Enter the matrix: factorization uncovers knowledge from omics

Names/Affiliations

Genevieve L. Stein-O'Brien[1,2,3], Raman Arora[4], Aedin C. Culhane[5,6], Alexander V. Favorov[1,7], Lana X. Garmire[8], Casey S. Greene[9,10], Loyal A. Goff[2,3], Yifeng Li[11], Aloune Ngom[12], Michael F. Ochs[13], Yanxun Xu[14], 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 University of Hawaii Cancer Center, Honolulu, HI

9 Department of Pharmacology, Perelman School of Medicine, University of Pennsylvania, PA

10 Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, PA

11 Digital Technologies Research Centre, National Research Council Canada

12 School of Computer Science, University of Windsor, Ontario, Canada

13 Department of Mathematics and Statistics, The College of New Jersey, Ewing, NJ,

14 Department of Applied Mathematics & Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD

*Correspondence: ejfertig@jhmi.edu (E.J. Fertig).

Keywords: genomics, unsupervised learning, matrix factorization, dimension reduction, integrated analyses, single cell

Abstract

Omics data contains signal from the molecular, physical, and kinetic inter- and intra-cellular interactions that control biological systems. Matrix factorization techniques can reveal low-dimensional structure from high-dimensional data that reflect these interactions. These techniques can uncover new biological knowledge from diverse high-throughput omics data in topics ranging from pathway discovery to time course analysis. We review exemplary applications of matrix factorization for systems-level analyses. We discuss appropriate application of these methods, their limitations, and focus on analysis of results to facilitate optimal biological interpretation. The inference of biologically relevant features with matrix factorization enables discovery from high-throughput data beyond the limits of current biological knowledge—answering questions from high-dimensional data that we have not yet thought to ask.

Determining the dimensions of biology from omics data

High-throughput technologies ushered in an era of big data in biology [1,2] and empowered *in silico* experimentation which is poised to characterize **complex biological processes (CBPs)** (RNAseq example given in Fig 1)[3]. The natural representation of high-dimensional biological data is a matrix of the measured values (expression counts, methylation levels, protein concentrations, etc.) in rows and individual samples in columns (Fig 2, Key Figure). Columns corresponding to experimental replicates, or samples with similar phenotypes will have values from the same distribution of biological variation. The related structure in the data is observed because they share one or more CBP. The activity of CBPs need not be identical in each sample. In these cases, the values of all molecular components that are associated with a CBP will change proportionally to the relative activity of that CBP. These phenotypes and CBP activities are often unknown a priori, requiring unsupervised computational techniques to learn them directly from the biological data.

The relationships between CBPs and similarities between samples constrain high-dimensional datasets to have low-dimensional structure. The number of genes, proteins, and pathways that are concurrently active within any cell is constrained by its energy and free-molecule limitations [4]. Only a characteristic subset of CBP will be active in any cell at a given time. Thus for a dataset where columns share CBP, a low dimensional structure can be extracted which is smaller than either the number of rows or the number of columns. Matrix Factorization (MF) is a class of unsupervised techniques provide a set of principled approaches to parsimoniously reveal the low-

dimensional structure while preserving as much information as possible from the original data (see Box 1).

When applied to high-throughput omics data, MF techniques learn two matrices: one describes the structure between features (e.g., genes) and another the structure between samples (Fig 2). Here, we call the former gene-level matrix the **Amplitude matrix** and the latter sample-level matrix the **Pattern matrix**. There are numerous approaches to MF, including both gradient-based and probabilistic methods (Box 1). Additional terms have been coined for the Amplitude and Pattern matrices based upon the MF problem applied and on the specific application to high-throughput biological data. Other reviews discuss the mathematical and technical details of MF techniques [5–8] and their applications to microarray data [9].

Here, we focus on the biological applications of MF techniques and the interpretation of their results since the advent of sequencing technologies. We describe a variety of MF techniques applied to high-throughput data analysis and compare and contrast their use for biological inference. Many techniques described are for sequencing data that is preprocessed with log transformation [10] or models of sequencing depth [11], while others directly model read counts [12]. We focus on examples in pathway analysis, subtype and clonal identification, time course analysis, multi-omics integration, and single cell data to present a field with much wider applications.

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

Genomics data are often interpreted by identifying molecular changes in sets of genes annotated to functionally related modules or pathways, called gene sets [13,14]. Often the association between gene sets and functions used are based upon human curation of the literature [15,16]. Such set-level interpretations often lack important contextual information [13,17,18] and cannot describe genes with unknown function or genes associated with new functional mechanisms.

The amplitude matrix from MF analysis can be used both for literature-based gene set analysis and to define new data-driven gene signatures (Fig 3). The values in each column of these amplitude matrices are continuous weights describing relative contribution of a molecule in each inferred factor. In cases where factors distinguish CBPs, the relative weights of these molecules can be associated with functional pathways. The same molecule may have high values in multiple columns of the amplitude matrix. Thus, MF techniques are able to account for the cumulative effect of genes that participate in multiple pathways. The properties of the amplitude matrix, and subsequently the interpretation of their values, depend critically on the specific MF problem and algorithm selected for analysis.

The three most prominent MF approaches are **Principal Component Analysis (PCA), Independent Component Analysis (ICA)**, and **Non-negative Matrix Factorization (NMF)**. Each of these techniques has a distinct mathematical formulation of a distinct MF problem that is described in the Box 2 and other reviews [5,8,19–22]. Briefly, PCA finds dominant sources of variation in high-dimensional datasets, inferring genes that distinguish samples. Maximizing the variability captured in certain factors, as

opposed to spreading relatively evenly among factors, may mix the signal from multiple CBPs in a single component. Therefore, PCA may conflate processes that sometimes occur and make interpretation of the amplitude matrix for define data-driven gene sets difficult.

To learn distinct processes, ICA learns factors that are statistically independent, resulting in more accurate associate with literature-derived gene sets [23–25]. Comparison analyses in Rotival et al [26] found that ICA could identify modules with known biological function. NMF methods constrain all elements of the amplitude and pattern matrices to be greater than or equal to zero [27,28]. NMF is well suited to transcriptional data, which is typically non-negative itself, and semi-NMF is also applicable to data that can have negative values. The assumptions of NMF model both the additive nature of CBPs and parsimony, generating solutions that are biologically intuitive to interpret [29].

The solutions from both ICA and NMF may vary depending upon the initialization of the algorithm, leading to disparate amplitude matrices. Therefore, it is critical to ensure that particular solution used for analysis provides an optimal and robust solution before using the amplitude matrix to define data-driven gene signatures. Bayesian techniques to solve NMF were found to have more robust amplitude matrices than gradient-based techniques, and thus more accurate associations of the values in the amplitude matrix with functional pathways [5,30]. These associations also depend critically on the input data. Therefore, to learn context dependent genes sets, MF can be applied to datasets with well-defined experimental perturbations [31].

Standard gene set analysis can be applied directly to the values in each column of the amplitude matrix to associate the inferred factors with literature-curated sets. New, context dependent gene sets can also be learned from the values in the amplitude matrix. Gene set annotations are often binary. Thresholding techniques to select which genes belong to a pathway from the amplitude matrix for binary membership provide an output similar to gene sets in databases [31,32]. Other studies also integrate the literature-derived gene signatures in these thresholds to refine the context of pathway databases [30,33]. The genes derived from these binarizations can be used as inputs to pathway analyses from differential expression statistics in independent datasets (Fig 3, right) and are analogous to the hierarchical-clustering based gene modules [34] and gene expression signatures from public domain studies in the MSigDB gene set database [35]. Another means of binarizing the data is to find genes that are most uniquely associated with a specific pattern to use as biomarkers of the cell type or process associated with that pattern [36,37]. Selecting genes based upon these statistics can facilitate visualization of the CBPs in high-dimensional data [36]. Whereas binarization of genes with high weights can associate a single gene with multiple CBPs, the statistics for unique associations link a gene with only one CBP. Therefore, these statistics also define specific genes that may be biomarkers of the cell type/state or a process [36] (Fig 3, right).

Although binary pathway models are substantially easier to interpret, continuous values from the original factorization provide a better model of the input data. Weighted gene signatures have been shown to be more robust to noise and missing values in the data [38]. If a gene's expression level is poorly measured in a sample, other genes in

the same factor can imply the actual expression level of the gene in question. By considering each gene in the context of all other genes, factorization improves the robustness of findings. Further, continuous signatures can be associated directly with other samples using projection methods [38,39] or profile correspondence methods [40].

MF has also been applied to learn functional gene modules on non-transcriptional datasets. Alexandrov et al applied NMF to the mutational landscape of tumors defined from the number of nucleotides that differ from the reference genome in the context of the preceding and following nucleotides [41]. This analysis defined mutational signatures associated with distinct cancer driving processes [42]. Additional extensions have been applied to learn allele combinations in phenotypes [43], transcript-regulation of genes [44], distributions of transcript lengths [45], and discriminate peptides in mass spectrometry proteomics [46].

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

Whereas each column of the amplitude matrix describes the relative contribution of molecules to a factor, each row of the pattern matrix describes the relative contribution of samples to a factor (Fig 2, Fig 4). Sample groups can be learned by comparing the relative weights in each row of the pattern matrix (Fig 4). The pattern matrix from MF can also be binarized to perform clustering [47,48] or kept as continuous values to define relationships between samples [49–51]. Just as molecules with high weights within a column of the amplitude matrix are associated with a common

pathway, samples with high weights within a row of the pattern matrix can be assumed to share a common phenotype or CBP. Although here we refer to each of these sample-level matrices as the Pattern matrix, numerous other terms have been adapted based on the MF method and its application (Box 3).

The application of PCA to SNP data from 3,000 European individuals [52] demonstrates inference of sample co-relationships using the pattern matrix and found that much of the variation in DNA sequence is explained by the longitude and latitude of an individual's country of origin. Additionally, statistical models can be formulated assuming that inheritance of an individual arises from proportions of ancestry in distinct populations through genetic admixture [52]. A MF based technique called sparse factor analysis also distinguishes these populations using GWAS data [53]. These analyses demonstrate that the ancestry of each individual is a dominant source variation in DNA sequence. Additional sources of variation in GWAS data arise from variants that give rise to disease risk, which can be shared among individuals with diverse genetic backgrounds [54]. These different sources of signal can give different sample groupings that reflect the biology of population genetics. For example, biologically a grouping inferred from GWAS which distinguishes ancestry is equally valid to a grouping inferred from the same GWAS data which distinguishes disease risk.

A single factorization of complex datasets can find multiple distinct sources of variation. For example, the power of MF to identify multiple sources of variation was seen when multiple technical factors from sample processing and biological factors were discovered in an ICA of gene expression profiles of 198 bladder cancer samples [55]. One factor in the pattern matrix of this analysis defined a CBP associated with

gender. Because ICA simultaneously accounts for multiple factors in the data as separate rows in the same matrix, each row can fully distinguish a single biological grouping from the data.

Applying multiple types of MF techniques to the same dataset or a single MF to distinct subsets of a dataset can also find distinct sources of variability. For example, an NMF-class algorithm separated tissue-specific patterns from gene expression data for postmortem samples in the Genotype-Tissue Expression (GTEx) project [56]. 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 called CoGAPS simultaneously separated these brain regions from GTEx with additional patterns that are associated with the individual who donated those samples [36,57]. Both of these algorithms are equally valid, and their distinct formulation gives rise to the distinct features observed in the data. Applications of multiple types of MF techniques, or even the same MF algorithm with different parameters, may infer several CBPs or phenotypes within a single dataset, in essence providing answers to different questions.

Analysis of a single dataset with one MF algorithm using different numbers of factors can reflect a hierarchy of biological processes. For example, applying CoGAPS to data from a set of head and neck tumors and controls for a range of dimensionalities was able to separate tumor and normal samples when limited to two patterns, but further decomposed the tumor samples into the two dominant clinical subtypes of head and neck cancer when identifying five patterns for the same data [58]. The hierarchical relationship between patterns has been used to assess the robustness of patterns to

quantify the optimality of the factorization [59] and learn the optimal dimensionality of the factorization [60]. Other algorithms use statistical metrics to estimate the number of factors [12,61]. While these algorithms quantify fit to the data, they may disregard the hierarchical nature of distinct CBPs learned by factoring biological data into multiple dimensions. This observation highlights the complexity of estimating the number of factors for optimal MF analysis of biological data (see Outstanding Questions).

Intra- and inter-tumor heterogeneity introduce a further degree of complexity to MF analysis of biological variation in their molecular data. **Computational microdissection** algorithms estimate the proportion of distinct cell types within a bulk samples by applying MF to genes whose expression are uniquely associated with each cell type [62]. Subsetting the data to different genes may give rise to different factors that represent different CBPs. Nonetheless, CoGAPS analysis of data subsets that were obtained by selecting equally sized sets of random genes found that the pattern matrices were consistent for each random gene set in expression data[36,63]. These results suggest that the dependency of a MF on the specific genes used for analysis may depend on the heterogeneity of the signal in the data matrix.

Even a pure tumor tissue can contain numerous subclones due to the accumulation of different driver events during tumor evolution. New MF techniques have been developed to estimate the proportion of the tumor that arises from each subclone [61,64–67]. Assumptions about the evolutionary mechanisms of the accumulation of molecular alterations can also be encoded in the factorization to model the resulting heterogeneity of these clones [12,61]. These studies demonstrate that encoding prior

knowledge into MF can focus the resulting factors to reflect one of the equally valid biological groupings within the data.

From snapshots to moving pictures: simplifying time course analysis

Entwined in the challenge of decomposing cell types and subpopulations is the fact that CBPs change over time. High-throughput time course datasets are emerging in the literature to account for the dynamics of biological systems. The central goal of time course analysis is to determine the extent to which 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 [68,69]. MF analyses can naturally infer changes in CBPs over time when applied to time course data because the continuous weights for each sample in the pattern matrix can vary across samples collected across distinct time points. The relative weights of rows of the pattern matrix can encode the timing of regulatory dynamics directly from the data (Fig 4A).

Both ICA and NMF were found to have signatures characterizing the yeast cell cycle and metabolism in early time-course microarray experiments [70,71]. The Sparse NMF techniques using Bayesian method had patterns that reflected the smooth dynamics of these phases [30,70]. This approach has been shown to simultaneously learn pathway inhibition and transitory response to chemical perturbation of cancer cells [72] and relate the the changes in phospho-proteomics trajectories between multiple therapies [73]. Similar analysis of healthy brain tissues learned the dynamics of

transcriptional alterations common to the ageing process from multiple individuals [63]. MF techniques designed for cancer subclones 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 [74]. These discrepancies by data modality suggest that different regulatory mechanisms may be responsible for initiating and stabilizing the malignant phenotype [74].

Integrated analysis of multiple omics data

Multi-omics data are generated in order to elucidate the molecular networks that govern phenotypes. MF can be applied to learn shared features between datasets [7,8,75]. Integration may occur between datasets with distinct samples measured with the same molecular type using different measurement technologies or in distinct technical batches. For example, an analysis of multiple microarray studies of the same cell lines across different platforms with an MF approach designed to find components that maximize covariance or correlated information discovers which microarray platforms have the most informative set of genes [76]. Gene regulation can also be inferred by applying MF to data with different molecular components of the same samples. For example, repression of gene expression by promoter methylation had been encoded to integrate gene expression and DNA methylation data of tumor samples [58]. Other techniques have used joint modeling of features learned separately

in different data types to perform integrated inference [77–79]. Techniques that extend this integrated MF framework, including Bayesian group factor analysis [8] and tensor decomposition [80] can also identify common and specific factors among different molecular levels [81]. Developing such data integration techniques is an active area of research in both genomics and computational sciences. When mature, these techniques will be able to formulate complete gene regulatory networks to further systems biology.

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

MF approaches are a natural choice in Single-cell RNA-sequencing (scRNA-seq) data analysis due to its high dimensionality and are used to identify and remove batch effects, summarize CBPs, and annotate cell types in the data [82–85]. Whereas the analysis in bulk data dissects groups from a small subset of samples, the analysis in scRNA-seq data aggregates cells into groups of common cell types or CBPs [37,86]. Often these analyses are performed on a subset of the data containing the most variable genes. Newer computationally efficient methods are being developed to enable factorization of large omics datasets for genome-wide analysis [79]. Biological knowledge can be encoded with a class of MF algorithms that summarizes factors using gene sets [82,87,88].

Most MF techniques developed for bulk omics data are linear, namely they assume the gene expression changes from CBPs are additive. This assumption is

violated in scRNA-seq data. One reason for the violation of the linearity assumption in MF is the the inability to distinguish true zeros from missing values. Imputation methods for preprocessing [84,89] or newer MF algorithms that model missing data are essential for scRNA-seq data. Branching of trajectories of cellular states and lack of synchronization of cell cycle in scRNA-seq data further violate the linearity assumption in MF. New non-linear factorization techniques are being developed to enhance visualization of trajectory structures in single cell data [83,90–92] in these cases.

Concluding remarks

MF is a versatile class of techniques with broad applications to unsupervised clustering, biological pattern discovery, component identification, and prediction. Since MF was first applied to microarray data analysis in the early 2000s [70,93–95], the breadth of MF problems and algorithms for high-throughput biology has grown with their broad applications. MF problems are ubiquitous in the computational sciences, with examples including unsupervised feature learning [96–101] clustering and metric learning [102–104], subspace learning [105–110], multiview learning [111], matrix completion [112], multi-task learning [113] semi-supervised learning [114], compressed sensing [115], and similarity-based learning [116,117]. Dimension reduction of biological data with MF highlights perspectives and questions that investigators have not yet considered, and also enables tractable exploration of otherwise massive datasets.

Different classes of techniques solve MF, including gradient-based and probabilistic methods (see Box 1). Distinct MF problems each aim 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 MF techniques 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 datasets and choosing those that are broad enough to cover the relevant sources of variability are essential. Advances to MF and related techniques will be essential to powering systems level analyses from big data (see Outstanding Questions).

Acknowledgements

We thank Orly Alter, J Brian Byrd, Michael Love, Irene Gallego Romero, Lillian Fritz-Laylin, Luciane Kagohara, Louise Klein, Craig Mak, Matthew Stephens, Daniela Witten, and other 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.A.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), K01ES025434 award by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (LXG), P20 COBRE GM103457 award by NIH/NIGMS (LXG), R01 LM012373 award by NLM (LXG), R01 HD084633 award by NICHD (LXG), 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

1 Bell, G. *et al.* (2009) Beyond the data deluge. *Science*

2 Sagoff, M. (2012) Data deluge and the human microbiome project. *Issues Sci. Technol.* at <http://www.jstor.org/stable/43315648>

3 Alter, O. (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc. Natl. Acad. Sci. U. S. A.* 103, 16063–16064

4 Heyn, P. *et al.* (2015) Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences. *Bioessays* 37, 148–154

5 Ochs, M.F. and Fertig, E.J. (2012) Matrix Factorization for Transcriptional Regulatory Network Inference. *… Bioinformatics and Computational Biology …*

6 Abdi, H. *et al.* (2013) Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Comp Stat* 5, 149–179

7 Meng, C. *et al.* (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* 17, 628–641

8 Li, Y. *et al.* (2016) A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.*

9 Devarajan, K. (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput. Biol.* 4, e1000029

10 Ritchie, M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47

11 Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* 11, R106

12 Xie, F. *et al.* (2017) BayCount: A Bayesian Decomposition Method for Inferring

Tumor Heterogeneity using RNA-Seq Counts. at <https://arxiv.org/abs/1702.07981>

13 Khatri, P. *et al.* (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8, e1002375

14 Irizarry, R.A. *et al.* (2009) Gene set enrichment analysis made simple. *Stat. Methods Med. Res.* 18, 565–575

15 Bauer-Mehren, A. *et al.* (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Mol. Syst. Biol.* 5, 290

16 Tsui, I.F.L. *et al.* (2007) Public databases and software for the pathway analysis of cancer genomes. *Cancer Inform.* 3, 379–397

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

18 Tan, J. *et al.* (2017) Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. *Cell systems* at <http://linkinghub.elsevier.com/retrieve/pii/S2405471217302314>

19 Meng, C. *et al.* (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* 17, 628–641

20 Berry, M.W. *et al.* (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.* 52, 155–173

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

22 Zhou, G. *et al.* (2014) Nonnegative Matrix and Tensor Factorizations : An algorithmic perspective. *IEEE Signal Process. Mag.* 31, 54–65

23 Lee, S.-I. and Batzoglou, S. (2003) Application of independent component analysis

to microarrays. *Genome Biol.* 4, R76

24 Engreitz, J.M. *et al.* (2010) Independent component analysis: Mining microarray data for fundamental human gene expression modules. *Journal of biomedical …* 43, 932–944

25 Teschendorff, A.E. *et al.* (2007) Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.* 3, e161

26 Rotival, M. *et al.* (2011) Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Trans-Regulated Gene Modules in Humans. *PLoS Genet.* 7, e1002367

27 Ochs, M.F. *et al.* (1999) A New Method for Spectral Decomposition Using a Bilinear Bayesian Approach. *J. Magn. Reson.* 137, 161–176

28 Lee, D.D. and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* at <http://search.proquest.com/openview/81c8bfec1d4e36de7aea730ec5c77816/1?pq-origsite=gscholar&cbl=40569>

29 Moloshok, T.D. *et al.* (2002) Application of Bayesian decomposition for analysing microarray data. *Bioinformatics* 18, 566–575

30 Kossenkov, A.V. *et al.* (2007) Determining transcription factor activity from microarray data using Bayesian Markov chain Monte Carlo sampling. *Stud. Health Technol. Inform.* 129, 1250–1254

31 Fertig, E.J. *et al.* (2012) 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

32 Kim, J.W. *et al.* (2017) Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. *Cell Syst* 5, 105–118.e9

33 Fertig, E.J. *et al.* (2012) , Identifying context-specific transcription factor targets from prior knowledge and gene expression data. , in *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1–6

34 Segal, E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.* at <http://search.proquest.com/openview/1a596f16c8b83504d735d79eab5b763c/1?pq-origsite=gscholar&cbl=33429>

35 Subramanian, A. *et al.* (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

36 Stein-O'Brien, G.L. *et al.* (2017) PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. *Bioinformatics* at <https://academic.oup.com/bioinformatics/article/2975325/PatternMarkers>

37 Zhu, X. *et al.* (2017) Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. *PeerJ* 5, e2888

38 DeTomaso, D. and Yosef, N. (2016) FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics* 17, 315

39 Fertig, E.J. *et al.* (2013) Identifying context-specific transcription factor targets from prior knowledge and gene expression data. *IEEE Trans. Nanobioscience* 12, 142–149

40 Irizarry, R.A. *et al.* (2005) Multiple-laboratory comparison of microarray platforms.

Nat. Methods 2, 345–350

41 Alexandrov, L.B. *et al.* (2013) Signatures of mutational processes in human cancer. *Nature* 500, 415–421

42 Alexandrov, L.B. *et al.* (2016) Mutational signatures associated with tobacco smoking in human cancer. *Science* 354, 618–622

43 Favorov, A.V. *et al.* (2005) A Markov chain Monte Carlo technique for identification of combinations of allelic variants underlying complex diseases in humans. *Genetics* 171, 2113–2121

44 Zakeri, M. *et al.* (2017) Improved data-driven likelihood factorizations for transcript abundance estimation. *Bioinformatics* 33, i142–i151

45 Bertagnolli, N.M. *et al.* (2013) SVD identifies transcript length distribution functions from DNA microarray data and reveals evolutionary forces globally affecting GBM metabolism. *PLoS One* 8, e78913

46 Peckner, R. *et al.* 08-Sep-(2017) , Specter: linear deconvolution as a new paradigm for targeted analysis of data-independent acquisition mass spectrometry proteomics. , *bioRxiv*, 152744

47 Yeung, K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987

48 Jiang, D. *et al.* (2004) Cluster analysis for gene expression data: a survey. *IEEE Trans. Knowl. Data Eng.* 16, 1370–1386

49 Venet, D. *et al.* (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics* at <https://academic.oup.com/bioinformatics/article-abstract/17/suppl_1/S279/262438>

50 Abbas, A.R. *et al.* (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* 4, e6098

51 Erkkilä, T. *et al.* (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. *Bioinformatics* 26, 2571–2577

52 Novembre, J. *et al.* (2008) Genes mirror geography within Europe. *Nature* 456, 98–101

53 Engelhardt, B.E. and Stephens, M. (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. *PLoS Genet.* 6, e1001117

54 McCarthy, M.I. *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev.* at <http://search.proquest.com/openview/a6e3158ffdfed42590298c6b633169bf/1?pq-origsite=gscholar&cbl=44267>

55 Biton, A. *et al.* (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

56 Dey, K.K. *et al.* (2017) Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet.* 13, e1006599

57 Fertig, E.J. *et al.* (2010) CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics* 26, 2792–2793

58 Fertig, E.J. *et al.* (2013) Preferential Activation of the Hedgehog Pathway by Epigenetic Modulations in HPV Negative HNSCC Identified with Meta-Pathway Analysis. *PLoS One* 8, e78127

59 Bidaut, G. (2010) Interpreting and Comparing Clustering Experiments Through Graph Visualization and Ontology Statistical Enrichment with the ClutrFree Package. In *link.springer.com* Chapter 19 vols.pp. 315–333, Springer US

60 Bidaut, G. *et al.* (2006) Determination of strongly overlapping signaling activity from microarray data. *BMC Bioinformatics* 7, 99

61 Xu, Y. *et al.* (2015) MAD Bayes for Tumor Heterogeneity—Feature Allocation With Exponential Family Sampling. *J. Am. Stat. Assoc.* 110, 503–514

62 Hackl, H. *et al.* (2016) Computational genomics tools for dissecting tumour-immune cell interactions. *Nat. Rev. Genet.* 17, 441–458

63 Fertig, E.J. *et al.* (2014) Pattern Identification in Time-Course Gene Expression Data with the CoGAPS Matrix Factorization. *Methods Mol. Biol.* 1101, 87–112

64 Nik-Zainal, S. *et al.* (2012) The Life History of 21 Breast Cancers. *Cell* 149, 994–1007

65 Roth, A. *et al.* (2014) PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* 11, 396–398

66 Deshwar, A.G. *et al.* (2015) PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* 16, 35

67 Lee, J. *et al.* (2016) Bayesian inference for intratumour heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. Ser. C Appl. Stat.* 65, 547–563

68 Bar-Joseph, Z. *et al.* (2012) Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* 13, 552–564

69 Liang, Y. and Kelemen, A. (2017) Dynamic modeling and network approaches for omics time course data: overview of computational approaches and applications.

Brief. Bioinform. at <https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbx036>

70 Moloshok, T.D. *et al.* (2002) Application of Bayesian decomposition for analysing microarray data. *Bioinformatics* 18, 566–575

71 Liebermeister, W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics* at <https://academic.oup.com/bioinformatics/article-abstract/18/1/51/243636>

72 Ochs, M.F. *et al.* (2009) Detection of Treatment-Induced Changes in Signaling Pathways in Gastrointestinal Stromal Tumors Using Transcriptomic Data. *Cancer Res.* 69, 9125–9132

73 Hill, S.M. *et al.* (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

74 Stein-O'Brien, G. *et al.* 01-Aug-(2017) , Integrated time-course omics analysis distinguishes immediate therapeutic response from acquired resistance. , *bioRxiv*, 136564

75 Huang, S. *et al.* (2017) More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Front. Genet.* 8, 84

76 Fagan, A. *et al.* (2007) A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics* 7, 2162–2171

77 Meng, C. *et al.* (2016) moCluster: Identifying Joint Patterns Across Multiple Omics

Data Sets. *J. Proteome Res.* 15, 755–765

78 Mo, Q. *et al.* (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the …* at <http://www.pnas.org/content/110/11/4245.short>

79 Hübschmann, D. *et al.* Deciphering programs of transcriptional regulation by combined deconvolution of multiple omics layers. DOI: 10.1101/199547

80 Hore, V. *et al.* (2016) Tensor decomposition for multiple-tissue gene expression experiments. *Nat. Genet.* 48, 1094–1100

81 Kolda, T. and Bader, B. (2009) Tensor Decompositions and Applications. *SIAM Rev.* 51, 455–500

82 Fan, J. *et al.* (2016-3) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat. Methods* 13, 241–244

83 Trapnell, C. *et al.* (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386

84 William Townes, F. *et al.* (2017) Varying-Censoring Aware Matrix Factorization for Single Cell RNA-Sequencing. *biorxiv* DOI: 10.1101/166736

85 Moon, K.R. *et al.* (2017) PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data. *bioRxiv*

86 Puram, S.V. *et al.* (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* 171, 1611–1624.e24

87 Buettner, F. *et al.* (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biol.* 18, 212

88 Buettner, F. *et al.* 15-Nov-(2016) , Scalable latent-factor models applied to single-

cell RNA-seq data separate biological drivers from confounding effects. , *bioRxiv*, 087775

89 van Dijk, D. *et al.* 25-Feb-(2017) , MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. , *bioRxiv*, 111591

90 Risso, D. *et al.* (2017) ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv* at <http://biorxiv.org/content/early/2017/04/06/125112.abstract>

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

92 Maaten, L. van der and Hinton, G. (2008) Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605

93 Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. U. S. A.* 97, 10101–10106

94 Fellenberg, K. *et al.* (2001) Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10781–10786

95 Brunet, J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U. S. A.* 101, 4164–4169

96 Abdi, H. and Williams, L.J. (2010) Principal component analysis. *WIREs Comp Stat* 2, 433–459

97 Hyvärinen, A. *et al.* (2004) *Independent Component Analysis*, John Wiley & Sons.

98 Hardoon, D.R. *et al.* (2004) Canonical correlation analysis: an overview with application to learning methods. *Neural Comput.* 16, 2639–2664

99 Scholkopf, B. *et al.* (1999) , Kernel principal component analysis. , in *ADVANCES IN KERNEL METHODS - SUPPORT VECTOR LEARNING*

100 Arora, R. and Livescu, K. (2012) Kernel CCA for multi-view learning of acoustic features using articulatory measurements. *Symposium on Machine Learning in Speech* at <http://www.isca-speech.org/archive/mlslp_2012/ml12_034.html>

101 Andrew, G. *et al.* (2013) , Deep Canonical Correlation Analysis. , in *International Conference on Machine Learning*, pp. 1247–1255

102 Ding, C. and He, X. (2004) , K-means Clustering via Principal Component Analysis. , in *Proceedings of the Twenty-first International Conference on Machine Learning*, Banff, Alberta, Canada, pp. 29–

103 Arora, R. *et al.* (2011) Clustering by left-stochastic matrix factorization. *Proceedings of the 28th International* at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.448.4587&rep=rep1&type=pdf>

104 Kulis, B. (2013) Metric Learning: A Survey. *Foundations and Trends® in Machine Learning* 5, 287–364

105 De la Torre, F. and Black, M.J. (2003) A Framework for Robust Subspace Learning. *Int. J. Comput. Vis.* 54, 117–142

106 [PDF]Computer Vision: Algorithms and Applications - Szeliski.org. at <http://szeliski.org/Book/drafts/SzeliskiBook_20100903_draft.pdf>

107 Candès, E.J. *et al.* (2011) Robust principal component analysis? *J. ACM* 58, 11

108 Arora, R. *et al.* (2012) , Stochastic optimization for PCA and PLS. , in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*,

pp. 861–868

109 Arora, R. *et al.* (2013) Stochastic Optimization of PCA with Capped MSG. In *Advances in Neural Information Processing Systems 26* (Burges, C. J. C. et al., eds), pp. 1815–1823, Curran Associates, Inc.

110 Goes, J. *et al.* (2014) , Robust Stochastic Principal Component Analysis. , in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 33, pp. 266–274

111 Bickel, S. and Scheffer, T. (2004) , Multi-view clustering. , in *ICDM*, 4, pp. 19–26

112 Candès, E.J. and Recht, B. (2009) Exact Matrix Completion via Convex Optimization. *Found. Comut. Math.* 9, 717

113 Argyriou, A. *et al.* (2007) Multi-Task Feature Learning. In *Advances in Neural Information Processing Systems 19* (Schölkopf, B. et al., eds), pp. 41–48, MIT Press

114 Ando, R.K. and Zhang, T. (2005) A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data. *J. Mach. Learn. Res.* 6, 1817–1853

115 Cleary, B. *et al.* (2017) Composite measurements and molecular compressed sensing for highly efficient transcriptomics. *bioRxiv* at <http://biorxiv.org/content/early/2017/01/02/091926.abstract>

116 Aha, D.W. *et al.* (1991) Instance-based learning algorithms. *Mach. Learn.* 6, 37–66

117 Arora, R. *et al.* (2013) Similarity-based clustering by left-stochastic matrix factorization. *J. Mach. Learn. Res.* at <http://www.jmlr.org/papers/volume14/arora13a/arora13a.pdf>

Box 1: Technical description of matrix factorization (400 words)

Matrix factorization approximates an input matrix X, as a product of two matrices, U and V, also referred to as factors.

In genomics, the input matrix X typically represents a preprocessed data matrix that has n molecular measurements as rows and p biological samples as columns (i.e., a matrix of size $n \times p$). Often, not all of the row and column vectors of the data are equally informative. Exploratory data analysis aims to find a set of k most informative dimensions or features, where k is smaller than either n or p. A k-dimensional representation that captures most of the "information" (say, for example, most of the variation) contained in the original data can then be used in its place in subsequent analysis.

MF solves this problem by approximating X as a product of two "factor" matrices: $X \approx UV$, where the factor U is of size $n \times k$ and the factor V is of size $k \times p$. Here, we refer to U as the amplitude matrix and V as the pattern matrix. We note that alternative variable names to U and V and terminology are often used in the literature, depending on the application.

Finding matrices U and V requires a quantitative definition of how well the product UV approximates X. One method for providing this definition is to use a loss function that quantifies the discrepancy between the approximation and the data, such as the mean square error and median absolute error. A common approach to solve for U and V is to apply an iterative approach such as a gradient-based method to minimize a loss

function. Bayesian approaches to MF are alternatives to approximate the probabilistic relationship between UV and X. Additional conditions can also be incorporated in both gradient-based and Bayesian MF to learn useful features, such as sparsity constraints to limit the number of non-zero matrix elements.

Both gradient-based and Bayesian MF approaches have numerous applications beyond genomics. These techniques are ubiquitous in unsupervised feature learning for big data analysis.

Box 2: PCA, ICA, and NMF (400 words)

MF can start with the same data matrix X and then solve different problems to learn amplitude (U) and pattern (V) matrices with specific properties. An MF method requires: 1) the desired dimension k; 2) a loss function that measures the approximation $X \approx UV$; and 3) constraints on U and/or V that enforce a desired structure on the low-dimensional representation. Changes to any of these give rises to different MF problems with different low-dimensional representations of the data. Three common MF approaches are applied to genomics: PCA, ICA, and NMF.

PCA finds a low-dimensional representation of the data that maximizes the variation contained in the original data. In PCA, the columns of U are orthogonal to each other describing non-overlapping structure in the data. Each column of represents the weight or loading of the data vector in the corresponding column of matrix X. In PCA, the columns of U and rows of V are ranked by the relative amount of variance they explain in the data. Thus, the first column of U and row of V explain most of the variation across molecular measurements and samples, respectively. Together, the *k* vectors learned with PCA maximize the total variance in the data captured using any rank *k* factorization. Geometrically, the vectors can be thought of as a set of orthogonal coordinate axes in high dimensional space, which represent the directions of maximal variation in the data.

ICA assumes that there are a set of k independent sources of variation that give rise to the observed data matrix X. This method enforces that the columns of U yields

components that are statistically independent of each other. ICA is solved by minimizing the total mutual information between the k estimated components. The resulting factors ideally represent independent sources of variation in the biological system.

Non-negative matrix factorization (NMF) is a group of algorithms that constrains all elements of the U and V matrices to be greater than or equal to zero. The non-negativity constraint makes the representation purely additive, with no sources that can explain the data by removing signal. This non-negativity often results in NMF producing a *sparse* representation of the data. The additivity and sparsity make the k features inferred from NMF easy to interpret as the set of *active* components of the data that will give the original data when added together.

Box 3: Common terminology in the literature (400 words)

Historically, the independent discovery of MF in multiple fields including mathematics, computer science, and statistics created distinct terminologies that are often used interchangeably as analytical orthologs in genomics. For example, the term factorization is often used interchangeably with decomposition. Other terms, such as features, components, latent variable, or latent factors are both used to to refer to relationships between molecular measurements or samples, depending on the context.

The specific terminology for the amplitude and pattern matrices also varies by method and preferably labeled with different variable name. In PCA, the amplitude matrix $U$is often called the score or rotation matrix; and pattern matrix V is called the loadings. In ICA, the amplitude matrix is called the unmixing matrix (labeled as A) and the pattern matrix is called the source matrix (labeled as S). In NMF, the amplitude matrix is commonly called the weights matrix (labeled as W) and the pattern matrix is called the features matrix (labeled as H).

Throughout this paper, we use the amplitude matrix to refer to matrix that contains vectors which represent relationships between molecular measurements. Other terms used in the literature for these molecular relationships include modules, meta-pathways, or signatures. Similarly, we use the pattern matrix to refer to the matrix that contains vectors which represent relationships between samples. Other literature terms for these sample-level relationships include patterns, metagenes, eigengenes, sources, or controlling factors.

Glossary (450 words)

Amplitude matrix The matrix learned from MF that contains molecules in rows and factors in columns. Each column represents the relative contribution of the genes in a factor, which can be used to define a molecular signature for a CBP.

Complex biological process (CBP) The coregulation or coordinated effect of multiple molecular species resulting in one or more phenotypes examples can range from activation of multiple proteins in a single cellular signaling pathway to epistatic regulation of development.

Computational microdissection A computational method to learn the composition of a heterogeneous sample, e.g., the cell types in a tissue sample.

Independent Component Analysis (ICA) A MF technique that learns statistically independent factors.

Matrix Factorization (MF) A technique to approximate a data matrix by the product of two matrices (see Box 1), one of which we call the amplitude matrix and the other the pattern matrix.

Non-negative Matrix Factorization (NMF) A MF technique for which all elements of the amplitude and pattern matrices are greater than or equal to zero.

Pattern matrix The matrix learned from MF that contains factors in rows and samples in columns. Each row represents the relative contribution of the samples in a factor, which can be used to define the relative activity of CBPs in each sample.

Principal Component Analysis (PCA) A MF technique that learns orthogonal factors ordered by the relative amount of variation of the data that they explain.

Figure Legends

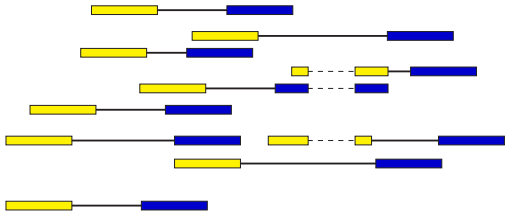
Figure 1 Pipeline for matrix factorization analysis of RNA-seq data. First, the RNA-seq data is preprocessed by alignment, gene-level quantification, and normalization. The data must aslo be log transformed for most matrix factorization methods, with the exception of specialized methods that have been developed for count data. Three dominant types of MF problems can be applied to analyze this data: Principle Component Analysis (PCA), Non-negative Matrix Factorization (NMF), or Independent Component Analysis (ICA). Each of these problems find distinct molecular and sample relationships in the amplitude and pattern matrices, respectively. The Amplitude matrix learned that reflects molecular relationships can be used for gene set analysis, pathway analysis, and biomarker discovery. The Pattern matrix that reflects sample relationships can be used for clustering, subtype / subclone discovery, and timecourse analysis.

Figure 2 Omics technologies yield a data matrix that has each sample as a column and each observed molecular value (expression counts, methylation levels, protein concentrations, etc.) as a row. This data matrix is preprocessed with techniques specific to each measurement technology and then input to an MF technique for analysis. MF decomposes the preprocessed data matrix into two related matrices that represent its sources of variation: an amplitude matrix and a pattern matrix. The rows of the amplitude matrix quantify the sources of variation among the molecular observations and the columns of the pattern matrix quantify the sources of variation among the samples. The matrix product of the amplitude and pattern matrices approximates the preprocessed input data matrix. The number of columns of the amplitude matrix equals the number of rows in the pattern matrix, and represents the number of dimensions in the low-dimensional representation of the data. Ideally, a pair of one column in the amplitude matrix and the corresponding row of the pattern matrix represents a distinct source of biological, experimental, and technical variation in each sample (called complex biological processes, CBPs). The values in the column of the amplitude matrix then represent the relative weight of each molecule in the CBP and the values in the row of the pattern matrix its relative role in each sample.

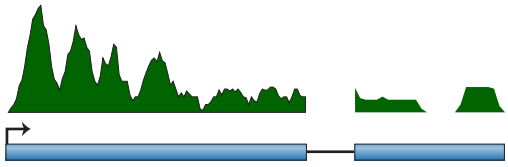
Figure 3 The amplitude matrix from MF can be used to derive data-driven molecular signatures associated with a CBP. The columns of the amplitude matrix contain continuous weights describing relative contribution of a molecule in a CBP (middle). The resulting molecular signature can be analyzed in a new dataset to determine the samples in which each previously detected CBP occurs, and thereby assess function in a new experiment. This comparison may be done by comparing the continuous weights in each column of the amplitude matrix directly to the new dataset (left). The amplitude matrix may also be used in traditional gene set analysis (right). Traditional gene set analysis using literature curated gene sets can be performed on the values in each column of the amplitude matrix to identify whether a CBP is occurring in the input data. Data-driven gene sets can also be defined from this matrix directly using binarization, and used in place of literature curated gene sets to query CBPs in a new dataset. Sets defined from molecules with high-weights in the amplitude matrix comprise signatures akin to many curated gene set resources whereas molecules that are most uniquely associated with a specific factor (purple box) may be biomarkers.

Figure 4 The pattern matrix from MF describes sample participation in a CBP. Trends or groupings of samples in a row of the pattern matrix can be tested against additional sample metadata to further define the biological details of a given CBP. Depending on the MF method used different visualizations can highlight specific aspects of the biology. **A.** Nonnegative matrix factorization method find the relative values of correlated or dependent measures just as pathway usage or time course data. If sample grouping are known *a priori* the relative usage of a CBP can then be inferred by comparing the weights of the pattern matrix between groups. Similarly, if samples correspond to time points the rows of the pattern matrix can be plotted as a function of time and sample condition to infer the dynamics of CBPs. **B.** Principal component analysis (PCA) on the other hand, uses an orthogonality constraint to maximize the amount of variance captured in each principal components (PC). Thus, it is often informative to plot samples projected into the first two PCs (right) to assess sample clustering.

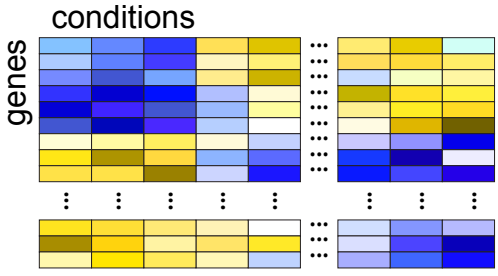
Raw RNA-Seq Reads



Alignment & quantification



Normalization & log-transformation



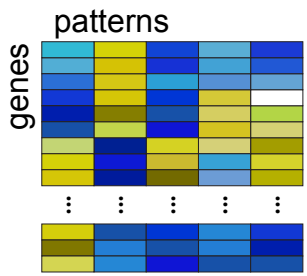
Matrix Factorization

PCA
- Maximal separation of (orthogonal) signal

ICA
- Independent signals

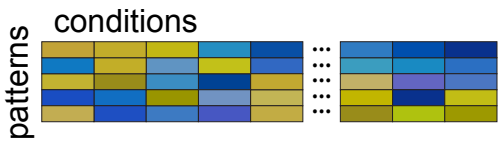
NMF
- Dependent signals
- Nonnegative values

Amplitude Matrix
(molecular relationships)

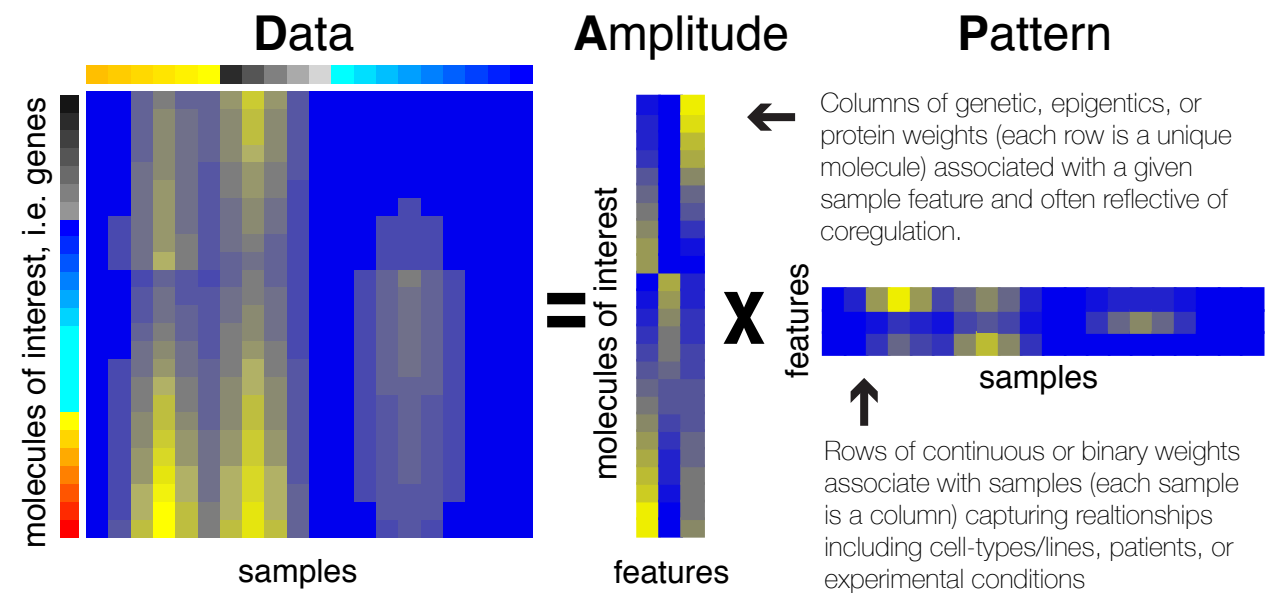


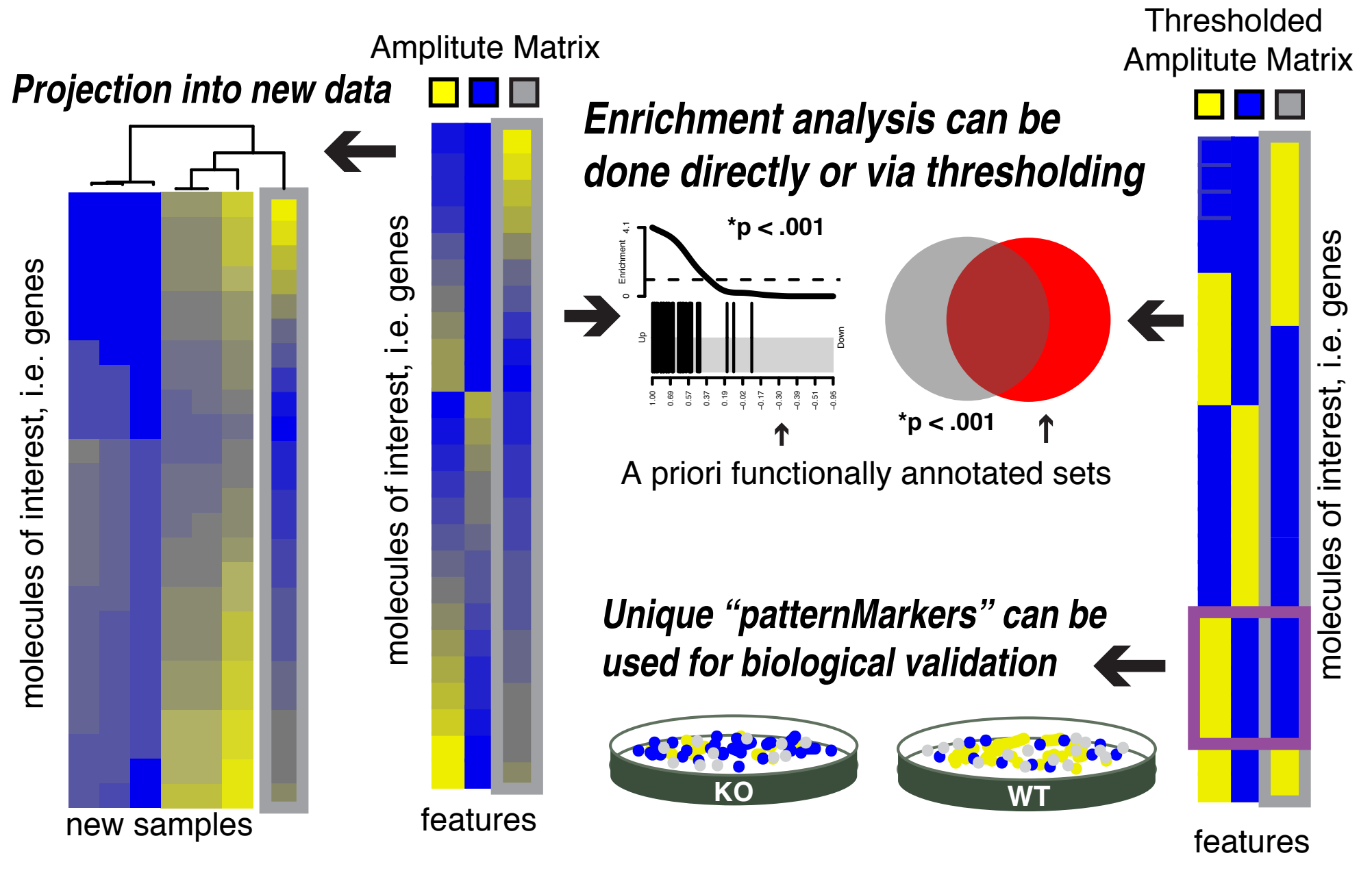
- Gene set discovery
- Pathway analysis
- Biomarker discovery

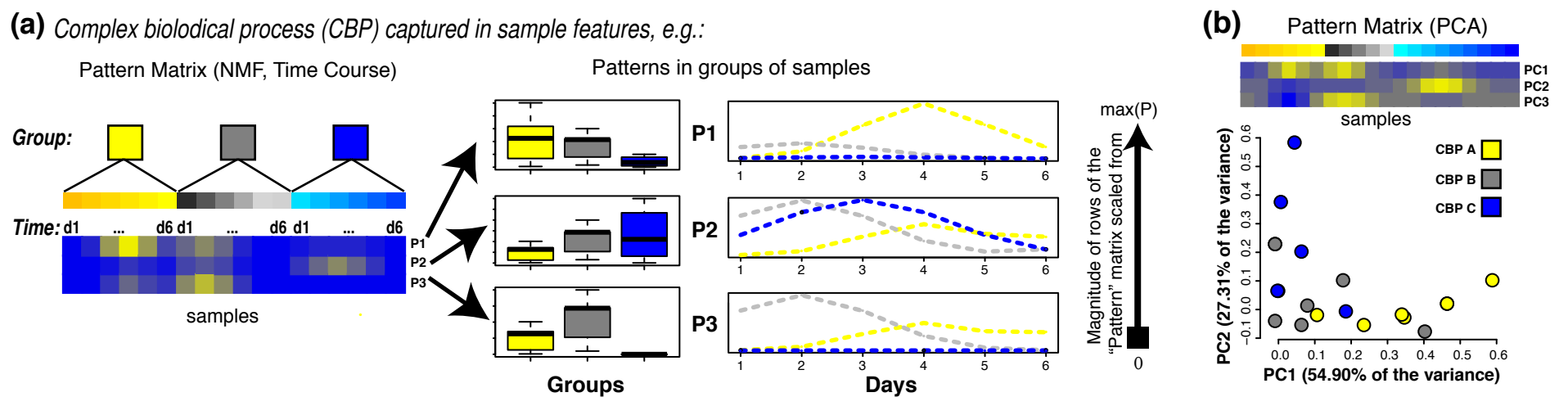
Pattern Matrix
(sample relationships)



- Clustering analysis
- Subtype / subclone discovery
- Timecourse analysis


Data
Amplitude
Pattern
molecules of interest, i.e. genes
samples
=
molecules of interest
features
X
Columns of genetic, epigentics, or protein weights (each row is a unique molecule) associated with a given sample feature and often reflective of coregulation.
features
samples
Rows of continuous or binary weights associate with samples (each sample is a column) capturing realtionships including cell-types/lines, patients, or experimental conditions


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


(a) Complex biolodical process (CBP) captured in sample features, e.g.:
Pattern Matrix (NMF, Time Course)
Group:
Time: d1 ... d6 d1 ... d6 d1 ... d6
samples
Patterns in groups of samples
P1
P2
P3
Groups
Days
max(P)
Magnitude of rows of the "Pattern" matrix scaled from
0
(b)
Pattern Matrix (PCA)
PC1
PC2
PC3
samples
CBP A
CBP B
CBP C
PC2 (27.31% of the variance)
PC1 (54.90% of the variance)