Multi-omics factor analysis disentangles patient heterogeneity in chronic lymphocytic leukaemia

Ricard Argelaguet1,\*, Britta Velten2,\*, Damien Arnol1, …, Sascha Dietrich, Thorsten Zenz, John C. Marioni1,3, Florian Buettner1, Wolfgang Huber2, Oliver Stegle1,2

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridge, UK
2. European Molecular Biology Laboratory (EMBL), Heidelberg, Germany.
3. Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge

\* Authors contributed equally and are sorted alphabetically.

Corresponding authors:

## Abstract

Complex phenotypes arise through interactions of multiple molecular layers, including the transcriptome, epigenome and proteome. Although numerous ‘omics’ are increasingly profiled in large patient cohorts, there is a lack of computational methods for integrating these different data modalities. Here, we present MultiOmics Factor Analysis (MOFA), a model for the unsupervised discovery of (latent) factors that drive variation in multi-omics data sets. Our model jointly infers factors that can either explain variability that is common to multiple ‘omics layers or specific to individual data modalities. We applied MOFA to a chronic lymphocytic leukemia (CLL) dataset comprising of 200 patients characterized by somatic mutation profiles, RNA sequencing, DNA methylation arrays as well as *ex-vivo* drug response profiling. MOFA recovered known axes of disease heterogeneity including immunoglobulin heavy chain variable region (IGHV) status or trisomy of chromosome 12, as well as previously underappreciated drivers of variation, some of which are predictive of clinical outcome or related to cancer-associated pathways such as oxidative stress and reactive oxygen species. Finally, we demonstrate that MOFA can be used to identify mislabeled samples and the model accurately imputes missing values.

# Introduction

Technological advances increasingly enable probing multiple biological layers in parallel, ranging from genomic sequencing, epigenome profiling to the transcriptome and the proteome (Hasin 2017).

The joint analysis of multiple molecular layers can deliver a more comprehensive view of relevant axes of molecular variability, each of which can manifest in individual or multiple data layers (Ritchie 2015). Such approaches are particularly appealing if the most relevant axes of variation between samples are not known *a priori*, and hence may be missed by studies that consider a single data modality.

*Note: There may be a point to made that the most interesting traits are ‘multi-omics’ traits that manifest in multiple views, i.e. following the central dogma of biology.*

Motivated by this, multi-omic profiling is increasingly applied across different domains, including studies to characterize molecular profiles across human cancers (e.g. TCGA 2017, Mertins 2016, Gerstung 2015, Iorio 2016), efforts to study cellular states in microbiology (e.g. Kim 2016) of in the identification of modulators of host-pathogen interactions (Söderholm, 2016). Most recently, it has become possible to also perform multi-omics analyses in single cells (Guo 2017, Angermueller 2016, Clark 2017, Macaulay 2015).

Multiple computational strategies for integrating different omics have been proposed (Ritchie 2015). One approach are local models to test for associations between different data modalities, for example discovering linkages between genetic variants and gene expression levels or other molecular layers (Add REF). While these strategies deliver insights into the coregulation of individual molecular features, they do not explain global phenotypic differences between samples. In contrast, global models use data across omics to detect global phenomena that generate sample to sample variation, for example using kernel- or graph-based modelling strategies (Lankriet 2004, Wang 2014), or by extending clustering to multiple data modalities (iCluster, Shen 2009, Mo 2012).

However, these existing strategies do not explicitly identify the biological and technical drivers that underpin variability in multi-omics studies. Moreover, these methods are hampered by computational scalability to larger datasets and they do not appropriately handle missing values as well as non-Gaussian data modalities such as binary readouts or count-based traits Finally, there are open challenges to define models that retain interpretable solutions to understand the etiology of individual drivers of variation.

# Results and discussion

To address these aforementioned challenges, we here propose MultiOmics Factor Analysis (MOFA), a model for integrating multi-omic data in an unsupervised fashion. MOFA automatically discovers a typically small set of latent factors that explain the driving sources of variation across different omics. Importantly, these factors can capture sources of variation that are unique to a single data modality or explain variation in multiple layers (**Fig. 1a**). These latent factors can for example correspond to cellular states or patient subgroups, which can be hard to characterise in the complex and noisy high-dimensional representation. Once trained, the model can be used for different downstream analyses, including the visualisation of samples in factor space, the automatic annotation of factors using (gene set) enrichment analysis, and the imputation of missing values (**Fig. 1b**).

MOFA is a Bayesian model that naturally builds upon the Group Factor Analysis statistical framework (Virtanen 2012, Klami 2015, Zhao 2016, Khan 2014, Bunte 2016), which we have here extended to enable its broad applicability in the context of multi omics studies (**Methods**, **Fig. S1**, **Table S1**). Briefly, MOFA combines a two-level sparsity, thereby automatically detecting the number of relevant factors and retaining interpretable solutions for active factors, with efficient variational approximations that scale to larger datasets. Notably, our inference framework also supports non-Gaussian data modalities, thereby enabling the integration of binary data (e.g. somatic mutations), count data (e.g. copy number variation) or binomial data types (e.g. methylation read ratios). Finally, MOFA accounts for missing values, including partially profiled samples. MOFA is implemented as accessible and user-friendly software, which facilitates a range of different downstream analyses. Taken together, these features enable the robust application of MOFA to a wide range of multi-omic studies (**Methods**).

## Application to Chronic Lymphocytic Leukaemia

To illustrate our model we applied MOFA to a recent study of chronic lymphocytic leukaemia (CLL) (Dietrich, Oles, Lu, submitted), which combines drug response measurements with somatic mutation status, transcriptome profiling and DNA methylation assays (**Fig. 2a**).

Notably, XX% of the sample were not profiled in all omics, a structure of missing values MOFA can cope with (**Methods**).

In this dataset, MOFA identified 10 relevant factors (minimum explained variance 3% in at least one view; **Methods**), which were robust to subsampling of the data (**Supp. Fig. S8**) and were largely orthogonal (**Supp. Fig. S9**). Among these, factor 1 and 2 were active in all views (defined as explained variance >3%), indicating their broad roles for multiple molecular layers. Other factors such as 3 or 5 were specific to a subset of data modalities, and others such as factor 4 were active in a single data modality. Cumulatively, all factors explained 41%, 38%, 24% and 24% of total variation in the drug response view, mRNA view, DNA methylation and mutation data.

MOFA implements different strategies for using the feature weights of individual factors to identify their etiology **(Fig. 1b)**. Inspection of the top weights for factors 1 and 2 in the somatic mutation view revealed high loadings for the mutation status of the immunoglobulin heavy-chain variable region gene (IGHV) and trisomy of chromosome 12 (**Fig. 2c**). This demonstrates that MOFA discovered two major major clinical markers in CLL biology (Fabbri 2016) in a completely unsupervised manner (**Fig. 2d**). Additionally, gene set enrichment analysis on loadings of the mRNA suggested plausible etiologies for other factors that have previously not been appreciated, including ones linked to immune response pathways, T-cell receptor signalling, oxidative stress and senescence as well as RNA regulation (**Fig. 2e**).

## Identification of a continuous IGHV state

Interestingly, the IGHV factor inferred by MOFA suggests a continuous IGHV phenotype rather than two discrete states. as generally assumed in clinical practice (REF). The ordering inferred by the model was consistent with previously documented subgroups (Oakes 2015, Queirós 2014). MOFA connects this sub structure to a continuous phenotype of the IGHV status (**Fig. 3a**), and reveals its global effect across multiple molecular layers. As expected, the factor is associated to previously known gene expression markers for IGHV status, and it suggests the existence of novel markers that show a continuous phenotype (**Fig. 3b,c**) (Vasconcelos 2005, Maloum ‎2009, Trojani 2012, Morabito 2015, Plesingerova 2017). Similarly, the weights on the drug response view highlight drugs whose response is influenced by the IGHV status. This includes many kinase inhibitors targeting the B-cell receptor pathway as well as CHK inhibitors **(Fig. 3d,e**)

## MOFA identifies mis annotated samples and accurately imputes missing values

As an additional use of the inferred factors, we explored their utility for refining clinical annotations that are based on a single marker . Among the 200 patients, we observed that MOFA predicts the IGHV status for 12 patients that lack mutational data, and it reclassifies the IGHV status of 12 patients (**Fig. S10a,b**). Upon inspection of the underlying molecular data, we observe clear substructure that supports the MOFA inference of this factor (**Fig. S10a,b**). This suggests that the IGHV alone has incomplete penetrance and additionally that clinically relevant subtypes are more accurately defined by combining molecular layer.

Motivated by this, we more generally assessed the ability of MOFA to leverage larger dataset, that however include samples with incomplete profiles (**Fig. SXX**). Moreover, we tested the ability of MOFA to fill in these missing values. Across different patterns of missing information (see **Methods**) we observed that our model yielded more accurate predictions than commonly used imputation strategies, including imputation by feature-wise mean, SoftImpute (Mazumder 2010) and a k-nearest neighbour method (Troyanskaya 2001) (**Fig. S11**). Taken together, this shows that MOFA is robust to incomplete data sets and can fill in missing values from sparse profiling datasets, which could massively reduce experimental costs.

## MOFA reveals previously unknown axes of variation in CLL attributed to oxidative stress

Despite their clinical importance, the IGHV and the trisomy 12 factors explain less than 20% of the explained variance by MOFA, suggesting that the model detected other substantial source of variation in this dataset.

One example is factor 5, which is active in the mRNA and drug response view, tagging genes that were highly enriched for oxidative stress and senescence pathways (**Fig. 2e, Fig. 4a**). In particular, the top weights correspond to heat shock proteins (HSPs) (**Fig. 4b**), which are known to be essential for protein stability and are up-regulated upon stress conditions such as oxidative stress (ADD REF). Interestingly, HSPs have previously been documented to be elevated in cancer contexts and may contribute to prolonged tumour cell survival (Dempsey et al. 2010a). Consistent with this annotation based on the mRNA view, we observed that the drugs with the strongest weights on factor 5 are associated with oxidative stress response, such as target reactive oxygen species (SD07, MIS-43, SD51), and DNA damage response (fludarabine, nutlin-3, doxorubicine) and apoptosis (BH3, Survivin) (**Fig.** **4c-d**).

## Latent factors inferred by MOFA are predictive of clinical outcomes

Finally, we explored the potential of using latent factor inferred by MOFA to predict clinical outcomes. Out of the 10 factors identified, we found 3 factors that are significantly linked to patient time to next treatment (Cox regression, **Methods**, P<1E-4, **Fig. 5a,b**), including the well-characterized IGHV factor and another factor associated o pre-treatment and TP53/del17p mutational status(**Fig. S8**).

Furthermore, accurate predictions of time to next treatment were obtained when combining all 10 MOFA factor into a single model (multivariate Cox regression, **Methods, Fig. 5b**). We also compared the prediction accuracy to using latent factors inferred using conventional principal component analysis either applied to individual data modalities as well as the joint datasets consisting of all data modalities, finding that MOFA factor were more predictive than alternative representations of the data. Analogous results were obtained when using all individual features instead of principal components (**Fig. S13**).

# Conclusions

Here, we proposed Multi-Omics Factor Analysis (MOFA), an unsupervised method for the integration of multi-omics data sets. In applications to CLL, we have demonstrated that MOFA is able disentangle patient heterogeneity using high-dimensional and incomplete multi-omics profiles. Most notably, our model identified previously known clinical markers as well as novel putative drivers of molecular heterogeneity, some of which were predictive of clinical outcome. This suggests that latent factors inferred using method such as MOFA could in the future replace biomarkers based on individual features or assays.

Although the application of factor models for integrating different data types is not new per se (Lankriet 2004, Shen 2009, Akavia 2010, Mo 2012), MOFA provides a set of unique features that are critical for its practical application: (i) fast inference based on a variational approximation, (ii) sparse solutions facilitating interpretation, (iii) handling of missing values, and (iv) the use of appropriate likelihoods to model different data types. This offers important advantages compared to existing methods and in particular allow for disentangling molecular variation into orthogonal components (**Supp. Fig. XX, Methods**). Coupled with user-friendly software this will foster the accessibility of these methods for wide a range of applications.

Although we have addressed important challenges for multi-omics applications, MOFA is not free of limitations. First, the model is intrinsically linear, which means that nonlinear relationships between inferred factors and the observed data may be missed (Buettner 2012, Campbell 2015). Non-linear extensions of MOFA may address this, although there will be tradeoffs between model complexity, computational efficiency and interpretability. A related area of future work is to incorporate prior information on the relationship between individual features. For example, pathway databases or genomic relationships between different views may inform model inference (Buettner, 2016).

# Methods

A description of the MOFA models and details of all analyses is presented as **Supp. methods.**

# Figure captions

**Figure 1 | MultiOmics Factor Analysis: model and motivation.**

**(a)** Model fitting: MOFA takes consideres M data matrices as input (Y1,..., YM), one of each data modality, with co-ocurrent samples but in general different numbers of features. MOFA decomposes these matrices jointly into a matrix of factor for each sample and M weight matrices, one for each view (loadings W1,.., WM). White cells correspond to zero values whereas the cross symbol denote missing data. **(b)** The fitted mofa model can be used for different downstream analyses, including (i) variance decomposition, assessing the proportion of variance explained by each factor in each view, (ii) the semi-automated factor annotation based on the inspection of loadings and gene set enrichment analysis, (iii) visualization of the samples in the factor space and (iv) imputation of missing values, including missing assays.

**Figure 2 | Overview of factor inferred by MOFA applied to the Chronic Lymphocytic Leukemia (CLL) data set. (a)** Study overview and datatypes. Data modalities are shown in different rows with missing samples shown using grey bars. **(b,c)**  Proportion of variance explained by individual factors for each view (R2) **(b)** and cumulative proportion of variance explained **(c)**. **(d)** Relative loadings of the top genes from the somatic mutation view for factors 1 and 2. **(e)** Ordination of samples using factors 1 and 2. Colors denote the IGHV status of patients according to clinical label (mutation based); symbol shape indicates chromosome 12 trisomy status. **(f)** Number of enriched Reactome gene sets per factor (FDR< 1%). Colors denote categories of related pathways defined as in Supp. Table XX.

**Figure 3 | Characterization of the inferred IGHV factor.**

**(a)** Inferred IGH factor (factor 1) for each sample with labels corresponding to previously defined patient classification (based on methylation arrays, (REF)). LP: lowly-programmed, IP: intermediately-programmed, HP: highly-programmed. **(b)** Relative loadings for the genes with the largest weights in the mRNA view. Genes highlighted in orange have previously been described as prognostic markers in CLL and associated to IGHV status (REF). **(c)** Heatmap of gene expression values for the genes with the largest weights as in **(b)**. Samples are ordered by their values of the IGHV factor. **(d)** Relative loadings of the drugs with the largest weights, grouped by target category. **(e)** Drug response curves for selected drugs. The two groups are defined using k-means clustering on the latent factor values.

**Figure 4 | Characterization of the oxidative stress response factor.**

**(a)** Gene set enrichment for the top Reactome pathways in the mRNA view (t-test, **Methods**). **(b)** Rank distribution of the relative loadings in the mRNA view. Selected top genes are labeled and annotated using black dots. **(c)**  Heatmap of gene expression values for selected top genes as in **(b)**. Samples are ordered by their values of the stress response factor. **(d)** Relative loadings of the drugs with the largest weights, grouped by target category. **(e)** Drug response curves for selected drugs. The two groups are defined using k-means clustering on the latent factor values.

**Figure 5 | Relationship between clinical data and latent factors.**

**(a)** Association of MOFA factors to time to next treatment using a univariate Cox model. Error bars denote the 95% confidence intervals. Numbers on the right are p-values for each predictor. **(b)** Kaplan Meier Plot for the individual MOFA factors. The cut-points on each factor were chosen using the maximally selected rank statistics. **(c)** Prediction of time to treatment using multivariate Cox regression trained using the 10 factors derived using MOFA, as well as the first 10 principal components of the corresponding single views and the full dataset. Shown are average values of the Harrell’s C index using 5-fold cross-validation. Error bars denote standard errors.

# Bibliography (still editing, ignore for now)

Wang, B. Mezlini, AM. Demir, F. Fiume, M. Tu, Z. Brudno, M. Haibe-Kains, Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* **11**, 333–337 (2014)

Hasin, Y. Seldin M. & Lusis, A. Multi-Omics Approaches to Disease. *Genome Biol*. **18(1)**, 83 (2017).

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nature Reviews Genetics.* **16**:85-97 (2015).

The Cancer Genome Atlas Research Network, Weinstein, J.N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. **45**;1113-20 (2013).

Mertins, P et al, Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*. **534(7605):**55-62 (2016).

Gerstung, M. et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat Commun*. **6:5901** (2015)

Iorio, Francesco, et al. "A landscape of pharmacogenomic interactions in cancer." Cell 166.3 (2016): 740-754.

Kim M, Rai N, Zorraquino N, Tagkopoulos I. Multi-omics integration accurately predicts celular satte in unexplored conditions for Escherichia coli. *Nat Commun*. **7:13090** (2016)

Macaulay, I. C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods.* **12**:519–522 (2015).

Angermueller, C. et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods*. **13**:229–232 (2016).

Clark et al. Joint Profiling Of Chromatin Accessibility, DNA Methylation And Transcription In Single Cells. *bioRxiv*. doi:10.1101/138685 (2017)

Chang X, Dacheng T, and Chao X. A Survey on Multi-view Learning. arXiv:1304.5634 (2013)

Schadt EE. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet*. **37(7):**710-7 (2005)

Lankriet 2004?

Buettner 2012

Wang, Bo, et al. "Similarity network fusion for aggregating data types on a genomic scale." Nature methods 11.3 (2014): 333-337.

Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. **25(22)**:2906-12 (2009)

Akavia UD et al. An integrated approach to uncover drivers of cancer. *Cell*. **143(6)**:1005-17 (2010)

Mo Q et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc Natl Acad Sci USA*. **110(11):**4245-50 (2013)

Virtanen, S et al. Bayesian group factor analysis”. *Proc. 15th Int. Conf. Artificial Intelligence and Statistics*, pp. 1269–77 (2012)

A. Klami, S. Virtanen, E. Lepp ̈aaho, and S. Kaski. Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136– 2147 (2015).

Zhao S, Gao C, Mukherjee S, and Engelhardt BE. Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research*, 17(196):1–47 (2016).

S. A. Khan, S. Virtanen, O. P. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics*, **30(17)**:497–504 (2014).

A. Klami, S. Virtanen, and S. Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(Apr):965–1003 (2013).

Bunte K, Leppaaho P, Saarinen I, and Kaski S. Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics*, **32(16)**:2457– 2463, 2016.

Kieran Campbell, Christopher Yau. Bayesian Gaussian Process Latent Variable Models for pseudotime inference in single-cell RNA-seq data. 2015 doi: https://doi.org/10.1101/026872

Buettner F, Pratanwanich N, Marioni JC, Stegle O. Scalable latent-factor models applied to single-cell RNA-seq data separate biological drivers from confounding effects. *bioRxiv*. doi:10.1101/087775

## Author contributions

RA and BV contributed equally and are listed alphabetically.

VB, WH and … processed the data

RA and BV performed the analysis.

RA, DA and BV implemented the model.

RA, BV, TZ, SD, JM, WH, FB, OS interpreted results

RA, BV, WH, FB, OS conceived the project and drafted the manuscript

WH, FB, JM and OS supervised the project

## Competing financial interest statement

The authors declare no competing financial interests.