekens, J.A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinsk, M., Jäger, N., Jones, D.T.W., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., López-Otín, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P.A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J.V., Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N.J., Valdés-Mas, R., van Buuren, M.M., van t Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Initiative, Australian Pancreatic Cancer Genome, Consortium, ICGC Breast Cancer, Consortium, ICGC MMML-Seq, PedBrain, I., Zucman-Rossi, J., Futreal, P.A., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R., 2013. Signatures of mutational processes in human cancer. Nature 500, 415–421.
Alter, O., Brown, P.O., Botstein, D., 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. U. S. A. 97, 10101–10106.
Bantscheff, M., Lemeer, S., Savitski, M.M., Kuster, B., 2012. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. Analytical and bioanalytical … 404, 939–965.
Bar-Joseph, Z., Gitter, A., Simon, I., 2012. Studying and modelling dynamic biological processes using time-series gene expression data. Nat. Rev. Genet. 13, 552–564.
Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C.L., Serova, N., Davis, S., Soboleva, A., 2013. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 41, D991–5.
Bauer-Mehren, A., Furlong, L.I., Sanz, F., 2009. Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol. Syst. Biol. 5, 290.
Bell, G., Hey, T., Szalay, A., 2009. Beyond the data deluge. Science.
Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J., 2007. Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. 52, 155–173.
Bidaut, G., 2010. Interpreting and Comparing Clustering Experiments Through Graph Visualization and Ontology Statistical Enrichment with the ClutrFree Package, in: Link.springer.com. Springer US, Boston, MA, pp. 315–333.
Bidaut, G., Suhre, K., Claverie, J.-M., Ochs, M.F., 2006. Determination of strongly overlapping signaling activity from microarray data. BMC Bioinformatics 7, 99.
Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A., Radvanyi, F., 2014. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 9, 1235–1245.

Bouwmans, T., Zahzah, E.H., 2014. Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance. Comput. Vis. Image Underst. 122, 22–34.
Brunet, J.P., Brunet, J.-P., Tamayo, P., Tamayo, P., Golub, T.R., Golub, T.R., Mesirov, J.P., Mesirov, J.P., 2004. Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A. 101, 4164–4169.
Buettner, F., Pratanwanich, N., Marioni, J.C., Stegle, O., 2016. Scalable latent-factor models applied to single-cell RNA-seq data separate biological drivers from confounding effects. bioRxiv. doi:10.1101/087775
Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., West, M., 2008. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. J. Am. Stat. Assoc. 103, 1438–1456.
Cleary, B., Cong, L., Lander, E., Regev, A., 2017. Composite measurements and molecular compressed sensing for highly efficient transcriptomics. bioRxiv.
De la Cruz, O., Holmes, S., 2011. THE DUALITY DIAGRAM IN DATA ANALYSIS: EXAMPLES OF MODERN APPLICATIONS. Ann. Appl. Stat. 5, 2266–2277.
Deshwar, A.G., Vembu, S., Yung, C.K., Jang, G.H., Stein, L., Morris, Q., 2015. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 16, 35.
DeTomaso, D., Yosef, N., 2016. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics 17, 315.
Devarajan, K., 2008. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Comput. Biol. 4, e1000029.
Dey, K.K., Hsiao, C.J., Stephens, M., 2017. Visualizing the structure of RNA-seq expression data using grade of membership models. PLoS Genet. 13, e1006599.
Di Pierro, M., Zhang, B., Aiden, E.L., Wolynes, P.G., Onuchic, J.N., 2016. Transferable model for chromosome architecture. Proc. Natl. Acad. Sci. U. S. A. 113, 12168–12173.
Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218.
Engreitz, J.M., Daigle, B.J., Jr., Marshall, J.J., Altman, R.B., 2010. Independent component analysis: Mining microarray data for fundamental human gene expression modules. Journal of biomedical … 43, 932–944.
Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P., Visakorpi, T., Shmulevich, I., Lähdesmäki, H., 2010. Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics 26, 2571–2577.
Escoufier, Y., 1987. The Duality Diagram: A Means for Better Practical Applications, in: Legendre, P., Legendre, L. (Eds.), Develoments in Numerical Ecology, NATO ASI Series. Springer Berlin Heidelberg, pp. 139–156.
Fan, J., Salathia, N., Liu, R., Kaeser, G.E., Yung, Y.C., Herman, J.L., Kaper, F., Fan, J.-B., Zhang, K., Chun, J., Kharchenko, P.V., 2016-3. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods 13, 241–244.
Favorov, A.V., Andreewski, T.V., Sudomoina, M.A., Favorova, O.O., Parmigiani, G., Ochs, M.F., 2005. A Markov chain Monte Carlo technique for identification of

combinations of allelic variants underlying complex diseases in humans. Genetics 171, 2113–2121.
Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., Vingron, M., 2001. Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. U. S. A. 98, 10781–10786.
Fertig, E.J., Ding, J., Favorov, A.V., Parmigiani, G., Ochs, M.F., 2010. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics 26, 2792–2793.
Fertig, E.J., Favorov, A.V., Ochs, M.F., 2013a. Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobioscience 12, 142–149.
Fertig, E.J., Favorov, A.V., Ochs, M.F., 2012a. Identifying context-specific transcription factor targets from prior knowledge and gene expression data, in: 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 1–6.
Fertig, E.J., Markovic, A., Danilova, L.V., Gaykalova, D.A., Cope, L., Chung, C.H., Ochs, M.F., Califano, J.A., 2013b. Preferential Activation of the Hedgehog Pathway by Epigenetic Modulations in HPV Negative HNSCC Identified with Meta-Pathway Analysis. PLoS One 8, e78127.
Fertig, E.J., Ren, Q., Cheng, H., Hatakeyama, H., Dicker, A.P., Rodeck, U., Considine, M., Ochs, M.F., Chung, C.H., 2012b. Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. BMC Genomics 13, 160.
Fertig, E.J., Stein-O'Brien, G., Jaffe, A., Colantuoni, C., 2014. Pattern Identification in Time-Course Gene Expression Data with the CoGAPS Matrix Factorization. Methods Mol. Biol. 1101, 87–112.
Gao, Y., Church, G., 2005. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics 21, 3970–3975.
Greene, C.S., 2016. The future is unsupervised. Sci. Transl. Med. 8, 346ec108–346ec108.
Hackl, H., Charoentong, P., Finotello, F., Trajanoski, Z., 2016. Computational genomics tools for dissecting tumour-immune cell interactions. Nat. Rev. Genet. 17, 441–458.
Hill, S.M., Heiser, L.M., Cokelaer, T., Unger, M., Nesser, N.K., Carlin, D.E., Zhang, Y., Sokolov, A., Paull, E.O., Wong, C.K., Others, 2016. Consortium HPN-DREAM. Mills GB, Gray JW, Kellen M, Norman T, Friend S, Qutub AA, Fertig EJ, Guan Y, Song M, Stuart JM, Spellman PT, Koeppl H, Stolovitzky G, Saez-Rodriguez J, Mukherjee S. Inferring causal molecular networks: empirical assessment through a community-based effort. Nat Methods 13, 310–318.
Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A.A., Van't Veer, L.J., Lopez-Bigas, N., Laird, P.W., Raphael, B.J., Ding, L., Robertson, A.G., Byers, L.A., Mills, G.B., Weinstein, J.N., Van Waes, C., Chen, Z., Collisson, E.A., Cancer Genome Atlas Research Network, Benz, C.C., Perou, C.M., Stuart, J.M., 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158, 929–944.

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M.I., Small, K., Marchini, J., 2016. Tensor decomposition for multiple-tissue gene expression experiments. Nat. Genet. 48, 1094–1100.
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.
Hyvärinen, A., Karhunen, J., Oja, E., 2004. Independent Component Analysis. John Wiley & Sons.
Irizarry, R.A., Chi Wang, Yun Zhou, Speed, T.P., 2009. Gene set enrichment analysis made simple. Stat. Methods Med. Res. 18, 565–575.
Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G.N., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S.C., Hoffman, E., Jedlicka, A.E., Kawasaki, E., Martínez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S.Q., Yu, W., 2005. Multiple-laboratory comparison of microarray platforms. Nat. Methods 2, 345–350.
Jiang, D., Tang, C., Zhang, A., 2004. Cluster analysis for gene expression data: a survey. IEEE Trans. Knowl. Data Eng. 16, 1370–1386.
Jiang, Y., Qiu, Y., Minn, A.J., Zhang, N.R., 2016. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc. Natl. Acad. Sci. U. S. A. 113, E5528–37.
Khatri, P., Sirota, M., Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol. 8, e1002375.
Kim, J.W., Abudayyeh, O.O., Yeerna, H., Yeang, C.-H., Stewart, M., Jenkins, R.W., Kitajima, S., Konieczkowski, D.J., Medetgul-Ernar, K., Cavazos, T., Mah, C., Ting, S., Van Allen, E.M., Cohen, O., Mcdermott, J., Damato, E., Aguirre, A.J., Liang, J., Liberzon, A., Alexe, G., Doench, J., Ghandi, M., Vazquez, F., Weir, B.A., Tsherniak, A., Subramanian, A., Meneses-Cime, K., Park, J., Clemons, P., Garraway, L.A., Thomas, D., Boehm, J.S., Barbie, D.A., Hahn, W.C., Mesirov, J.P., Tamayo, P., 2017. Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. Cell Syst 5, 105–118.e9.
Kim, P.M., Tidor, B., 2003. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 13, 1706–1718.
Kolesnikov, N., Hastings, E., Keays, M., 2014. ArrayExpress update—simplifying data submissions. Nucleic acids ….
Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., Huang, X., 2008. A review of independent component analysis application to microarray gene expression data. Biotechniques 45, 501–520.
Kossenkov, A.V., Ochs, M.F., 2010. Matrix factorisation methods applied in microarray data analysis. Int. J. Data Min. Bioinform. 4, 72–90.
Kossenkov, A.V., Ochs, M.F., 2009. Chapter 3 Matrix Factorization for Recovery of Biological Processes from Microarray Data, in: (null). Elsevier, pp. 59–77.
Kossenkov, A.V., Peterson, A.J., Ochs, M.F., 2007. Determining transcription factor activity from microarray data using Bayesian Markov chain Monte Carlo sampling. Stud. Health Technol. Inform. 129, 1250–1254.
Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation

network analysis. BMC Bioinformatics 9, 559.
Lê Cao, K.-A., Martin, P.G.P., Robert-Granié, C., Besse, P., 2009. Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics 10, 34.
Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature.
Lee, J., Müller, P., Sengupta, S., Gulukota, K., Ji, Y., 2016. Bayesian inference for intratumour heterogeneity in mutations and copy number variation. J. R. Stat. Soc. Ser. C Appl. Stat. 65, 547–563.
Leek, J.T., 2011. Asymptotic conditional singular value decomposition for high-dimensional genomic data. Biometrics 67, 344–352.
Leek, J.T., Storey, J.D., 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet. 3, e161.
Lee, S.-I., Batzoglou, S., 2003. Application of independent component analysis to microarrays. Genome Biol. 4, R76.
Liang, Y., Kelemen, A., 2017. Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications. Brief. Bioinform.
Liebermeister, W., 2002. Linear modes of gene expression determined by independent component analysis. Bioinformatics.
Liu, H., Beck, T.N., Golemis, E.A., Serebriiskii, I.G., 2014. Integrating in silico resources to map a signaling network. Methods Mol. Biol. 1101, 197–245.
Li, Y., Wu, F.-X., Ngom, A., 2016. A review on machine learning principles for multi-view biological data integration. Brief. Bioinform. bbw113.
Li, Z., Safo, S.E., Long, Q., 2017. Incorporating biological information in sparse principal component analysis with application to genomic data. BMC Bioinformatics 18, 332.
Maaten, L. van der, Hinton, G., 2008. Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.
Maindonald, J., 2009. Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning. Journal of Statistical Software, Book Reviews 29, 1–3.
Ma, S., Dai, Y., 2011. Principal component analysis based methods in bioinformatics studies. Brief. Bioinform. 12, 714–722.
McCarthy, M.I., Abecasis, G.R., Cardon, L.R., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev.
Meng, C., Helm, D., Frejno, M., Kuster, B., 2016a. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. J. Proteome Res. 15, 755–765.
Meng, C., Kuster, B., Culhane, A.C., Gholami, A.M., 2014. A multivariate approach to the integration of multi-omics datasets. BMC Bioinformatics 15, 162.
Meng, C., Kuster, B., Peters, B., Culhane, A.C., Gholami, A.M., 2016b. moGSA: integrative single sample gene-set analysis of multiple omics data. bioRxiv 046904.
Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M., Culhane, A.C., 2016c. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 17, 628–641.
Meng, C., Zeleznik, O.A., Thallinger, G.G., Kuster, B., Gholami, A.M., Culhane, A.C.,

2016d. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform. 17, 628–641.
Moloshok, T.D., Klevecz, R.R., Grant, J.D., Manion, F.J., Speier, W.F., 4th, Ochs, M.F., 2002a. Application of Bayesian decomposition for analysing microarray data. Bioinformatics 18, 566–575.
Moloshok, T.D., Klevecz, R.R., Grant, J.D., Manion, F.J., Speier, W.F., Ochs, M.F., 2002b. Application of Bayesian decomposition for analysing microarray data. Bioinformatics 18, 566–575.
Moon, K.R., van Dijk, D., Wang, Z., Chen, W., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G., Krishnaswamy, S., 2017. PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data. bioRxiv 120378.
Mo, Q., Wang, S., Seshan, V.E., Al, E., 2013. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proceedings of the ….
Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S.L., Hinton, J., Menzies, A., Stebbings, L.A., Leroy, C., Jia, M., Rance, R., Mudie, L.J., Gamble, S.J., Stephens, P.J., McLaren, S., Tarpey, P.S., Papaemmanuil, E., Davies, H.R., Varela, I., McBride, D.J., Bignell, G.R., Leung, K., Butler, A.P., Teague, J.W., Martin, S., Jönsson, G., Mariani, O., Boyault, S., Miron, P., Fatima, A., Langerød, A., Aparicio, S.A.J.R., Tutt, A., Sieuwerts, A.M., Borg, Å., Thomas, G., Salomon, A.V., Richardson, A.L., Børresen-Dale, A.-L., Futreal, P.A., Stratton, M.R., Campbell, P.J., 2012. The Life History of 21 Breast Cancers. Cell 149, 994–1007.
Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M., Bustamante, C.D., 2008. Genes mirror geography within Europe. Nature 456, 98–101.
Nyamundanda, G., Eason, K., Sadanandam, A., 2017. A Next Generation Clustering Tool Enables Identification of Functional Cancer Subtypes with Associated Biological Phenotypes. bioRxiv. doi:10.1101/175307
Ochs, M.F., Fertig, E.J., 2012. Matrix Factorization for Transcriptional Regulatory Network Inference. … Bioinformatics and Computational Biology … 1–10.
Ochs, M.F., Rink, L., Tarn, C., Mburu, S., Taguchi, T., Eisenberg, B., Godwin, A.K., 2009. Detection of Treatment-Induced Changes in Signaling Pathways in Gastrointestinal Stromal Tumors Using Transcriptomic Data. Cancer Res. 69, 9125–9132.
Ochs, M.F., Stoyanova, R.S., Arias-Mendoza, F., Brown, T.R., 1999. A New Method for Spectral Decomposition Using a Bilinear Bayesian Approach. J. Magn. Reson. 137, 161–176.
Parker, H.S., Leek, J.T., Favorov, A.V., Considine, M., Xia, X., Chavan, S., Chung, C.H., Fertig, E.J., 2014. Preserving biological heterogeneity with a permuted surrogate variable analysis for genomics batch correction. Bioinformatics 30, 2757–2763.
Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philos. Mag. 2, 559–572.

Pierson, E., Yau, C., 2015. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 16, 241.

Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Gottgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S.K., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundeberg, J., Majumder, P., Marioni, J., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Philipakis, A., Ponting, C.P., Quake, S.R., Reik, W., Rozenblatt-Rosen, O., Sanes, J.R., Satija, R., Shumacher, T., Shalek, A.K., Shapiro, E., Sharma, P., Shin, J., Stegle, O., Stratton, M., Stubbington, M.J.T., van Oudenaarden, A., Wagner, A., Watt, F.M., Weissman, J.S., Wold, B., Xavier, R.J., Yosef, N., Human Cell Atlas, 2017. The Human Cell Atlas. bioRxiv. doi:10.1101/121202

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., Vert, J.P., 2017. ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. bioRxiv.

Roden, J.C., King, B.W., Trout, D., Mortazavi, A., Wold, B.J., Hart, C.E., 2006. Mining gene expression data by interpreting principal components. BMC Bioinformatics 7, 194.

Ronan, T., Qi, Z., Naegle, K.M., 2016. Avoiding common pitfalls when clustering biological data. Sci. Signal. 9, re6–re6.

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., Shah, S.P., 2014. PyClone: statistical inference of clonal population structure in cancer. Nat. Methods 11, 396–398.

Rotival, M., Zeller, T., Wild, P.S., Maouche, S., Szymczak, S., Schillert, A., Castagné, R., Deiseroth, A., Proust, C., Brocheton, J., Godefroy, T., Perret, C., Germain, M., Eleftheriadis, M., Sinning, C.R., Schnabel, R.B., Lubos, E., Lackner, K.J., Rossmann, H., Münzel, T., Rendon, A., Cardiogenics Consortium, Erdmann, J., Deloukas, P., Hengstenberg, C., Diemert, P., Montalescot, G., Ouwehand, W.H., Samani, N.J., Schunkert, H., Tregouet, D.-A., Ziegler, A., Goodall, A.H., Cambien, F., Tiret, L., Blankenberg, S., 2011. Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Trans-Regulated Gene Modules in Humans. PLoS Genet. 7, e1002367.

Sagoff, M., 2012. Data deluge and the human microbiome project. Issues Sci. Technol.

Segal, E., Friedman, N., Koller, D., Regev, A., 2004. A module map showing conditional activity of expression modules in cancer. Nat. Genet.

Stein-O'Brien, G., Kagohara, L.T., Li, S., Thakar, M., Ranaweera, R., Ozawa, H., Cheng, H., Considine, M., Favorov, A., Danilova, L., Califano, J.A., Izumchenko, E., Gaykalova, D.A., Chung, C.H., Fertig, E.J., 2017. Integrated time-course omics analysis distinguishes immediate therapeutic response from acquired resistance. bioRxiv. doi:10.1101/136564

Stein-O'Brien, G.L., Carey, J.L., Lee, W.-S., Considine, M., Favorov, A.V., Flam, E., Guo, T., Li, S., Marchionni, L., Sherman, T., Sivy, S., Gaykalova, D.A., McKay, R.D., Ochs, M.F., Colantuoni, C., Fertig, E.J., 2017. PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. Bioinformatics.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545–15550.

Takane, Y., Hwang, H., Abdi, H., 2008. Regularized Multiple-Set Canonical Correlation Analysis. Psychometrika 73, 753–775.

Tan, J., Doing, G., Lewis, K.A., Price, C.E., Chen, K.M., Cady, K.C., Perchuk, B., Laub, M.T., Hogan, D.A., Greene, C.S., 2017. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. Cell systems.

Tenenhaus, A., Philippe, C., Guillemot, V., Le Cao, K.-A., Grill, J., Frouin, V., 2014. Variable selection for generalized canonical correlation analysis. Biostatistics 15, 569–583.

Tenenhaus, A., Tenenhaus, M., 2014. Regularized generalized canonical correlation analysis for multiblock or multigroup data analysis. Eur. J. Oper. Res. 238, 391–403.

Teschendorff, A.E., Journée, M., Absil, P.A., Sepulchre, R., Caldas, C., 2007. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput. Biol. 3, e161.

The GTEx Consortium, 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348, 648–660.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., Rinn, J.L., 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386.

Tsui, I.F.L., Chari, R., Buys, T.P.H., Lam, W.L., 2007. Public databases and software for the pathway analysis of cancer genomes. Cancer Inform. 3, 379–397.

Vattikuti, S., Lee, J.J., Chang, C.C., Hsu, S.D.H., Chow, C.C., 2014. Applying compressed sensing to genome-wide association studies. Gigascience 3, 10.

Venet, D., Pecasse, F., Maenhaut, C., Bersini, H., 2001. Separation of samples into their constituents using gene expression data. Bioinformatics.

Wall, M.E., Rechtsteiner, A., Rocha, L.M., 2003. Singular Value Decomposition and Principal Component Analysis, in: Berrar, D.P., Dubitzky, W., Granzow, M. (Eds.), A Practical Approach to Microarray Data Analysis. Springer US, pp. 91–109.

Wang, Y.-X., Zhang, Y.-J., 2013. Nonnegative Matrix Factorization: A Comprehensive Review. IEEE Trans. Knowl. Data Eng. 25, 1336–1353.

Welsh, S., McVean, G., Leslie, S., Donnelly, P., Marchini, J., 2017. Genome-wide genetic data on~ 500,000 UK Biobank participants. bioRxiv.

William Townes, F., Hicks, S.C., Aryee, M.J., Irizarry, R.A., 2017. Varying-Censoring Aware Matrix Factorization for Single Cell RNA-Sequencing. biorxiv. doi:10.1101/166736

Witten, D.M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.

Wouters, L., Göhlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G., Lewi, P.J., 2003. Graphical exploration of gene expression data: a comparative study of three multivariate methods. Biometrics 59, 1131–1139.

Xie, F., Zhou, M., Xu, Y., 2017. BayCount: A Bayesian Decomposition Method for Inferring Tumor Heterogeneity using RNA-Seq Counts.

Xu, Y., Müller, P., Yuan, Y., Gulukota, K., Ji, Y., 2015. MAD Bayes for Tumor Heterogeneity—Feature Allocation With Exponential Family Sampling. J. Am. Stat. Assoc. 110, 503–514.

Yao, F., Coquery, J., Lê Cao, K.-A., 2012. Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. BMC Bioinformatics 13, 24.

Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L., 2001. Model-based clustering and data transformations for gene expression data. Bioinformatics 17, 977–987.

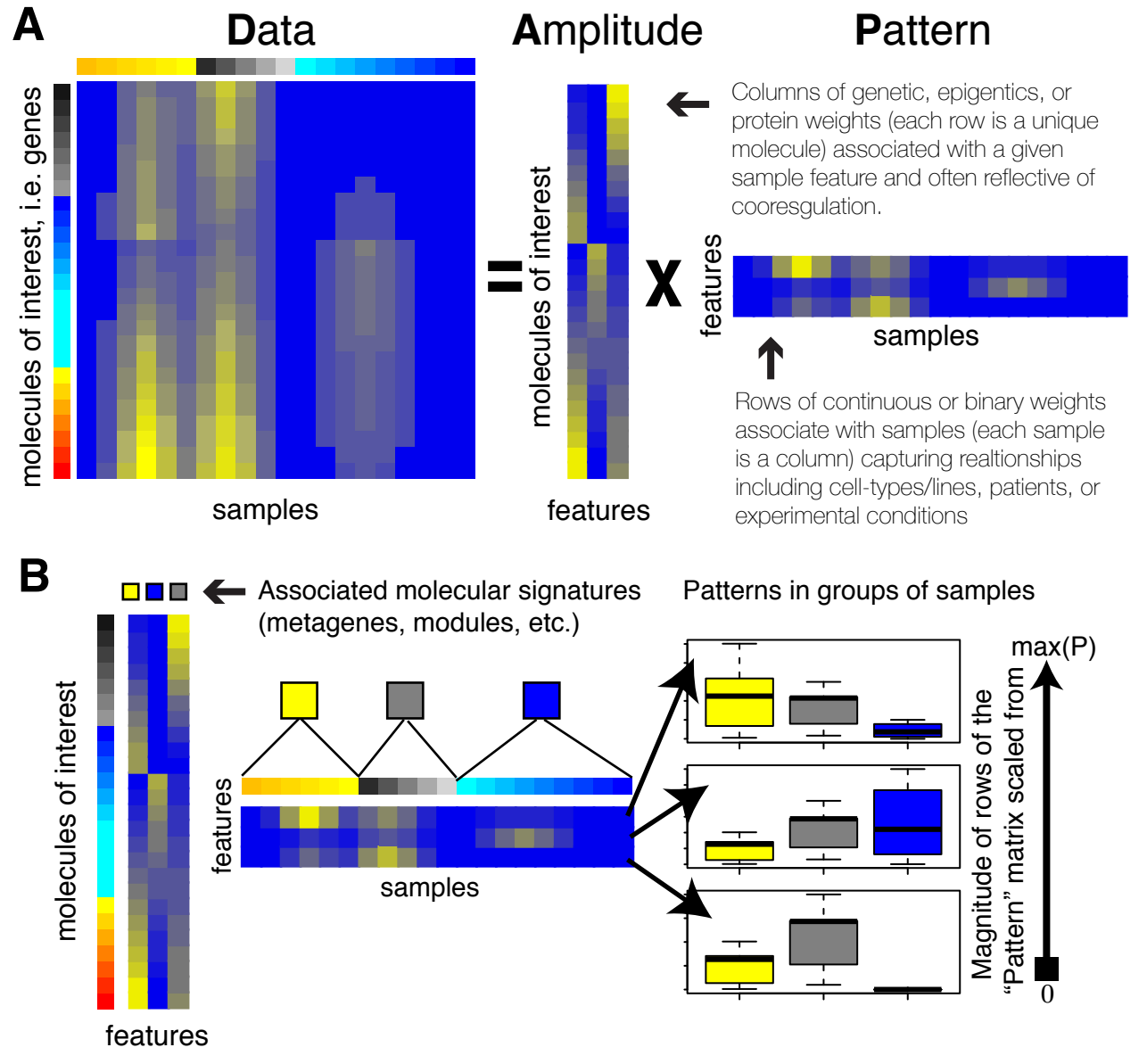
Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., Quackenbush, J., Saadatpour, A., Schroeder, T., Shivdasani, R., Tirosh, I., 2017. Challenges and emerging directions in single-cell analysis. Genome Biol. 18, 84.

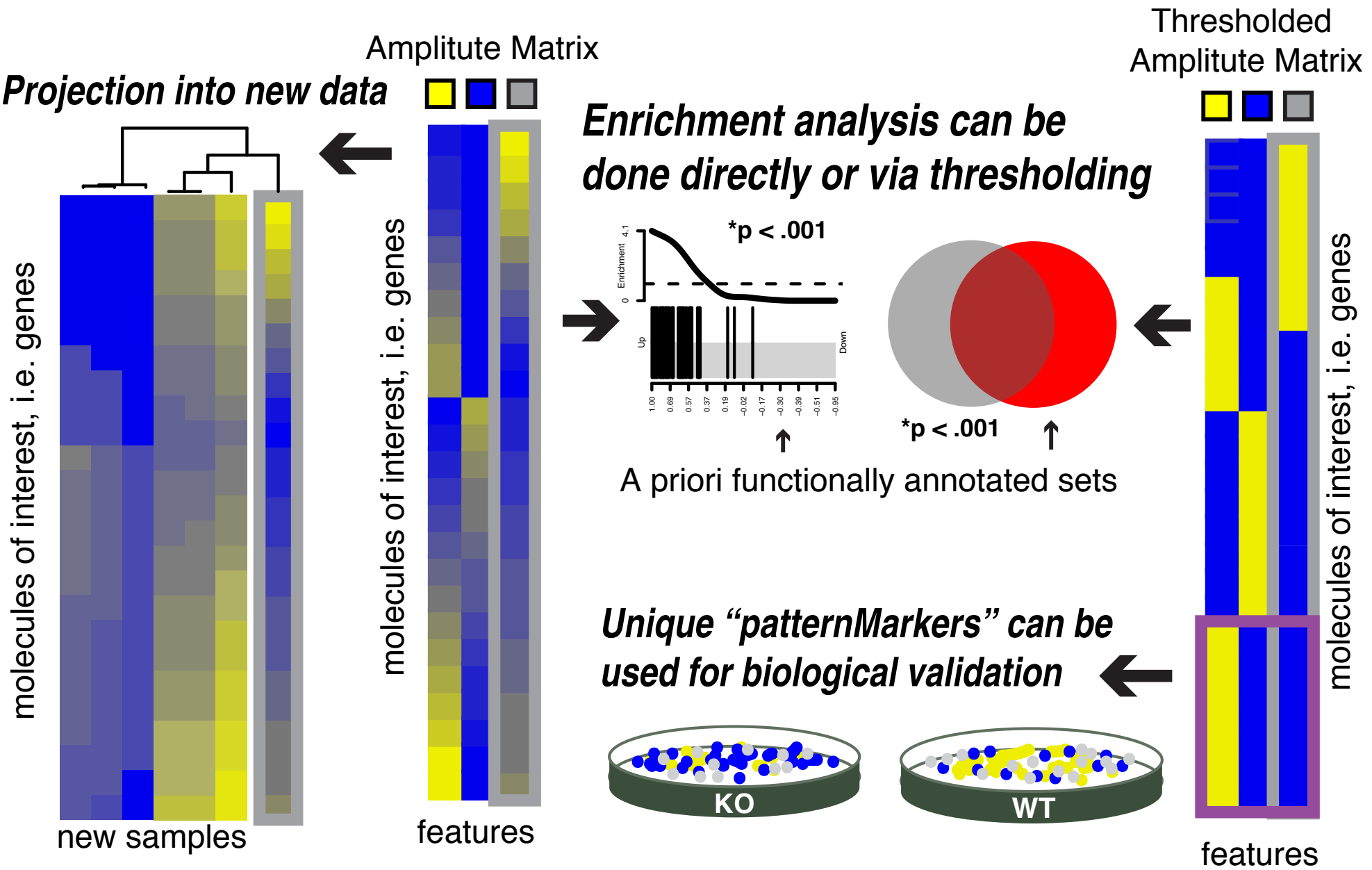
Zakeri, M., Srivastava, A., Almodaresi, F., Patro, R., 2017. Improved data-driven likelihood factorizations for transcript abundance estimation. Bioinformatics 33, i142–i151.

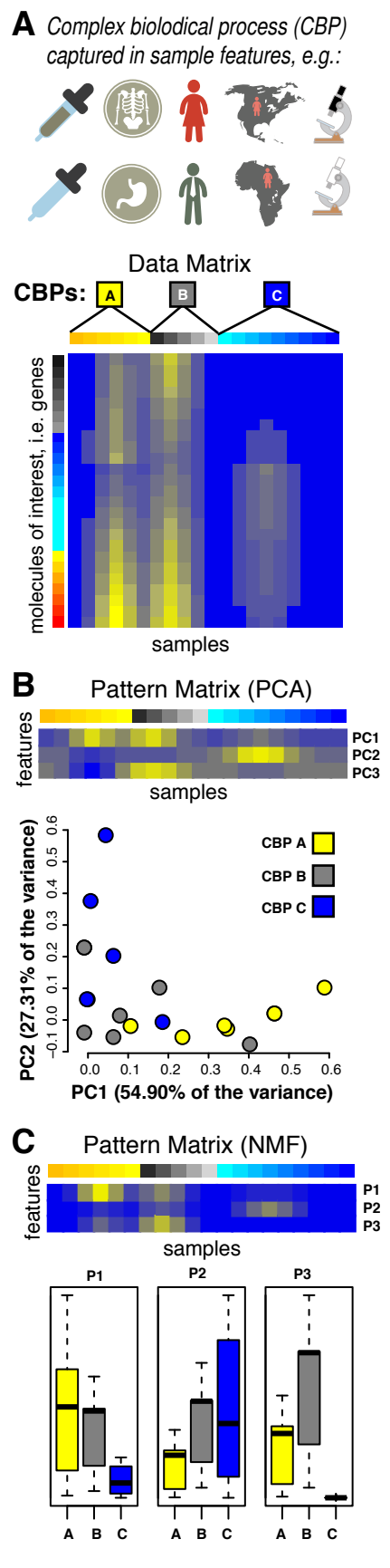
Zhang, Z.Y., Li, T., Ding, C., Ren, X.W., Zhang, X.S., 2010. Binary matrix factorization for analyzing gene expression data. Data Mining and Knowledge.

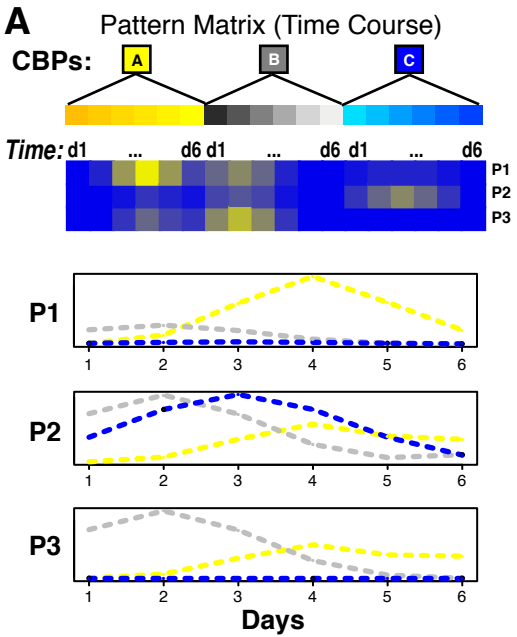
Zhou, G., Cichocki, A., Zhao, Q., Xie, S., 2014. Nonnegative Matrix and Tensor Factorizations : An algorithmic perspective. IEEE Signal Process. Mag. 31, 54–65.

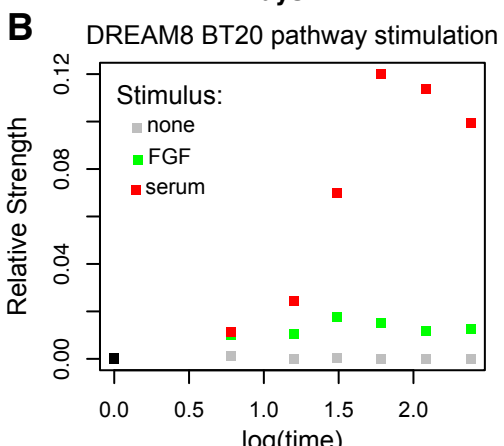
Zhou, G., Cichocki, A., Zhao, Q., Xie, S., 2014. Nonnegative Matrix and Tensor Factorizations : An algorithmic perspective. IEEE Signal Process. Mag. 31, 54–65.


A
Data
Amplitude
Pattern
molecules of interest, i.e. genes
samples
=
molecules of interest
features
X
features
samples
Columns of genetic, epigentics, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of cooresgulation.
Rows of continuous or binary weights associate with samples (each sample is a column) capturing realtionships including cell-types/lines, patients, or experimental conditions
B
Associated molecular signatures (metagenes, modules, etc.)
Patterns in groups of samples
molecules of interest
features
features
samples
max(P)
Magnitude of rows of the "Pattern" matrix scaled from
0


Projection into new data
Amplitute Matrix
Thresholded Amplitute Matrix
Enrichment analysis can be done directly or via thresholding
*p < .001
Enrichment
Up
Down
*p < .001
A priori functionally annotated sets
Unique "patternMarkers" can be used for biological validation
KO
WT
molecules of interest, i.e. genes
new samples
features
molecules of interest, i.e. genes
features
molecules of interest, i.e. genes


A
Complex biolodical process (CBP) captured in sample features, e.g.:
Data Matrix
CBPs:
A
B
C
molecules of interest, i.e. genes
samples
B
Pattern Matrix (PCA)
features
samples
PC1
PC2
PC3
CBP A
CBP B
CBP C
PC2 (27.31% of the variance)
PC1 (54.90% of the variance)
C
Pattern Matrix (NMF)
features
samples
P1
P2
P3
A
B
C


A
Pattern Matrix (Time Course)
CBPs:
A
B
C
Time: d1 ... d6 d1 ... d6 d1 ... d6
P1
P2
P3
P1
P2
P3
1 2 3 4 5 6
Days


B
DREAM8 BT20 pathway stimulation
Stimulus:
none
FGF
serum
Relative Strength
0.00
0.04
0.08
0.12
0.0 0.5 1.0 1.5 2.0
log(time)

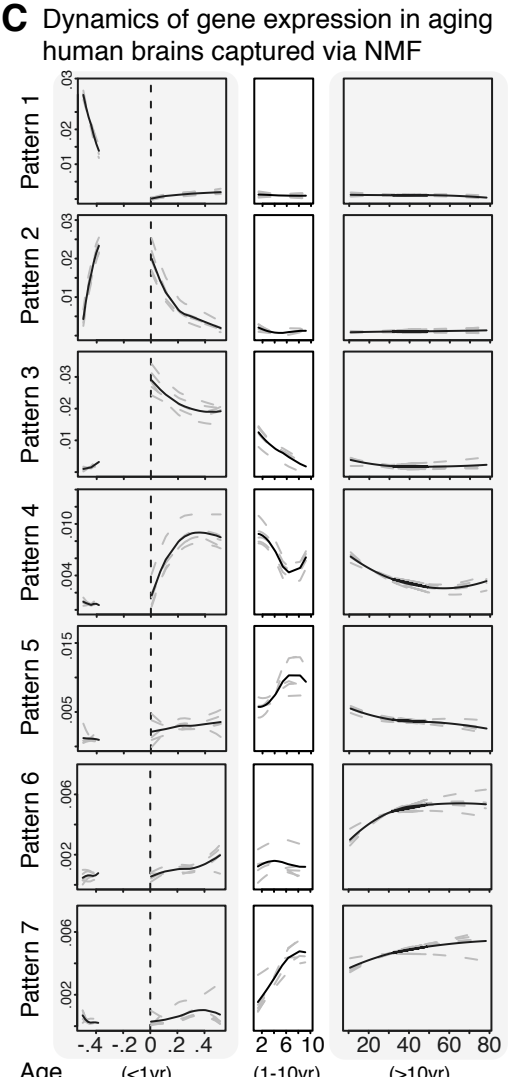

C
Dynamics of gene expression in aging human brains captured via NMF
Pattern 1
Pattern 2
Pattern 3
Pattern 4
Pattern 5
Pattern 6
Pattern 7
-.4 -.2 0 .2 .4
2 6 10
20 40 60 80
Age
(<1yr)
(1-10yr)
(>10yr)

Figure Captions

Figure 1
A) MF analyzes a data matrix that has each sample as a column and each observed value (expression counts, methylation levels, protein concentrations, etc.) as a row. MF seeks to identify the sources of variation within that data matrix by decomposing that data into two related matrices. B) The rows of the amplitude matrix contain the observations and the columns of the pattern matrix contain the samples. The corresponding profiles in the columns of the amplitude matrix or profiles in the rows of the sample matrix are both referred to as "features" or "components". Ideally, the relationships between samples reflect the comparative use of distinct sources of biological, experimental, and technical variation in each sample in the dataset.

Figure 2
MF can derive data-driven gene signatures which contain the context of the gene activity in the data matrix. The output of these algorithms are then an amplitude matrix with columns reflecting the correlation structure between genes. Each column can then be used as a pathway reflecting coregulation of the genes contained within (middle). In this way, MF techniques are able to account for gene reuse in multiple pathways. The values of most amplitude matrices are continuous weights describing relative contribution of a gene to a given pathway. However, as marker genes are often used to discretize biological phenomena for experimentation, gene set annotations are often binary (left). Methods based on high-weight genes comprise signatures akin to many curated gene set resources whereas genes that are most uniquely associated with a specific pattern may be biomarkers of the cell type or process associated with that pattern.

Figure 3
A) Example sources of variation are pharmacological perturbations, tissue/cell types, sex, inter-individual genetic variation, and biases in measurement technology, protocol, or sample batches. B) PCA learns features that explain most of the variation in the data, the algorithm will necessarily combine signal from the multiple sources of biological and technical factors in the data. This aspect of PCA can be exploited to learn features that explain most of the variance independent of known biological phenotypes to remove technical artifacts for supervised analysis or to learn features that explain most of the variance independent of the technical artifacts from known batches to retain biological signal for unsupervised analyses. C) Non-negative matrix factorization (NMF) methods instead constrain all elements of the amplitude and pattern matrices to be greater than zero which enables algorithms to learn shared components in multiple features thereby modeling co-regulation and dependent processes in biological systems.

Figure 4
A) The continuous weights for each sample in the pattern matrix can vary across samples collected across distinct time points. These weights can encode the timing of regulatory dynamics directly from the data. B) The relative strength of inhibition and stimulation experiments was inferred by comparing the magnitude of the rows of the pattern matrix for all experimental conditions at each time point of the HPN-DREAM challenge phospho-proteomics data. C) The pattern matrix learned from NMF of microarray data from post-mortem brain samples describes both the dynamics and heterogeneity of the molecular processes associated with aging.