HHS Public Access

Author manuscript

Trends Genet. Author manuscript; available in PMC 2019 October 01.

Published in final edited form as:

Trends Genet. 2018 October ; 34(10): 790–805. doi:10.1016/j.tig.2018.07.003.

Enter the Matrix: Factorization Uncovers Knowledge from Omics

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], and **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, USA

[2]Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, USA

[3]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA

[4]Department of Computer Science, Institute for Data Intensive Engineering and Science, Johns Hopkins University, Baltimore, MD, USA

[5]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, USA

[6]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, USA

[7]Vavilov Institute of General Genetics, Moscow, Russia

[8]University of Hawaii Cancer Center, Honolulu, HI, USA

[9]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, PA, USA

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

[11]Digital Technologies Research Centre, National Research Council of Canada, Ottawa, ON, Canada

[12]School of Computer Science, University of Windsor, Windsor, ON, Canada

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

[14]Department of Applied Mathematics and Statistics, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD, USA

Abstract

Omics data contain signals from the molecular, physical, and kinetic inter- and intracellular interactions that control biological systems. Matrix factorization (MF) techniquescan reveal low-dimensional structure from high-dimensional data that reflect these interactions. These techniques

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

Supplemental Information

Supplemental information associated with this article can be found, in the online version, at https://doi.org/10.1016/j.tig.2018.07.003.

can uncover new biological knowledge from diverse high-throughput omics data in applications ranging from pathway discovery to timecourse analysis. We review exemplary applications of MF for systems-level analyses. We discuss appropriate applications of these methods, their limitations, and focus on the analysis of results to facilitate optimal biological interpretation. The inference of biologically relevant features with MF 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 have 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; see Glossary) [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 (Figure 1). Columns corresponding to experimental replicates, or samples with similar phenotypes, will have values from the same distribution of biological variation. Related structures in the data are 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 computational techniques for **unsupervised learning** to discover CBPs 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 CBPs 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 that provide a set of principled approaches to parsimoniously reveal the low-dimensional structure while preserving as much information as possible from the original data. MF is also referred to as matrix decomposition,and the corresponding inference problem as deconvolution. Other reviews discuss the mathematical and technical details of MF techniques [5–8] and their applications to microarray data [9]. We focus here 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 from bulk and single-cell data.

Workflow for MF Analysis

After data preprocessing, most high-throughput molecular datasets can be represented as a matrix in which each element contains the measurement of a single molecule in a single experimental condition. In the example of RNA sequencing (RNA-seq), the number of short reads from each gene are summarized into gene level counts. The resulting high-dimensional

dataset is formulated by representing these gene level counts for each sample as a column in the data matrix (Figure 1). MF methods can then be applied to these count matrices to learn CBPs from the data. Many MF techniques described are for RNA-seq data that are preprocessed using log transformation [10] or models of sequencing depth [11], while others directly model read counts [12].

Most examples featured in this review are based on such preprocessed RNA-seq analysis of gene counts or log-transformed gene counts. We note that applications of MF are not limited to this data modality or this preprocessing pipeline. For example, MF has also been applied to define mutational signatures in cancer [13,14], allele combinations in phenotypes [15], transcript regulation of genes [16], and distributions of transcript lengths [17], as well as to discriminate peptides in mass spectrometry proteomics [18]. To apply MF to other data modalities, they must also be properly preprocessed into a data matrix with a distribution appropriate to the MF analysis method.

When applied to high-throughput omics data, MF techniques learn two matrices: one describes the structure between features (e.g., genes) and another describes the structure between samples (Figure 2, Key Figure). We call the former feature-level matrix the **amplitude matrix** and the latter sample-level matrix the **pattern matrix**. 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 (Box 1). The values in each column of the amplitude matrix are continuous weights describing the relative contribution of a molecule in each inferred factor. In cases where factors distinguish between 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.

Whereas each column of the amplitude matrix describes the relative contributions of molecules to a factor, each row of the pattern matrix describes the relative contributions of samples to a factor (Figure 2). Sample groups can be learned by comparing the relative weights in each row of the pattern matrix. The pattern matrix from MF can also be binarized to perform clustering [19,20] or kept as continuous values to define relationships between samples [21–23]. In the same way 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.

The optimal number of columns of the amplitude matrix and rows of the pattern matrix is often referred to as the dimensionality of the MF problem, and learning this value remains an open problem for the MF research community [24,25]. We also note that MF is not a single computational method. Instead, there is a wide body of literature on numerous MF techniques that have been applied throughout computational biology. The properties of both the amplitude and pattern matrices, and subsequently the interpretation of their values, depend crucially on the specific MF problem and the algorithm selected for analysis.

MF Techniques: PCA, ICA, and NMF

There are numerous approaches to MF. 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 (as described in the supplemental information online and in other reviews [5,8,26–29]).

Briefly, PCA finds dominant sources of variation in high-dimensional datasets, inferring genes that distinguish between samples. Maximizing the variability captured in specific 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 complicate interpretation of the amplitude matrix for defining data-driven gene sets or the inference of specific CBPs.

ICA and NMF learn distinct processes from an input data matrix using different techniques. ICA learns factors that are statistically independent, resulting in more accurate association with literature-derived gene sets [30–32]. NMF methods constrain all elements of the amplitude and pattern matrices to be greater than or equal to zero [33,34]. Whereas the features in PCA can be ranked by the extent to which they explain the variation in the data, the features in both ICA and NMF are assumed to have equal weight. 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 [35].

The solutions from both ICA and NMF may vary depending upon the initialization of the algorithm, leading to disparate amplitude and pattern matrices. Therefore, it is crucial to ensure that particular solution used for analysis provides an optimal and robust solution before using the results of the factorization to interpret CBPs. We previously found that Bayesian techniques to solve NMF have more robust amplitude matrices than gradient-based techniques, and thus generate more accurate associations between the values in the amplitude matrix and functional pathways [5,36]. Other studies have found that gradient-based approaches have similar computational performance to their Bayesian counterparts [37], and new techniques are being developed to enhance the stability of the factorization [38]. Further integrated computational/ experimental investigation is necessary to assess the biological relevance of solutions from both classes of techniques and their robustness. These associations also depend crucially on the input data. Therefore, to learn CBPs from data, MF must be applied to datasets with sufficient measurements of the experimental perturbations or conditions relevant to the specific biological problem being addressed.

Sample Application to Genotype-Tissue Expression (GTEx) Project Gene Expression Data of Postmortem Tissues

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. Selecting which MF method

to use to learn relevant CBPs from a dataset is then crucial, and developing standardized metrics to choose between them is an active area of research in computational biology. To illustrate the differences between MF methods, we apply PCA, ICA (CRAN package fastICA [39]), and NMF (R/Bioconductor package CoGAPS [40,41]) to a single dataset from postmortem samples in the GTEx project [42] (Figure 3). Specifically, we select a subset of GTEx data containing 12 brain tissues in the GTEx data for seven individuals for which we had previously performed only NMF analysis [41] (codes are provided in the supplemental information online). We select this problem because the CBPs (tissue and individual) are readily separated and are known a *priori,* providing a known ground truth to facilitate comparison of methods.

First we apply PCA to this sample GTEx dataset (Figure 3A). The components learned from PCA can be ranked by the amount of variation that they explain in the data, with the first two components explaining 89.6% of the variation in this dataset. The amount of variation explained by these components can be used to determine the optimal dimensionality for PCA [24]. A typical PCA analysis explores the association between these components and biological covariates in low-dimensional plots in which each axis is defined by the weights in one row of the pattern matrix (Figure 3A). Similarly to clustering analyses, these plots can be used to determine which biological features can be separated from the data. In this application, we observe that the cerebellum (light blue) and first cervical spinal cord (yellow) cluster separately from all other brain tissues (PC1 and PC2, or rows 1 and 2 of the pattern matrix, respectively). No separation between individuals is observed in this PCA analysis.

In contrast to PCA, the components of ICA and NMF cannot be ranked by percent variation explained. Instead, each row of the pattern matrix is an equally important CBP in the data. Therefore, these patterns are plotted independently relative to biological covariates, and not relative to one another as in PCA analyses. When applied to the sample GTEx dataset, rows of the pattern matrix from both ICA (Figure 3B) and NMF (Figure 3C) also distinguish the cerebellum, similarly to PCA. Whereas PC1 has large positive values for the cerebellum and large negative values for the other brain tissues, both NMF and ICA have large absolute value only for the tissue of interest and are near zero for other tissues. As a result, both ICA and NMF provide tissue-specific patterns and PCA provides tissue-segregating patterns. We note that the magnitude of all these patterns is unit-less, reflecting the relative weights of each sample in the pattern and not a measurable quantity. Because gene weights can be either positive or negative in ICA, the patterns weights can also. By contrast, by construction the NMF values are all non-negative. As a result, the pattern corresponding to the cerebellum samples from ICA is still specific to that tissue, with genes overexpressed in this tissue having negative weights in the corresponding column of the amplitude matrix, and genes underexpressed in that tissue having positive weights. Indeed, the gene weights in the column of the amplitude matrix corresponding to the NMF cerebellum pattern are significantly anti-correlated with the gene weights in the column of the amplitude matrix corresponding to the ICA cerebellum pattern ($R = -0.72$, $P < 2 \times 10^{-16}$).

In contrast to PCA, both ICA and NMF also infer patterns that distinguish individuals with common weights across all tissues. In the NMF analysis, each individual has a separate

pattern. The ICA analysis has a single pattern that has large positive values for one of the individuals distinguished in one NMF pattern, and large negative values for another individual distinguished in different NMF pattern. This discrepancy highlights the difference between inferring independent sources of variation with ICA and NMF. Specifically, ICA may combine multiple CBPs using common genes whose sign changes by experimental condition, whereas NMF will find CBPs that are additive, corresponding only to overexpression of genes in that condition.

PCA, ICA, and NMF 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. We further note that the specific techniques for PCA, ICA, and NMF selected for analysis in this example are only one of a multitude of variants of techniques which have been developed for computational biology. 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. However, the general properties of these methods will remain and be consistent with our example. Briefly, we observe in this example that PCA finds sources of separation in the data, whereas both ICA and NMF find independent sources of variation. ICA can find both over- and underexpression of genes in a single CBP, whereas NMF can find only overexpressed genes in a single CBP. As a result, ICA may better model both repression and activation than NMF, but as a side effect may have greater mixture of CBPs than NMF.

Regardless of the technique selected, the results will also be sensitive to the input data. For example, a different NMF-class algorithm called 'grade of membership' (GOM) was also applied to a larger set of postmortem samples in GTEx. 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 only to tissue samples from the brain. Thus, 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 that are essential for exploratory data analysis.

Further Example Applications To Represent Cell Types, Disease Subtypes, Population Stratification, Tissue Composition, and Tumor Clonality with the Pattern Matrix

Exactly as we observed in the GTEx example, 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 [43]. 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.

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 normal 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 [44]. The hierarchical relationship between patterns has been used to assess the robustness of patterns to quantify the optimality of the factorization [45] and learn the optimal dimensionality of the factorization [46]. Other algorithms use statistical metrics to estimate the number of factors [12,47]. 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).

Moreover, application of different MF algorithms to the same dataset can give different sample groupings that reflect biology. For example, in population genetics a grouping inferred from GWAS which distinguishes ancestry is equally valid to a grouping inferred from the same GWAS data which distinguishes disease risk. The application of PCA to SNP data from 3000 European individuals [48] demonstrates inference of sample corelationships using the pattern matrix, and found that much of the variation in DNA sequence is explained by the longitude and latitude of the country of origin of an individual. In addition, statistical models can be formulated assuming that the inheritance of an individual arises from proportions of ancestry in distinct populations through genetic admixture [48]. An MF-based technique called sparse factor analysis also distinguishes between these populations using GWAS data [49]. These analyses demonstrate that the ancestry of each individual is a dominant source of variation in DNA sequence. At the same time, 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 [50]. In the same way as we observed in our GTEx example, the application of multiple MF techniques is essential to determine each source of variation in GWAS studies.

Mixtures of cell types in biological samples 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 sample by applying MF to genes whose expression is uniquely associated with each cell type [51]. Subsetting the data to different genes may give rise to different factors that represent different CBPs. Nonetheless, CoGAPS NMF 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 geneset in the expression data [41,52]. These results suggest that the dependency of an MF on the specific genes used for analysis may depend on the heterogeneity of the signal in the data matrix.

Cellular and molecular heterogeneity poses a particular challenge to MF analysis in cancer genomics. Even a pure tumor tissue can contain numerous subclones owing 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

[47,53–56]. 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 [12,47]. 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 Timecourse Analysis

Entwined in the challenge of decomposing cell types and subpopulations is the fact that CBPs change over time. High-throughput timecourse datasets are emerging in the literature to account for the dynamics of biological systems. The central goal of timecourse 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 timecourse analysis [57,58]. MF analyses can naturally infer changes in CBPs over time when applied to timecourse data because the continuous weights for each sample in the pattern matrix can vary among samples collected across distinct timepoints. The relative weights of rows of the pattern matrix can encode the timing of regulatory dynamics directly from the data (Figure 4). Nonetheless, most MF algorithms for timecourse analysis do not encode the known timepoints or retain their relative ordering. Methods that specifically use these temporal data are currently an active area of research.

Both ICA and NMF were found to have signatures characterizing the yeast cell cycle and metabolism in early timecourse microarray experiments [59,60]. The sparse NMF techniques using Bayesian methods had patterns that reflected the smooth dynamics of these phases [36,59]. This approach has been shown to simultaneously learn pathway inhibition and transitory responses to chemical perturbation of cancer cells [61] and relate the changes in phospho-proteomic trajectories between multiple therapies [62]. Similar analysis of healthy brain tissues learned the dynamics of transcriptional alterations that are common to the aging process in multiple individuals [52]. 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, thereby elucidating the molecular mechanisms that give rise to therapeutic resistance and metastasis. Even if there are the same number of biological features, the rate or timing of related features in different molecular modalities may be offset [63]. These discrepancies by data modality suggest that different regulatory mechanisms may be responsible for initiating and stabilizing the malignant phenotype [63].

Data-Driven Gene Sets from MF Provide Context-Dependent Coregulated Gene Modules and Pathway Annotations

Genomic data are often interpreted by identifying molecular changes in sets of genes annotated to functionally related modules or pathways, called gene sets [64,65]. Often the associations between gene sets and functions are based upon manual curation of the literature [66,67]. Such set-level interpretations often lack important contextual information [64,68,69], and cannot describe genes of 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 (Figure 5). 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 [70,71]. Other studies also integrate the literature-derived gene signatures in these thresholds to refine the context of pathway databases [36,72]. The genes derived from these binarizations can be used as inputs to pathway analyses from differential expression statistics in independent datasets (Figure 5, right), and are analogous to the hierarchical clustering-based gene modules [73] and gene expression signatures from public domain studies in the MSigDB gene-set database [74]. 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 [41,75]. Selecting genes based upon these statistics can facilitate visualization of the CBPs in high-dimensional data [41]. 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 [41] (Figure 5, 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 [76]. If the expression level of a gene 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 the findings. Further, continuous signatures can be associated directly with other samples using projection methods [76,77] or profile correspondence methods [78].

MF Enables Unbiased Exploration of Single-Cell Data for Phenotypes and Molecular Processes

MFapproaches are a natural choice in single-cell RNA-seq (scRNA-seq) data analysis owing to the high dimensionality of the data, and are used to identify and remove batch effects, summarize CBPs, and annotate cell types in the data [79–82]. Whereas MF analysis of 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 [75,83]. 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 [84]. Biological knowledge can be encoded with a class of MF algorithms that summarize factors using gene sets [79,85,86].

Most MF techniques developed for bulk omics data assume that the gene expression changes from CBPs are additive. This assumption is violated in scRNA-seq data. One reason for the

violation of the additive assumption in MF is the inability to distinguish true zeros from missing values. Imputation methods for preprocessing [81,87] or newer MF algorithms that model missing data are essential for scRNA-seq data. Branching of trajectories of cellular states and lack of cell-cycle synchronization in scRNA-seq data further violate the additive assumption in MF. New nonlinear factorization techniques are being developed to enhance visualization of trajectory structures in single-cell data [80,88–90] in these cases. However, the results from these methods cannot be interpreted in the same way as those from MF algorithms. In particular, the low-dimensional solutions from these methods are not necessarily useable to reconstruct the original high-dimensional data.

Concluding Remarks

MF encompasses 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 [59,91–93], 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 [94–99], clustering and metric learning [100–102], latent Dirichlet allocation [103], subspace learning [104–109], multiview learning [110], matrix completion [111], multitask learning [112], semi-supervised learning [113], compressed sensing [114], and similarity-based learning [115,116]. 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. As the size of these datasets grow, it is crucial to develop new algorithms to solve MF problems that scale with the ever-increasing size of omics data [37,41,117,118]. MF algorithms can also be extended for simultaneous analysis of data from multiple data modalities, enabling genomic data integration [7,8,119]. Techniques that extend this integrated MF frame-work, including Bayesian group factor analysis [8] and tensor decomposition [120–123], can also analyze datasets across different molecular levels [124]. Developing such data-integration techniques is an active area of research in both genomics and computational sciences.

Different classes of techniques solve MF, including gradient-based and probabilistic methods (supplemental information online). Distinct MF problems each aim to identify specific 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 both the dataset and the question in selecting exactly the right technique for that question. Most such comparisons in the literature have been made by investigators who are developing MF methods. Unbiased assessments of the relative performance of different MF algorithms for different exploratory data analysis problems are essential to determine the relative strengths and weaknesses of each method for distinct biological problems. MF algorithms can be further tailored to the biological problem of interest using methods that also encode prior biological knowledge of the system underlying the measured dataset [125–127].

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 is essential. Advances in MF and related techniques will be essential for powering systems-level analyses from big data (see Outstanding Questions).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Orly Alter, David Berman, 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 the National Institutes of Health (NIH) National Cancer institute (NCI) and the National Libary of Medicine (NLM) (grants NCI 2P30CA006516–52 and 2P50CA101942–11 to A.C.C., NCI R01CA177669 E.J.F., NCI U01CA212007 to E.J.F., NLM R01LM011000 to M.F.O., and NCI P30 CA006973), Johns Hopkins University Catalyst and Discovery Awards to E.J.F., a Johns Hopkins University Institute for Data Intensive Engineering and Science (IDIES) Award to E.J.F. and R.A., a 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 C.S.G., Alex's Lemonade Stand Foundation's Childhood Cancer Data Lab (C.S.G.), award K01ES025434 from the National Institute of Environmental Health Sciences (NIEHS) through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (L.X.G.), award P20 COBRE GM103457 from the NIH/National Institute of General Medical Sciences (NIGMS; to L.X.G.), R01 LM012373 from the NLM (L.X.G.), R01 HD084633 from the National Institute of Child Health and Human Development (NICHD; to L.X.G.), the Department of Defense Breast Cancer Research Program (BCRP; award BC140682P1, A.C.C.), the National Science and Engineering Council of Canada (NSERC; DG grant number RGPIN-2016–05017, A.N.), the Windsor-Essex County Cancer Centre Foundation (Seeds4Hope grant 814221, A.N.), Hopkins inHealth and Booz Allen Hamilton (90056858) to Y.X., the Russian Foundation for Basic Research (KOMFI 17–00-00208) and NIH (NCI P30 CA006973) to A.V.F., and the National Research Council of Canada to Y.L. This project has been made possible in part by grants 2018–183444 (E.J.F.), 2018–128827 (L.A.G.), and 2018–182718 (C.S.G.) from the Chan Zuckerberg Initiative Donor-Advised Fund (DAF), an advised fund of the Silicon Valley Community Foundation. The views and opinions of, and endorsements by, the author(s) do not reflect those of the US Army or the Department of Defense.

Glossary

Amplitude matrix
the matrix learned from MF that contains molecules in rows and factors in columns. Each column represents the relative contributions of the genes to a factor, and these 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, for example the cell types in a tissue sample.

Independent component analysis (ICA)
an MF technique that learns statistically independent factors.

Matrix factorization (MF)

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

Non-negative matrix factorization (NMF)
an 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 contributions of the samples to a factor, and these can be used to define the relative activity of CBPs in each sample.

Principal component analysis (PCA)
an MF technique that learns orthogonal factors ordered by the relative amount of variation of the data that they explain.

Unsupervised learning
computational techniques to discover features from high-dimensional data, without reliance on prior knowledge of low-dimensional covariates.

References

1. Bell G et al. (2009) Beyond the data deluge. Science 323, 1297–1298 [PubMed: 19265007]
2. Sagoff M (2012) Data deluge and the human microbiome project. Issues Sci. Technol 28 http://issues.org/28-4/sagoff-3/
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 [PubMed: 17060616]
4. Heyn P et al. (2015) Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences. Bioessays 37, 148–154 [PubMed: 25400101]
5. Ochs MF and Fertig EJ (2012) Matrix factorization for transcriptional regulatory network inference. IEEE Symp. Comput. Intell. Bioinforma. Comput. Biol. Proc 2012, 387–396 [PubMed: 25364782]
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 [PubMed: 26969681]
8. Li Y et al. (2016) A review on machine learning principles for multi-view biological data integration. Brief. Bioinform 19, 325–340
9. Devarajan K (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLoS Com-put. Biol 4, e1000029
10. Ritchie ME et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43, e47 [PubMed: 25605792]
11. Anders S and Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11, R106 [PubMed: 20979621]
12. Xie F et al. (2017) BayCount: a Bayesian decomposition method for inferring tumor heterogeneity using RNA-Seq counts. *bioRxiv* Published online November 13, 2017. 10.1101/218511
13. Alexandrov LB et al. (2013) Signatures of mutational processes in human cancer. Nature 500, 415–421 [PubMed: 23945592]
14. Alexandrov LB et al. (2016) Mutational signatures associated with tobacco smoking in human cancer. Science 354, 618–622 [PubMed: 27811275]

15. Favorov AV 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 [PubMed: 16118183]
16. Zakeri M et al. (2017) Improved data-driven likelihood factorizations for transcript abundance estimation. Bioinformatics 33, i142–i151 [PubMed: 28881996]
17. Bertagnolli NM 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 [PubMed: 24282503]
18. Peckner R et al. (2017) Specter: linear deconvolution as a new paradigm for targeted analysis of data-independent acquisition mass spectrometry proteomics. *bioRxiv* Published online September 8, 2017. 10.1101/152744
19. Yeung KY et al. (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics 17,977–987 [PubMed: 11673243]
20. Jiang D et al. (2004) Cluster analysis for gene expression data: a survey. IEEE Trans. Knowl. Data Eng 16, 1370–1386
21. Venet D et al. (2001) Separation of samples into their constituents using gene expression data. Bioinformatics 17, S279–S287 [PubMed: 11473019]
22. Abbas AR et al. (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS One 4, e6098 [PubMed: 19568420]
23. Erkkilä T et al. (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues. Bioinformatics 26, 2571–2577 [PubMed: 20631160]
24. Leek JT (2011) Asymptotic conditional singular value decomposition for high-dimensional genomic data. Biometrics 67, 344–352 [PubMed: 20560929]
25. Kelton CJ. The estimation of dimensionality in gene expression data using nonnegative matrix factorization; 2015 IEEE Internationa/ Conference on Bioinformatics and Biomedicine (BIBM; 2015. 1642–1649.
26. Meng C et al. (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Brief. Bioinform 17, 628–641 [PubMed: 26969681]
27. Berry MW et al. (2007) Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal 52, 155–173
28. Wang Y-X and Zhang Y-J (2013) Nonnegative matrix factorization: a comprehensive review. IEEE Trans. Knowl. Data Eng 25, 1336–1353
29. Zhou G et al. (2014) Nonnegative matrix and tensor factorizations: an algorithmic perspective. IEEE Signal Process. Mag 31, 54–65
30. Lee S-I and Batzoglou S (2003) Application of independent component analysis to microarrays. Genome Biol. 4, R76 [PubMed: 14611662]
31. Engreitz JM et al. (2010) Independent component analysis: mining microarray data for fundamental human gene expression modules. J. Biomed. Bioinf 43, 932–944
32. Teschendorff AE et al. (2007) Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput. Biol 3, e161 [PubMed: 17708679]
33. Ochs MF et al. (1999) A new method for spectral decomposition using a bilinear Bayesian approach. J. Magn. Reson 137, 161–176 [PubMed: 10053145]
34. Lee DD and Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 [PubMed: 10548103]
35. Moloshok TD et al. (2002) Application of Bayesian decomposition for analysing microarray data. Bioinformatics 18,566–575 [PubMed: 12016054]
36. Kossenkov AV et al. (2007) Determining transcription factor activity from microarray data using Bayesian Markov chain Monte Carlo sampling. Stud. Health Technol. Inform 129, 1250–1254 [PubMed: 17911915]
37. Mairal J et al. (2010) Online learning for matrix factorization and sparse coding. J. Mach. Learn. Res 11, 19–60

38. Wu S et al. (2016) Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks. Proc. Natl. Acad. Sci. U. S. A 113, 4290–4295 [PubMed: 27071099]
39. Hyvärinen A and Oja E (2000) Independent component analysis: algorithms and applications. Neural Netw. 13, 411–430 [PubMed: 10946390]
40. Fertig EJ et al. (2010) CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. Bioinformatics 26, 2792–2793 [PubMed: 20810601]
41. Stein-O'Brien GL et al. (2017) PatternMarkers & GWCoGAPS for novel data-driven biomarkers via whole transcriptome NMF. Bioinformatics 33, 1892–1894 [PubMed: 28174896]
42. Dey KK et al. (2017) Visualizing the structure of RNA-seq expression data using grade of membership models. PLoS Genet. 13, e1006599 [PubMed: 28333934]
43. 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 [PubMed: 25456126]
44. Fertig EJ 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 [PubMed: 24223768]
45. Bidaut G et al. (2010) Interpreting and comparing clustering experiments through graph visualization and ontology statistical enrichment with the ClutrFree Package In Biomedical Informatics for Cancer Research (Ochs MF, ed.), pp. 315–333, Springer
46. Bidaut G et al. (2006) Determination of strongly overlapping signaling activity from microarray data. BMC Bioinformatics 7, 99 [PubMed: 16507110]
47. Xu Y et al. (2015) MAD Bayes for tumor heterogeneity - feature allocation with exponential family sampling. J. Am. Stat. Assoc 110, 503–514 [PubMed: 26170513]
48. Novembre J et al. (2008) Genes mirror geography within Europe. Nature 456, 98–101 [PubMed: 18758442]
49. Engelhardt BE and Stephens M (2010) Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS Genet. 6, e1001117 [PubMed: 20862358]
50. McCarthy MI et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet 9, 356–369 [PubMed: 18398418]
51. Hackl H et al. (2016) Computational genomics tools for dissecting tumour-immune cell interactions. Nat. Rev. Genet 17, 441–458 [PubMed: 27376489]
52. Fertig EJ et al. (2014) Pattern identification in time-course gene expression data with the CoGAPS matrix factorization. Methods Mol. Biol 1101, 87–112 [PubMed: 24233779]
53. Nik-Zainal S et al. (2012) The life history of 21 breast cancers. Cell 149, 994–1007 [PubMed: 22608083]
54. Roth A et al. (2014) PyClone: statistical inference of clonal population structure in cancer. Nat. Methods 11, 396–398 [PubMed: 24633410]
55. Deshwar AG et al. (2015) PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 16, 35 [PubMed: 25786235]
56. 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
57. Bar-Joseph Z et al. (2012) Studying and modelling dynamic biological processes using time-series gene expression data. Nat. Rev. Genet 13, 552–564 [PubMed: 22805708]
58. Liang Y and Kelemen A (2017) Dynamic modeling and network approaches for omics time course data:overview of computational approaches and applications. Brief. Bioinform. Published online April 18, 2017. 10.1093/bib/bbx036
59. Moloshok TD et al. (2002) Application of Bayesian decomposition for analysing microarray data. Bioinformatics 18, 566–575 [PubMed: 12016054]
60. Liebermeister W (2002) Linear modes of gene expression determined by independent component analysis. Bioinformatics 18, 51–60 [PubMed: 11836211]

61. Ochs MF et al. (2009) Detection of treatment-Induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. Cancer Res. 69, 9125–9132 [PubMed: 19903850]
62. Hill SM et al. (2016) Inferring causal molecular networks: empirical assessment through a community-based effort. Nat. Methods 13, 310–318 [PubMed: 26901648]
63. Stein-O'Brien G et al. (2017) Integrated time-course omics analysis distinguishes immediate therapeutic response from acquired resistance. *bioRxiv* Published online August 1, 2017. 10.1101/136564
64. Khatri P et al. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput. Biol 8, e1002375 [PubMed: 22383865]
65. Irizarry RA et al. (2009) Gene set enrichment analysis made simple. Stat. Methods Med. Res 18, 565–575 [PubMed: 20048385]
66. Bauer-Mehren A et al. (2009) Pathway databases and tools for their exploitation: benefits, current limitations and challenges. Mol. Syst Biol 5, 290 [PubMed: 19638971]
67. Tsui IFL et al. (2007) Public databases and software for the pathway analysis of cancer genomes. Cancer Inform. 3, 379–397 [PubMed: 19455256]
68. The GTEx Consortium (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 348, 648–660 [PubMed: 25954001]
69. Tan J et al. (2017) Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. Cell Syst. 5, 63–71 [PubMed: 28711280]
70. Kim JW et al. (2017) Decomposing oncogenic transcriptional signatures to generate maps of divergent cellular states. Cell Syst. 5, 105–18.e9 [PubMed: 28837809]
71. Fertig EJ 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 [PubMed: 22549044]
72. Fertig EJ et al. (2012) Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobioscience 12, 142–149
73. Segal E et al. (2004)A module map showing conditional activity of expression modules in cancer. Nat. Genet 36, 1090–1098 [PubMed: 15448693]
74. Subramanian A et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci 102, 15545–15550 [PubMed: 16199517]
75. Zhu X et al. (2017) Detecting heterogeneity in single-cell RNA-Seq data by non-negative matrix factorization. Peer J. 5, e2888 [PubMed: 28133571]
76. DeTomaso D and Yosef N (2016) FastProject: a tool for lowdimensional analysis of single-cell RNA-Seq data. BMC Bioinformatics 17, 315 [PubMed: 27553427]
77. Fertig EJ et al. (2013) Identifying context-specific transcription factor targets from prior knowledge and gene expression data. IEEE Trans. Nanobiosci 12, 142–149
78. Irizarry RA et al. (2005) Multiple-laboratory comparison of microarray platforms. Nat. Methods 2, 345–350 [PubMed: 15846361]
79. Fan J et al. (2016) Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat. Methods 3, 241–244
80. 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 [PubMed: 24658644]
81. Townes FW et al. (2017) Varying-censoring aware matrix factorization for single cell RNA-sequencing *bioRxiv* Published online July 21, 2017. 10.1101/166736
82. Moon KR et al. (2017) PHATE: a dimensionality reduction method for visualizing trajectory structures in high-dimensional biological data. bioRxiv Published online March 24,2017. 10.1101/120378
83. Puram SV et al. (2017) Single-vell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. Cell 171, 1611–1624 [PubMed: 29198524]

84. Hübschmann D et al. (2017) Deciphering programs of transcriptional regulation by combined deconvolution of multiple omics layers. *bioRxiv* Published online October 8, 2017. 10.1101/199547
85. Buettner F et al. (2017) f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq. Genome Biol. 18, 212 [PubMed: 29115968]
86. Buettner F et al. (2016) Scalable latent-factor models applied to single-cell RNA-seq data separate biological drivers from confounding effects. *bioRxiv* Published online November 15, 2016. 10.1101/087775
87. van Dijk D et al. (2017) MAGIC: a diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* Published online February 25, 2017. 10.1101/111591
88. Risso D et al. (2017) ZINB-WaVE: a general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv* Published online November 2, 2017. 10.1101/125112
89. Pierson E and Yau C (2015) ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 16, 241 [PubMed: 26527291]
90. van der ML and Hinton G (2008)Visualizing data using t-SNE. J. Mach. Learn Res 9, 2579–2605
91. 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 [PubMed: 10963673]
92. Fellenberg K et al. (2001) Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. U. S. A 98, 10781–10786 [PubMed: 11535808]
93. Brunet JP et al. (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. U. S. A 101, 4164–4169 [PubMed: 15016911]
94. Abdi H and Williams LJ (2010) Principal component analysis. WIREs Comp. Stat 2, 433–459
95. Hyvärinen A et al. (2004) Independent Component Analysis, John Wiley & Sons
96. Hardoon DR et al. (2004) Canonical correlation analysis: an overview with application to learning methods. Neural. Comput 16, 2639–2664 [PubMed: 15516276]
97. Schölkopf B et al. (1999) Kernel principal component analysis In Advances In Kernel Methods - Support Vector Learning (Schölkopf B, ed.), MIT Press, pp. 327–000
98. Arora R and Livescu K (2012) Kernel CCA for multi-view learning of acoustic features using articulatory measurements. In Symposium on Machine Learning in Speech and Language Processing. MSLP
99. Andrew G. Deep canonical correlation analysis; Proceedings of the 30th International Conference on Machine Learning; 2013. 1247–1255.
100. Ding C and He X (2004) K-means clustering via principal component analysis. Proceedings of the 21st International Conference on Machine Learning 29
101. Arora R et al. (2011) Clustering by left-stochastic matrix factorization. Proceedings of the 28th International Conference on Machine Learning 28, 761–768
102. Kulis B (2013) Metric learning: a survey. Found. Trends Mach. Learn 5, 287–364
103. Pritchard JK et al. (2000) Inference of population structure using multilocus genotype data. Genetics 155, 945–959 [PubMed: 10835412]
104. De la Torre F and Black MJ (2003) A framework for robust subspace learning. Int J. Comput. Vis 54, 117–142
105. Szeliski R (2010) Computer Vision: Algorithms and Applications. http://szeliski.org/Book/drafts/SzeliskiBook_20100903_draft.pdf
106. Candes EJ et al. (2011) Robust principal component analysis? J. ACM 58, 11
107. Arora R. Stochastic optimization for PCA and PLS; 50th Annual Allerton Conference on Communication Control, and Computing; Allerton: 2012. 861–868.
108. Arora R et al. (2013) Stochastic optimization of PCA with capped MSG In Advances in Neural Information Processing Systems (Vol. 26) (Burges CJC, ed.), In pp. 1815–1823, Curran Associates
109. Goes J et al. (2014) Robust stochastic principal component analysis In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (Kaski S and Corander J, eds), pp. 266–274, PMLR

110. Bickel S and Scheffer T (2004) Multi-view clustering. Proceedings of the IEEE International Conference on Data Mining 19–26
111. Candes EJ and Recht B (2009) Exact matrix completion via convex optimization. Found Comut. Math 9, 717
112. Argyriou A et al. (2007) Multi-task feature learning. Adv. Neural. Inf. Process. Syst 19, 41–48
113. Ando RK and Zhang T (2005) A framework for learning predictive structures from multiple tasks and unlabeled data. J. Mach. Learn. Res 6, 1817–1853
114. Cleary B (2017) Composite measurements and molecular compressed sensing for highly efficient transcriptomics. *bioRxiv* Published online January 2, 2017. 10.1101/091926
115. Aha DW et al. (1991) Instance-based learning algorithms. Mach. Learn 6, 37–66
116. Arora R et al. (2013) Similarity-based clustering by left-stochastic matrix factorization. J. Mach. Learn. Res 14, 1715–1746
117. Liao R et al. (2014) CloudNMF: a MapReduce implementation of nonnegative matrix factorization for large-scale biological datasets. Genomics Proteomics Bioinf. 12, 48–51
118. de Campos CP et al. (2013) Discovering subgroups of patients from DNA copy number data using NMF on compacted matrices. PLoS One 8, e79720 [PubMed: 24278162]
119. Huang S et al. (2017) More is better: recent progress in multi-omics data integration methods. Front. Genet 8, 84 [PubMed: 28670325]
120. Hore V et al. (2016) Tensor decomposition for multiple-tissue gene expression experiments. Nat. Genet 48, 1094–1100 [PubMed: 27479908]
121. Durham TJ et al. (2018) PREDICTD parallel epigenomics data Imputation with cloud-based tensor decomposition. Nat. Commun 9, 1402 [PubMed: 29643364]
122. Zhu Y et al. (2016) Constructing 3D interaction maps from 1D epigenomes. Nat. Commun 7, 10812 [PubMed: 26960733]
123. Wang M et al. (2017) Three-way clustering of multi-tissue multi-individual gene expression data using constrained tensor decomposition. *bioRxiv* Published online December 5, 2017. 10.1101/229245
124. Kolda T and Bader B (2009) Tensor decompositions and applications. SIAM Rev. 51, 455–500
125. Mao W et al. (2017) Pathway-level information extractor (PLIER): a generative model for gene expression data. bioRxiv Published online December 16 2017. 10.1101/116061
126. Hofree M et al. (2013) Network-based stratification of tumor mutations. Nat. Methods 10, 1108 [PubMed: 24037242]
127. Liao JC et al. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. U. S. A 100, 15522–15527 [PubMed: 14673099]

Box 1. Common Terminology in the Literature

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 interchangeablywith 'decomposition'. Other terms, such as 'features', 'components', 'latent variable', or 'latent factors' are used to refer to relationships between either molecular measurements or samples, depending on the context.

The specific terminology for the amplitude and pattern matrices also varies according to the method, and are preferably labeled with different variable names. In PCA, the amplitude matrix is often called the score or rotation matrix, and pattern matrix called the loadings. In ICA, the amplitude matrix is called the unmixing matrix and the pattern matrix is called the source matrix. In NMF, the amplitude matrix is commonly called the weights matrix and the pattern matrix is called the features matrix.

Throughout this paper we use 'amplitude matrix' to refer to the matrix that contains vectors which represent relationships between molecular measurements. Other terms used in the literature for these molecular relationships include modules, meta-pathways, and signatures. Similarly, we use '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, and controlling factors.

Outstanding Questions

How can the optimal number of factors be quantified? Increasing the number of factors in MF improves its approximation, but may cause overfitting. Computational metrics that balance these properties can be used to estimate the number of factors. Biologically driven metrics are also essential to assess whether this number represents the number of CBPs.

How can different factorizations be compared? Different MF algorithms identify different properties and CBPs from a single dataset. New techniques will be necessary to compare and merge the disparate but equally valid factorizations.

What techniques are best for efficient and optimal factorization? Large omics data-sets are common, particularly with single-cell technologies. Fast computational approaches to MF are required for the analysis of these data. These algorithms also require computational criteria to ensure that the solutions are robust.

How can fully nonlinear factorizations be performed on omics-scale data? Nonlinear extensions to MF have notable applications to single-cell data. Researchers often use low-dimensional representations of omics data learned from linear MF as inputs to nonlinear MF to account for both computational efficiency and the underlying mathematical assumptions of these nonlinear methods. Breakthroughs in theories for techniques such as kernel MF, deep MF, and manifold learning will enable fully nonlinear factorizations for large-scale single-cell data.

How can both gene regulatory relationships and distinct sources of technical variation be encoded in integrative analysis of data from different measurement technologies? Different measurement technologies yield data that follow unique distributions. CBPs and their timing also vary between types of measures. Integrative techniques must account for both the technical and biological sources of variation between disparate data modalities, including sparse or missing data. Such techniques are crucial to learning the regulatory relationships that drive CBPs and to modeling the multiscale nature of biological systems from omics data.

Highlights

MFs techniques infer low-dimensional structure from high-dimensional omics data to enable visualization and inference of complex biological processes (CBPs).

Different MFs applied to the same data will learn different factors. Exploratory data analysis should employ multiple MFs, whereas a specific biological question should employ a specific MF tailored to that problem.

MFs learn two sets of low-dimensional representations (in each matrix factor) from high-dimensional data: one defining molecular relationships (amplitude) and another defining sample-level relationships (pattern).

Data-driven functional pathways, biomarkers, and epistatic interactions can be learned from the amplitude matrix.

Clustering, subtype discovery, *in silico* microdissection, and timecourse analysis are all enabled by analysis of the pattern matrix.

MF enables both multi-omics analyses and analyses of single-cell data.

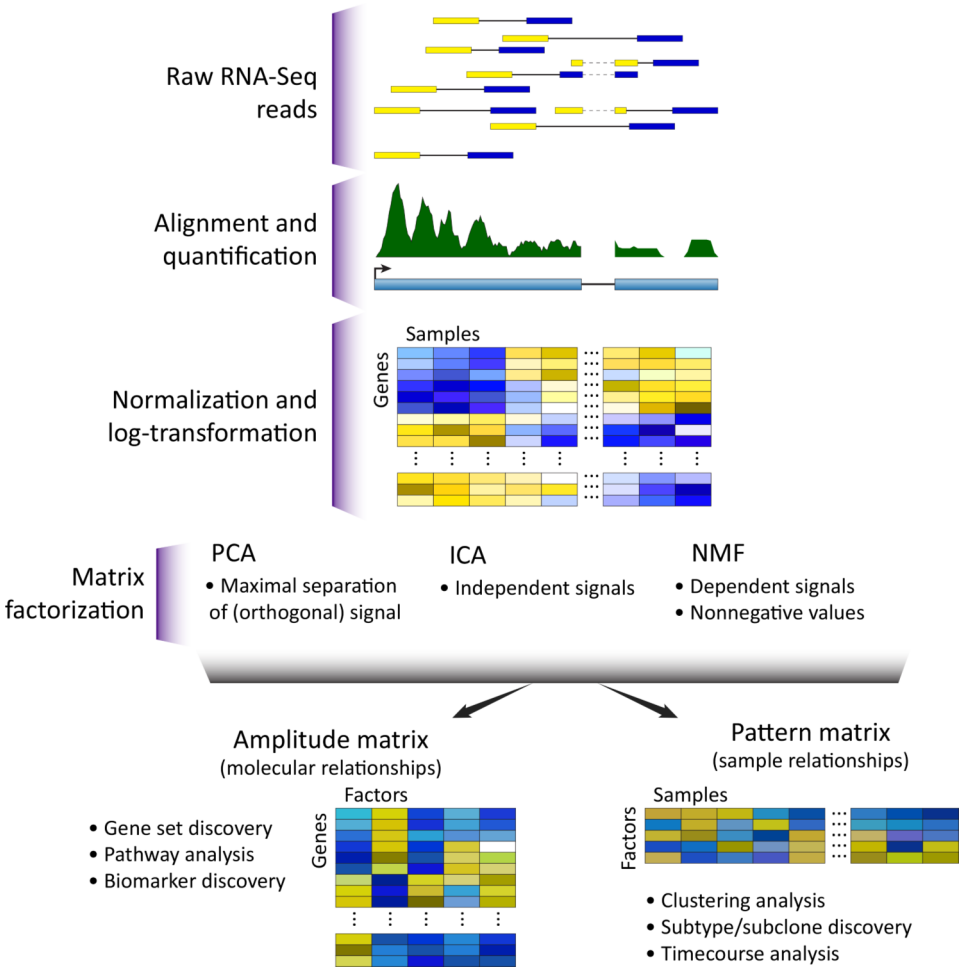

Figure 1. Omics Technologies Yield a Data Matrix That Can Be Interpreted through MF.

The data matrixfrom omics 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 is then input to a matrix factorization (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. Abbreviations: ICA, independent component analysis; NMF, non-negative matrix factorization; PCA, principal component analysis.

Key Figure

The Matrix Product of the Amplitude and Pattern Matrices Approximates the Preprocessed Input Data Matrix (A)

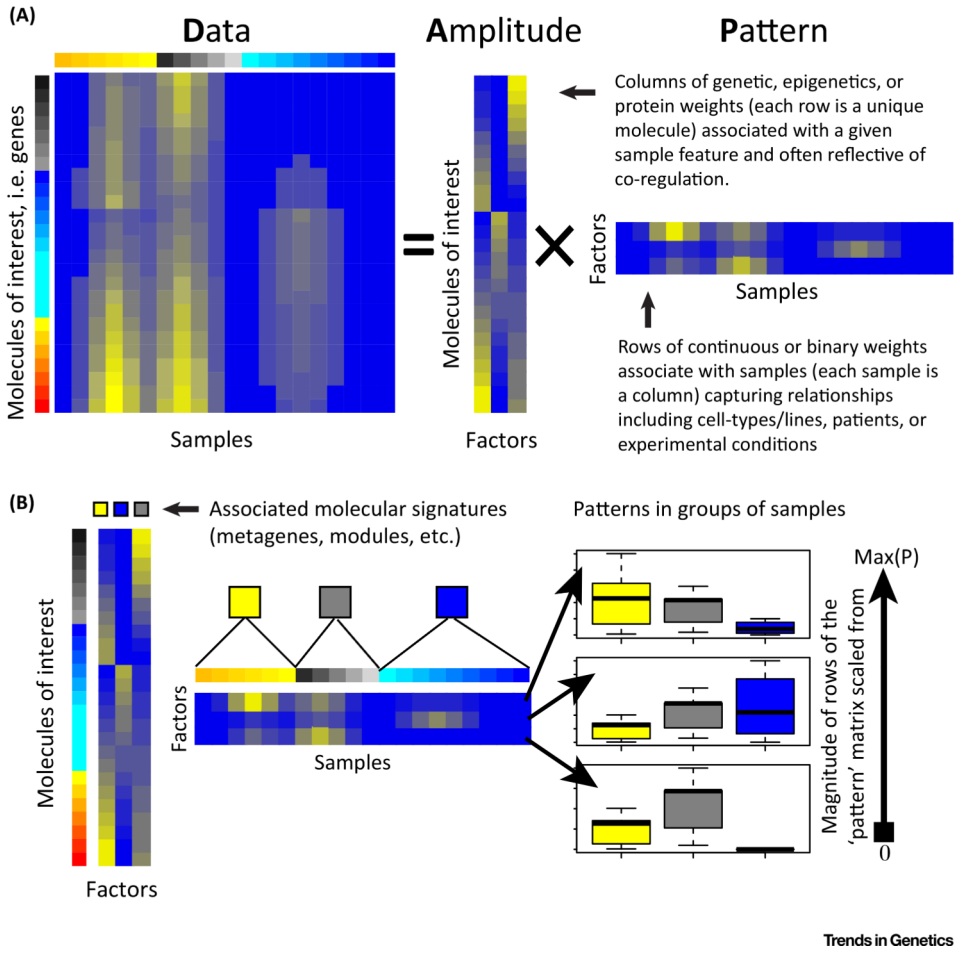

Figure 2.

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). (B) The values in the column of the amplitude matrix then represent the relative weights of each molecule in the CBP, and the values in the row of the pattern matrix represent its relative role in each sample. Plotting of the values of each pattern for a pre-determined sample grouping (here indicated by yellow, grey, and blue) in a boxplot as an example of a visualization technique for the pattern matrix. Abbreviation: Max(P), maximum value of each row of the pattern matrix.

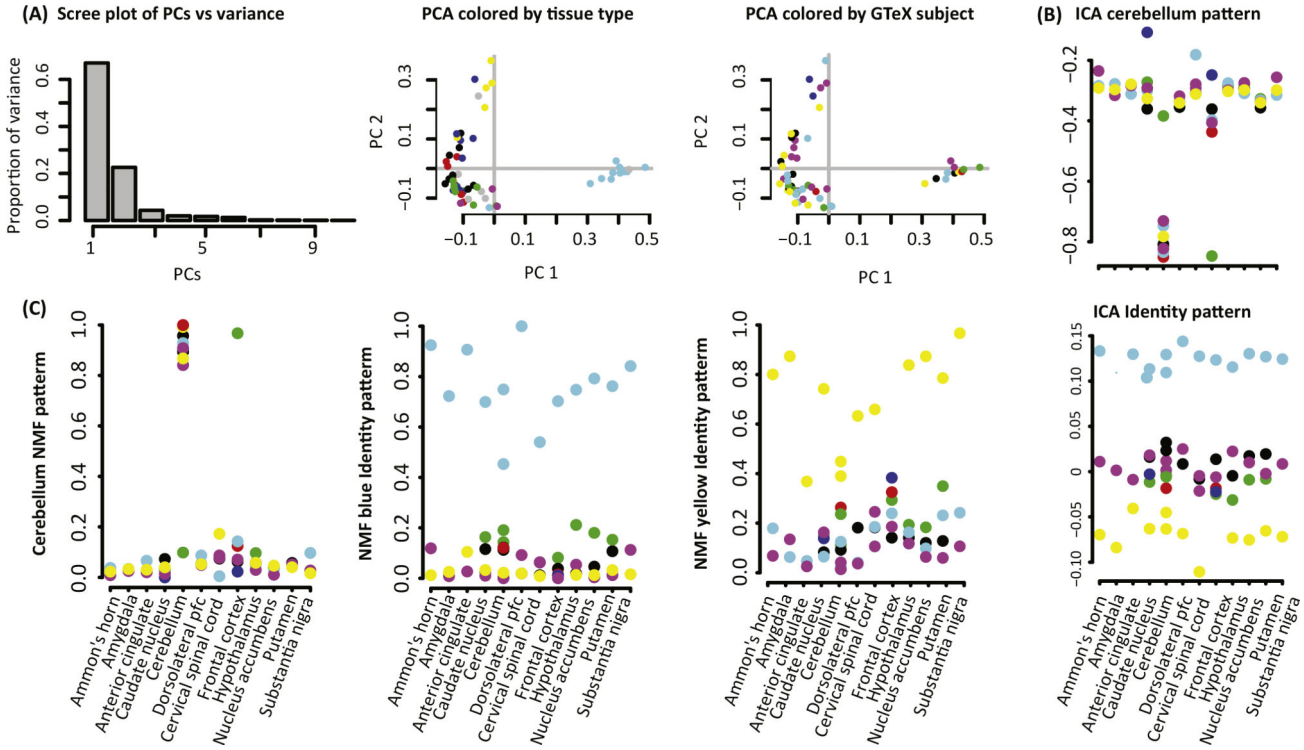

Figure 3. Comparison of Pattern Matrix From Matrix Factorization (MF) in Postmortem Tissue Samples from GTEx.

(A) PCA finds factors in rows of the pattern matrix that can be ranked by the amount of variation that they explain in the data, as illustrated in a scree plot. PCA analyses typically plot the first two principal components (PCs; rows of the pattern matrix) to assess sample clustering. Points are colored by tissue type annotations from GTEx (left), where Ammon's horn refers to the hippocampus, and donor (right). In GTEx data, the cerebellum (light blue) and first cervical spinal cord (yellow) cluster separately from all other brain tissues, but no separation between individuals is observed. (B) ICA finds factors associated with independent sources of variation, and therefore cannot be ranked in a scree plot. The relative absolute value of the magnitude of each element in the pattern matrix indicates the extent to which that sample contributes to the corresponding source of variation. The sign of the values indicate over- or underexpression in that factor depending on the sign of the corresponding gene weights in the amplitude matrix. As a result, the values can be plotted on the y axis against known covariates on the x axis to directly interpret the relationship between samples. When applied to GTEx, we observe one pattern associated with cerebellum, another pattern that has large positive values for one donor and large negative values for another donor, and eight other patterns associated with other sources ofvariation (supplemental information online). (C) NMF findsfactors that are both non-negative and not ranked by relative importance, similarly to ICA. The value of the pattern matrix indicates the extent to which each sample contributes to an inferred source of variation and is associated with overexpression of corresponding gene weights in the amplitude matrix. Values of the pattern matrix can be plotted similarly to ICA. When applied to GTEx, we observe one pattern associated with cerebellum, two more patterns associated with the two donors that were assigned to a single pattern in ICA, and seven other patterns associated with other

sources of variation (supplemental information online). Abbreviations: GTEx, Genotype-Tissue Expression (GTEx) project; ICA, independent component analysis; NMF, non-negative matrix factorization; PCA, principal component analysis.

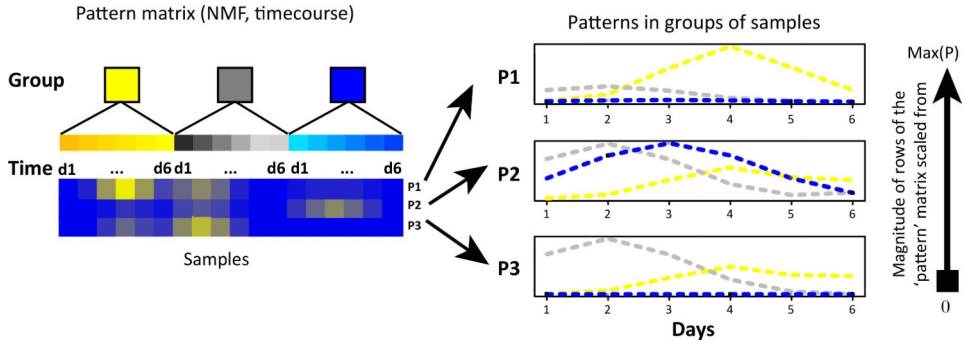

Figure 4. Samples Correspond to Timepoints; the Rows of the Pattern Matrix Can Be Plotted as a Function of Time and Sample Condition To Infer the Dynamics of Complex Biological Processes (CBPs).

Abbreviations: d1-d6, days 1–6; max(P), maximum value of each row of the pattern matrix; NMF, non-negative matrix factorization; P1–3, patterns 1–3.

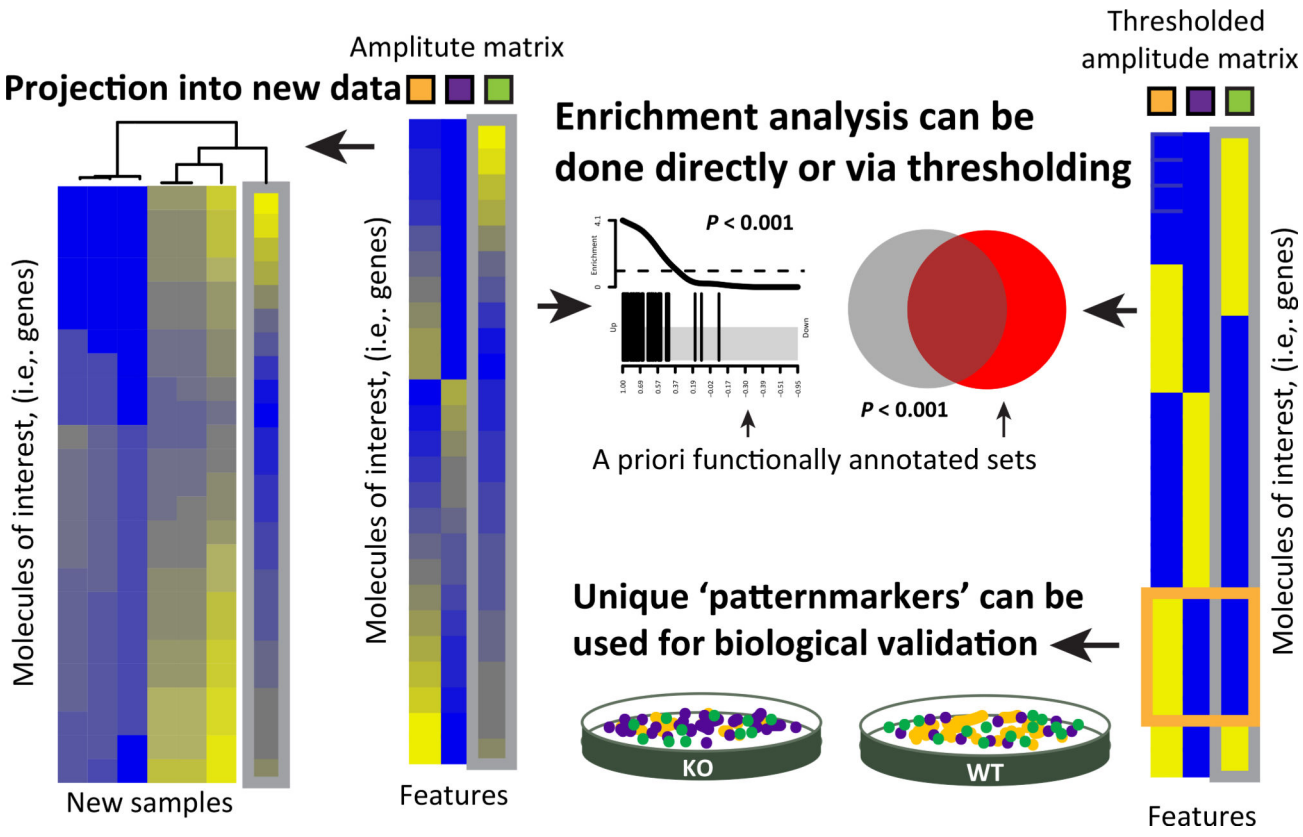

Figure 5. The Amplitude Matrix from Matrix Factorization (MF) Can Be Used to Derive Data-Driven Molecular Signatures Associated with a Complex Biological Process (CBP).

The columns of the amplitude matrix contain continuous weights describing the relative contribution of a molecule to a CBP (center panel; indicated by the orange, purple, and green boxes). 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. Abbreviations, KO, knockout; WT, wild type.