Multi-Omics factor analysis disentangles heterogeneity in blood cancer

Ricard Argelaguet[1,*], Britta Velten[2,*], Damien Arnol[1], Sascha Dietrich[4], Thorsten Zenz[4,5], John C. Marioni[1,3,6], Florian Buettner[1,7#], Wolfgang Huber[2#], Oliver Stegle[1,2]

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, UK
4. Heidelberg University Hospital, Heidelberg, Germany
5. German Cancer Research Center (dkfz) and National Center for Tumor Diseases (NCT), Heidelberg, Germany & Hematology, University Hospital Zurich and University of Zurich, 8091 Zurich, Switzerland
6. Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK
7. Helmholtz Zentrum München–German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany

* Authors contributed equally and are sorted alphabetically

# Author order determined by coin flip

Corresponding authors: Florian Buettner (fbuettner.phys@gmail.com), Wolfgang Huber (wolfgang.huber@embl.de) and Oliver Stegle (oliver.stegle@ebi.ac.uk)

Abstract

Multi-omic studies in large cohorts promise to characterize biological processes across molecular layers including genome, transcriptome, epigenome, proteome and perturbation phenotypes. However, methods for integrating multi-omic datasets in an unsupervised manner are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in a multi-omics dataset. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability across data modalities, thereby enabling a variety of downstream analyses, including factor annotation, data imputation and the detection of outlier samples. We applied MOFA to a study of 200 patient samples of chronic lymphocytic leukemia (CLL) profiled for somatic mutations, RNA expression, DNA methylation and *ex-vivo* responses to a panel of 63 drugs. MOFA discovered known dimensions of disease heterogeneity, including immunoglobulin heavy chain variable region (IGHV) status and trisomy of chromosome 12, as well as previously underappreciated drivers of variation, such as response to oxidative stress. These learnt factors capture key dimensions of inter-patient heterogeneity and enhance prediction accuracy of clinical outcomes.

Introduction

Technological advances increasingly enable multiple biological layers to be probed in parallel, ranging from genome, epigenome, transcriptome, proteome and metabolome to phenome profiling[1]. Integrative analyses that use information across these data modalities promise to deliver more comprehensive insights into the biological systems under study. Motivated by this, multi-omic profiling is increasingly applied in biological domains including cancer genomics[2-5], microbiology[6] and host-pathogen interactions[7]. Most recently, it has also become possible to perform multi-omics analyses in single cells[8-11]. A common aim of such applications is to understand the nature of any extant heterogeneity between samples, as manifested in one or several of the omic data types[12]. Multi-omics approaches are particularly appealing if not all relevant axes of variation are known *a priori*, and hence may be missed by studies that consider a single data modality or by hypothesis-driven approaches.

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is QTL-analysis, where large numbers of marginal association tests are performed between genetic variants and gene expression levels or chromatin marks [13]. While eminently useful for variant annotation, such association lists do not provide a coherent global map of the molecular differences between samples. A second, multivariate strategy is to use kernel- or graph-based methods to combine different data types into a single similarity network between samples[14, 15]; however, it is difficult to identify the principal molecular drivers of heterogeneity from such networks. Finally, there exist generalizations of clustering methods to reconstruct discrete groups samples based on multiple data modalities [16, 17].

A key limitation of existing methods is lack of interpretability, in particular because the underlying factors that drive the variation are not explicitly reconstructed - a necessary step to establish associations between specific sources of variation and external data such as organismal phenotypes or (clinical) covariates. Additionally, available methods are hampered by computational scalability to larger datasets and commonly do not appropriately handle missing values and non-Gaussian data modalities, such as binary readouts or count-based traits.

Results and discussion

We present Multi-Omics Factor Analysis (MOFA), a statistical method for integrating multiple modalities of omic data in an unsupervised fashion. Intuitively, MOFA can be viewed as a versatile and statistically rigorous generalization of principal component analysis (PCA) to multi-omics data. Given several data matrices with measurements of multiple ‘omics data types on the same or on overlapping sets of samples, MOFA infers an interpretable low-dimensional data representation in terms of (hidden) factors. These learnt factors represent the driving sources of variation across data modalities, thus facilitating the identification of cellular states or disease subgroups. Importantly, MOFA disentangles whether the underlying axes of heterogeneity are unique to a single data modality or are manifested in multiple modalities (**Fig. 1**), thereby identifying links between the different ‘omics. Once trained, the model output can be used for a range of downstream analyses, including the visualisation of samples in factor space, the automatic annotation of factors using (gene set) enrichment analysis, the identification of outliers (e.g. due to sample swaps) and the imputation of missing values (**Fig. 1**).

Technically, MOFA builds upon the statistical framework of Group Factor Analysis[18-22], which we have extended to the specific requirements of multi-omics studies (**Supp. Fig. S1**, **Supp. Table 1**). In particular, MOFA combines i) the inference of interpretable factors by using sparse representations of factor weights, ii) the regularization of the relevance of factors for individual data types, iii) scalable inference using variational Bayesian approximations to permit applications to larger datasets, iv) the support of non-Gaussian data modalities, including binary and count data and v) explicit modelling of missing data, including samples for which some, but not all data modalities were acquired. We used simulated data to systematically test the MOFA model and validate its handling of missing values and modelling of non-Gaussian data types (**Supp. Fig S2-S5**). We compared MOFA to other factor models, finding increased accuracy for identifying true drivers (**Supp. Fig. S6**). MOFA is implemented as well-documented open-source software that facilitates a range of different downstream analyses, including automatic characterization of the inferred factors (**Methods**). Taken together, these features provide a powerful and versatile tool for disentangling sources of variation in multi-omic studies.

Application to Chronic Lymphocytic Leukaemia

We applied MOFA to a study of chronic lymphocytic leukaemia (CLL), which combined ex-vivo drug response measurements with somatic mutation status, transcriptome profiling and DNA methylation assays (**Fig. 2a**). Notably, nearly 40% of the 200 samples were profiled with some but not all ‘omics types; such a missing value scenario is not uncommon in large cohort studies, and MOFA is designed to cope with it (**Methods; Supp. Fig S2**).

MOFA identified 10 factors (minimum explained variance 3% in at least one view; **Methods**). These were robust to algorithm initialisation as well as subsampling of the data (**Supp. Fig. S7,8**). The factors were largely orthogonal, capturing independent sources of variation (**Supp. Fig. S9**). Among these, Factors 1 and 2 were active in most views, indicating broad roles in multiple molecular layers (**Fig. 2b**). However, other factors such as Factor 3 or Factor 5 were specific to two data modalities, and Factor 4 was active in a single data modality only. Cumulatively, the factors explained 41% of variation in the drug response data, 38% in the mRNA data, 24% in the DNA methylation data and and 24% in the mutation data.

MOFA identifies and refines known clinical markers in CLL

As part of the downstream pipeline, MOFA provides different strategies to use the loadings of the features on each factor in order to identify their etiology **(Fig. 1)**. For example, based on the top weights in the mutation view, Factor 1 was aligned with the somatic mutation status of the immunoglobulin heavy-chain variable region gene (IGHV), while Factor 2 aligned with trisomy of chromosome 12 (**Fig. 2c**). Thus, MOFA correctly pinpointed the two most important clinical markers in CLL and identified them with two major axes of molecular disease heterogeneity, which were active at multiple ‘omics layers [23, 24] (**Fig. 2d**).

Factor 1 aligned largely with IGHV status, a surrogate of the differentiation state of the tumor’s cell of origin and the level of activation of the B-cell receptor. While in clinical practice this axis of variation is generally considered binary[23], our results indicate a more complex substructure (**Fig. 3a**). At the current resolution, this factor is consistent with three subgroup models such as proposed by[25, 26] (**Supp. Fig. S10**), although there is suggestive evidence for an underlying continuum. MOFA robustly connected this factor to multiple molecular layers (**Supp. Fig. 11, 12**), including changes in the expression of genes previously linked to IGHV status[27-31] (**Fig. 3b,c**) and with drugs that target kinases in or downstream of the B-cell receptor **(Fig. 3d,e**).

Taken together, Factor 1 captures a global cell state that is reflected in almost all molecular layers and represents the differentiation state of the cell of origin and reliance on B-cell receptor signalling,

MOFA reveals a previously underappreciated axis of variation in CLL attributed to oxidative stress

Despite their clinical importance, the IGHV and the trisomy 12 factors account for less than 20% of the variance explained by MOFA, suggesting the existence of other sources of variation. One example is Factor 5, which is active in the mRNA and drug response views. This factor tagged a set of genes that were highly enriched for oxidative stress and senescence pathways (**Fig. 2e, Fig. 4a**), with the top weights corresponding to heat shock proteins (HSPs) (**Fig. 4b,c**), a group of genes that are essential for protein folding and are up-regulated upon stress conditions[32, 33]. The expression levels of HSPs are known to be elevated in some cancers and may contribute to prolonged tumour cell survival[34], but so far have received little attention as biomarkers in CLL. Consistent with the annotation based on the mRNA view, we observed that the drugs with the strongest weights on Factor 5 were associated with response to oxidative stress, such as target reactive oxygen species, DNA damage response and apoptosis (**Fig. 4d,e**). Overall, our results suggest that changes in oxidative stress levels are a heterogeneous feature of CLL and could underpin a functionally relevant phenotype.

For other factors, gene set enrichment analysis on the loadings of the mRNA view suggested etiologies such as immune response pathways, T-cell receptor signalling (suggesting possible T-cell contamination), and senescence. (**Fig. 2e**)

MOFA identifies outlier samples and accurately imputes missing values

Next, we explored the relationship between inferred factors and clinical annotations, which can be missing, mis-annotated or inaccurate, since they are frequently based on single markers or imperfect surrogates[35]. As in clinical practice patients are divided into two major disease subgroups on the basis of IGHV status, we assessed the consistency between the inferred continuous Factor 1 and these groups. For 176 out of 200 patients, the MOFA factor was in agreement with the clinical IGHV status, and MOFA further allowed for classifying 12 patients that lacked clinically measured IGHV status (**Fig. S13a,b**). Interestingly, MOFA assigned 12 patients to a different group than suggested by their clinical IGHV label. Upon inspection of the underlying molecular data, nine of these cases showed intermediate molecular signatures, suggesting that they are borderline cases that are not appropriately captured by the binary classification; the remaining three cases were clearly discordant, highlighting that the binary IGHV status is not a perfect biomarker of CLL biology (**Fig. S13c,d**). Overall, this suggests that latent factors inferred from the multi-omic data can improve biology-based disease stratification.

As incomplete data is a common problem in studies that combine multiple high-throughput assays, we assessed the ability of MOFA to fill in missing values within assays as well as when entire data modalities are missing for some of the samples. For both imputation tasks, MOFA yielded more accurate predictions than established imputation strategies, including imputation by feature-wise mean, SoftImpute[36] and a k-nearest neighbour method[37] (**Fig. S14, Fig. S15**). These results demonstrate that MOFA can leverage information from multiple omics layers to accurately impute missing values from sparse profiling datasets.

Latent factors inferred by MOFA are predictive of clinical outcomes

Finally, we explored the potential of using the latent factors inferred by MOFA to predict clinical outcomes. Three of the 10 factors identified by MOFA were significantly associated with time to next treatment (Cox regression, **Methods**, $P<10^{-4}$, **Fig. 5a,b**): the cell of origin related Factor 1, and two factors associated with TP53/del17p mutational status (Factor 7; **Fig. S16**) and whether the patient was treated with chemo-immunotherapy before sample collection (Factor 8) respectively.

We also assessed the prediction performance when combining the 10 MOFA factors in a multivariate Cox regression model (assessed using cross-validation, **Methods, Fig. 5c**). Notably, this model yielded higher prediction accuracy than models using factors derived from conventional PCA (**Fig. 5c**) or using the individual features (**Fig. S17**).

Conclusions

Multi-Omics Factor Analysis (MOFA) is an unsupervised method for decomposing the sources of heterogeneity in multi-omics data sets. We have applied MOFA to high-dimensional and incomplete multi-omics profiles collected from patient-derived tumour samples, demonstrating that our method is able to identify the drivers of variation in CLL, a clinically and biologically heterogeneous disease. Most notably, our model identified all previously known clinical markers as well as novel putative molecular drivers of heterogeneity, some of which were predictive of clinical outcome. This indicates that latent factors inferred by methods such as MOFA are a powerful complement to biomarkers that are based on individual features and single assays. Additionally, because MOFA factors capture variations of multiple features and data modalities, these inferred factors can help to mitigate assay noise and technical variability, thereby increasing the sensitivity for identifying molecular signatures compared to using individual features or assays.

While applications of factor models for integrating different data types have been reported previously[14, 16, 17, 38], MOFA has unique features that are critical for applying factor methods to 'omics data in practice: (i) fast inference based on a variational approximation, (ii) inference of sparse solutions facilitating interpretation, (iii) handling of missing values, and (iv) the flexible combination of different data likelihood models to integrate diverse data types, including non-normal data. This offers important advantages compared to existing methods and in particular allows for disentangling molecular variation into distinct components (**Methods**). MOFA is fully automated and available as open source software. Taken together, this will foster the accessibility of these methods for a wide range of applications.

Although we have addressed important challenges for multi-omics applications, MOFA is not free of limitations. The model is linear, which means that it can miss strongly non-linear relationships between features within and across views[39]. Non-linear extensions of MOFA may address this, although as with any models in high-dimensional spaces, there will be trade-offs between model complexity, computational efficiency and interpretability[40]. A related area of work is to incorporate prior information on the relationships between individual features. For example, future extensions could make use of pathway databases within views[41] or priors that reflect relationships that capture the 'dogma of molecular biology'.

Methods and availability of software

CLL study

We collected peripheral blood samples from 200 patients chronic lymphocytic leukemia (CLL). Blood was separated by a Ficoll gradient (GE Healthcare, Freiburg, Germany), and mononuclear cells were cryopreserved. Samples were profiled for somatic mutations (combination of targeted and whole exome sequencing), RNA expression (RNA-Seq), DNA methylation (Illumina arrays) and cell viability *ex-vivo* after 48h exposure to a panel of 63 drugs at five concentrations each. Cell viability was read out using the ATP-based CellTiter Glo assay (Promega, Fitchburg, WI, USA), and luminescence was measured with a Tecan Infinite F200 Microplate Reader (Tecan Group AG, Männedorf, Switzerland). Details of the data collection will be described elsewhere.

MOFA model

Briefly, MOFA builds on the statistical framework of Group Factor Analysis, which we adapted to the specific requirements of multi-omics studies with non-Gaussian data types, sparse solutions and fast inference. A complete description of the MOFA model and details of all analyses is presented as **Supp. methods.** An open source implementation of MOFA is available from https://github.com/PMBio/MOFA.

Figures

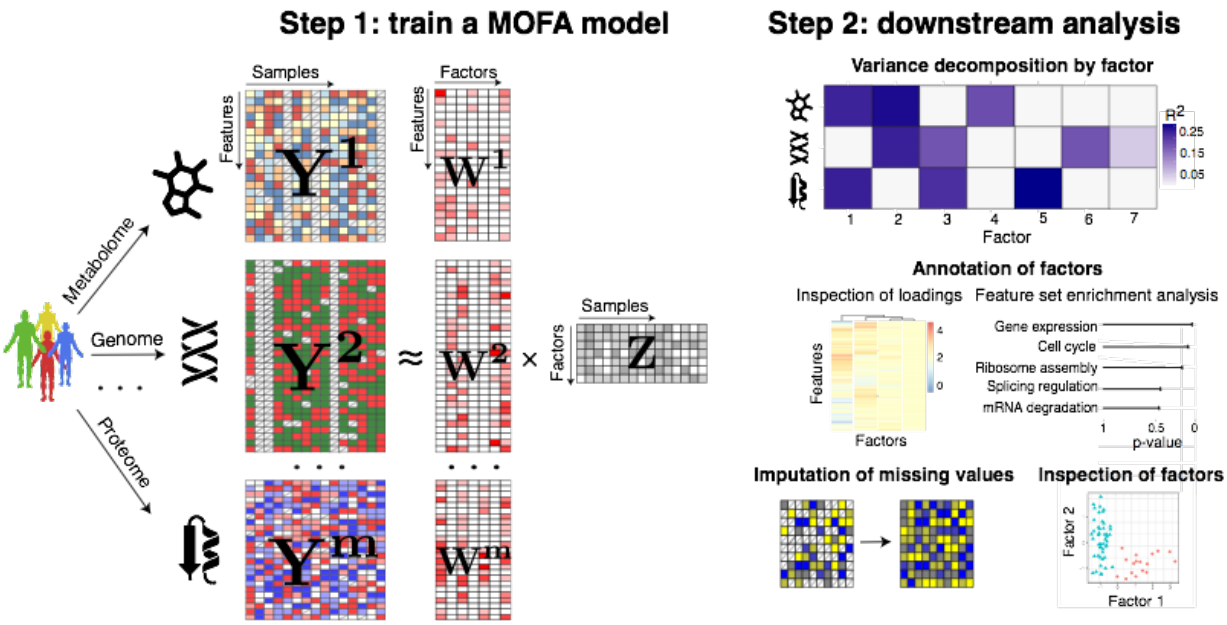

Figure 1 | Multi-Omics Factor Analysis: model and motivation.

Step 1: Model fitting: MOFA takes an arbitrary number of M data matrices as input (Y^1,..., Y^M), one or more from each data modality, with co-occurrent samples but features that are in general unrelated and that differ in numbers. MOFA decomposes these matrices into a matrix of factors, Z, for each sample and M weight matrices, one for each view (loadings W^1,.., W^M). White cells in the weight matrices correspond to zeros, i.e. inactive features, whereas the cross symbol in the data matrices denote missing values. **Step 2:** The fitted MOFA model can be queried for different downstream analyses, including (i) variance decomposition, assessing the proportion of variance explained by each factor in each view, (ii) 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.

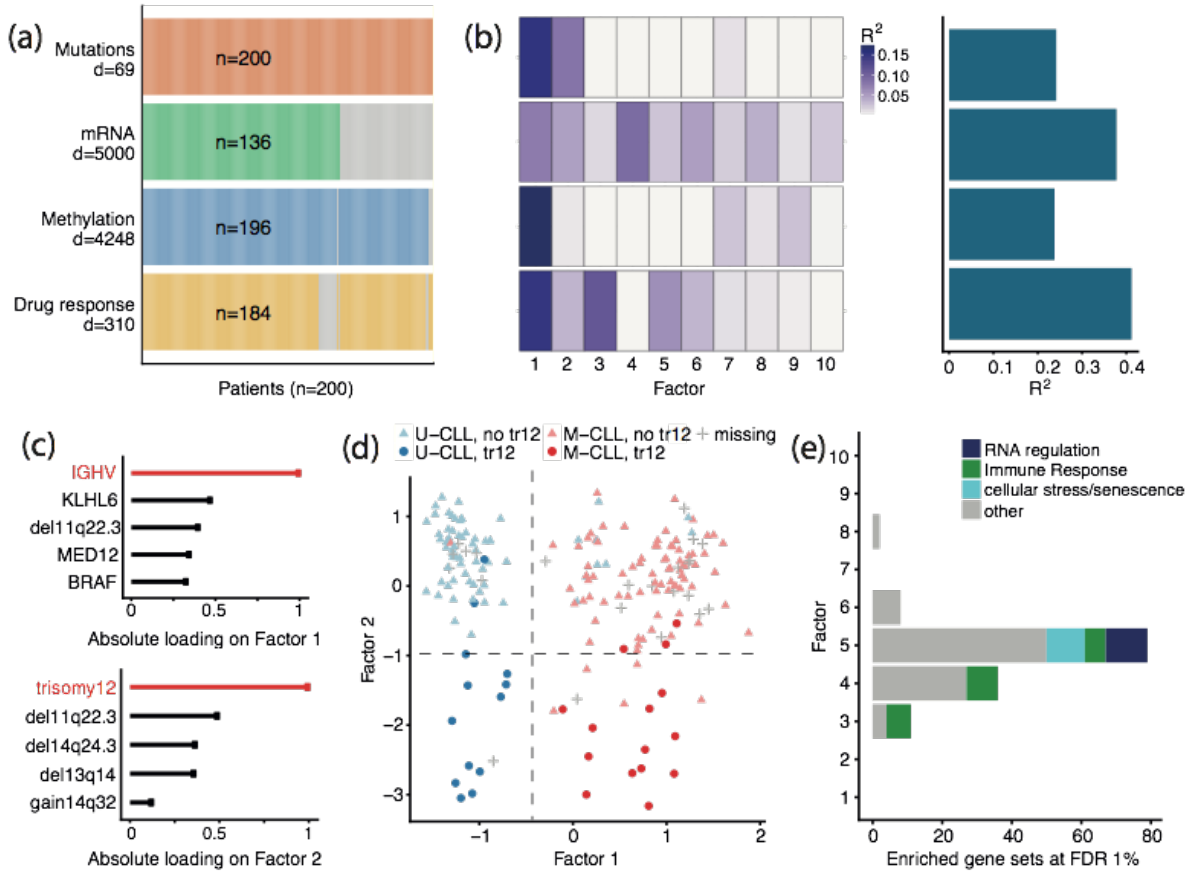

Figure 2 | Application of MOFA to a study of chronic lymphocytic leukemia.

(a) Study overview and datatypes. Data modalities are shown in different rows (d = number of features) and samples in columns, with missing samples shown using grey bars. **(b)** (Left) Proportion of total variance explained by individual factors for each view (R^2) and (Right) cumulative proportion of total variance explained. **(c)** Relative loadings of the top features of Factors 1 and 2 in the somatic mutation view. **(d)** Visualisation of samples using Factors 1 and 2. Colors denote the IGHV status of patients according to clinical label (determined by targeted sequencing); symbol shape and color indicate chromosome 12 trisomy status. **(e)** Number of enriched Reactome gene sets per factor based on the gene expression view (FDR< 1%). Colors denote categories of related pathways defined as in **Supp. Table S2**.

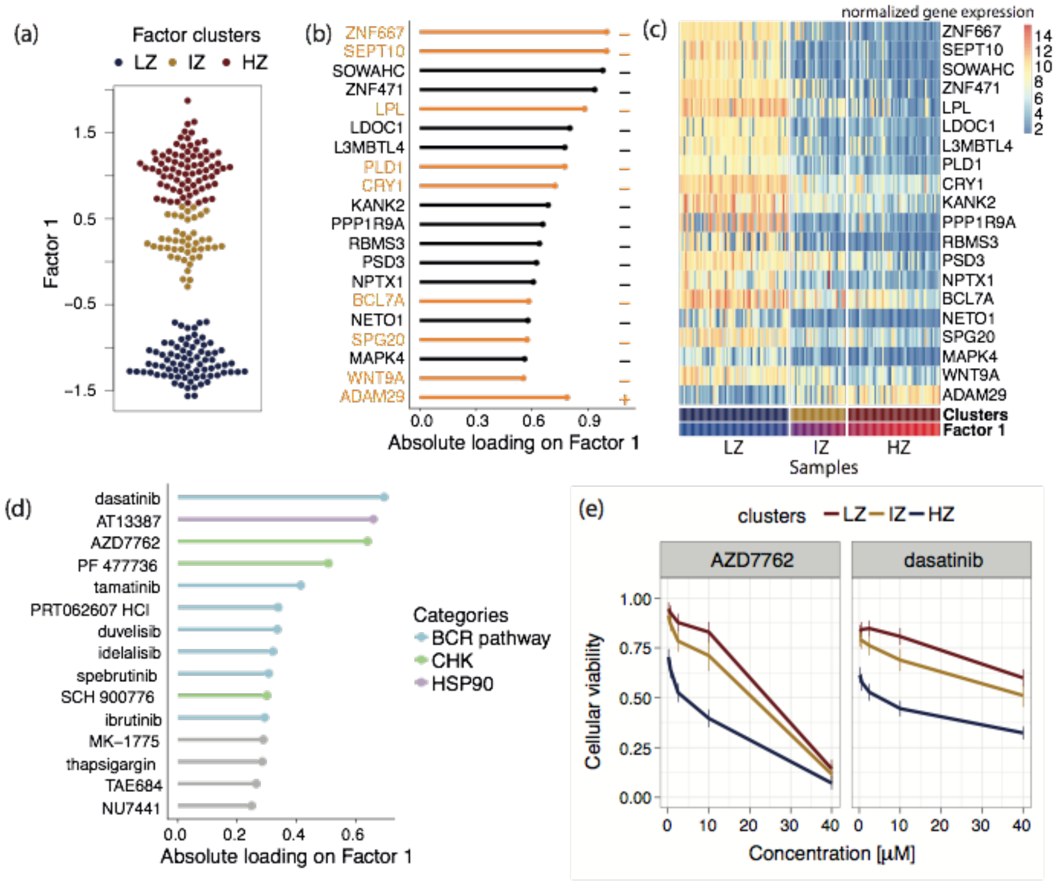

Figure 3 | Characterization of the inferred factor associated to the differentiation state of the cell of origin.

(a) Factor 1 values for each sample with colors corresponding to three clusters found by 3-means clustering with low factor values (LZ), intermediate factor values (IZ) and high factor values (HZ). **(b)** Scaled absolute loadings for the genes with the largest absolute weights in the mRNA view. Plus or minus symbols on the right indicate the sign of the loading. Genes highlighted in orange were previously described as prognostic markers in CLL and associated with IGHV status[27-31]. **(c)** Heatmap of gene expression values for genes with the largest weights as in **(b)**. **(d)** Scaled absolute loadings of the drugs with the largest weights, annotated by target category. **(e)** Drug response curves for two of the drugs, stratified by the clusters in **(a)**.

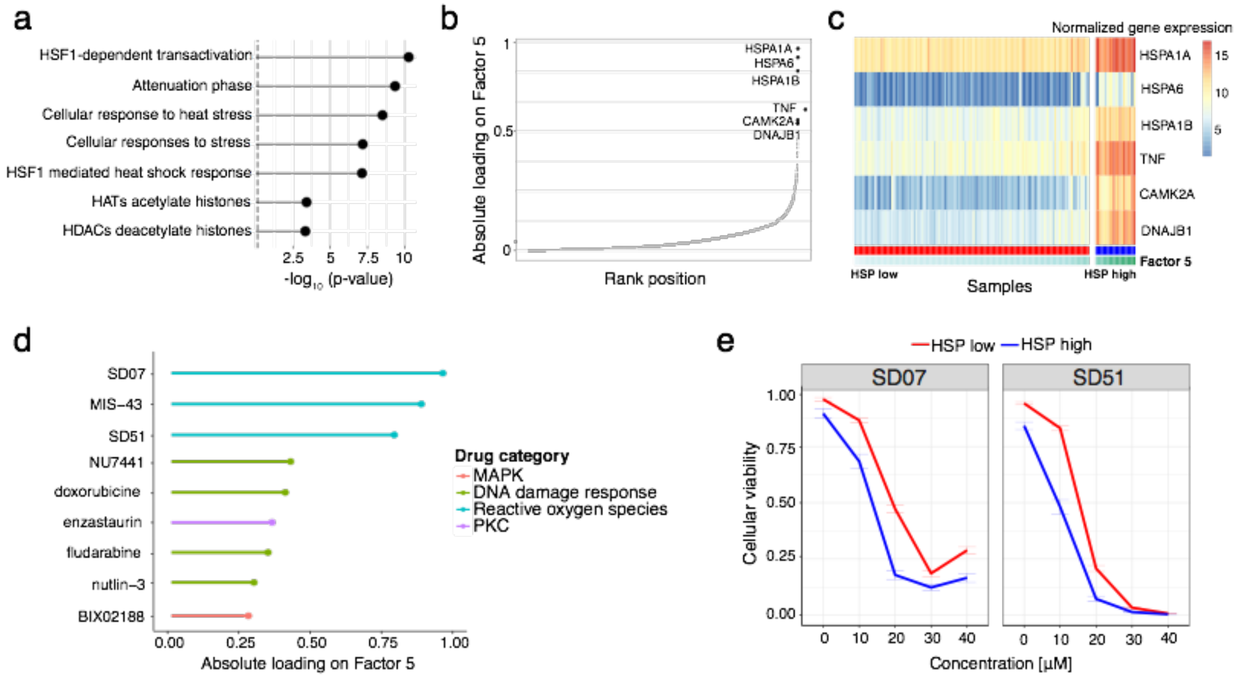

Figure 4 | Characterization of the oxidative stress response factor.

(a) Gene set enrichment for the top Reactome pathways in the mRNA view (t-test, **Methods**). **(b)** Rank distribution of the scaled absolute loadings in the mRNA view. Selected top genes are highlighted. **(c)** Heatmap of gene expression values for the six genes from Panel **(b)**. Samples are ordered by their values of the stress response factor. **(d)** Scaled absolute loadings of the drugs with the largest weights, annotated by target category. **(e)** Drug response curves for selected drugs. Sample groups are defined using 2-means clustering on the latent factor values.

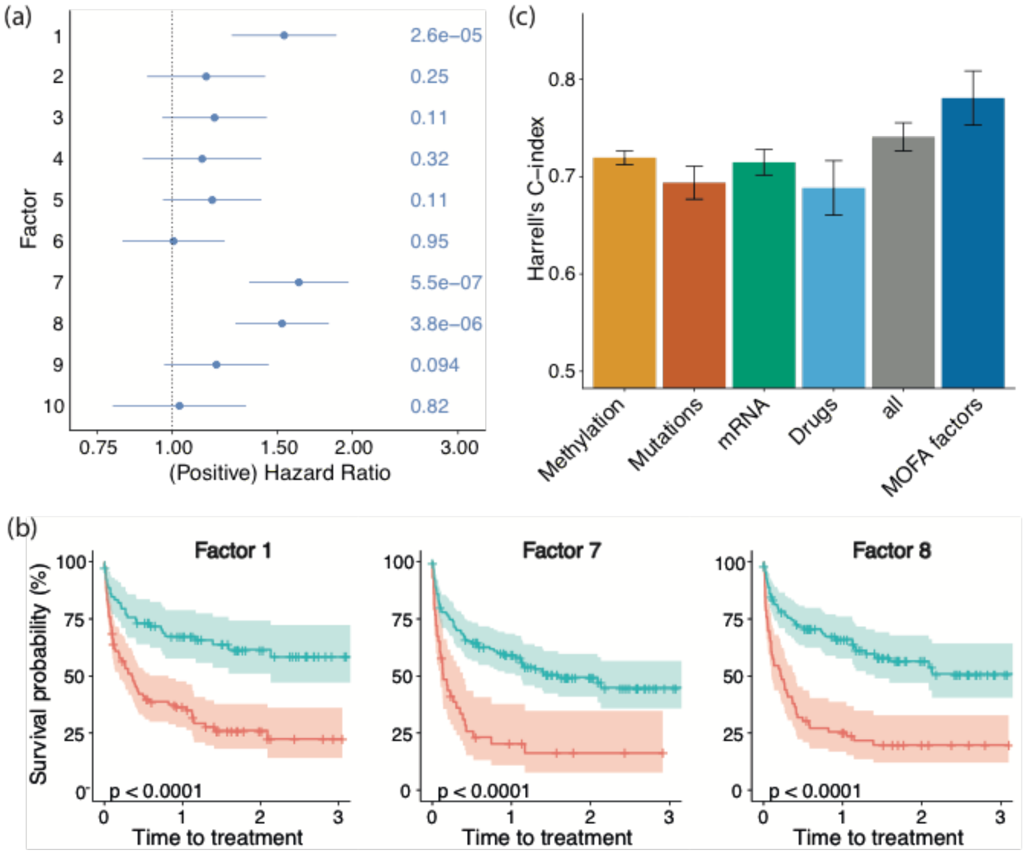

Figure 5 | Relationship between clinical data and latent factors.

(a) Association of MOFA factors to time to next treatment using a univariate Cox models. Error bars denote 95% confidence intervals. Numbers on the right denote p-values for each predictor. **(b)** Kaplan-Meier plots for the individual MOFA factors. The cut-points on each factor were chosen using maximally selected rank statistics[42], and p-values were calculated using a Log-rank test on the resulting groups. **(c)** Prediction accuracy of time to treatment using multivariate Cox regression trained using the 10 factors derived using MOFA, as well using the first 10 components obtained from PCA applied to the corresponding single views and the full dataset (assessed on hold-out data). Shown are average values of Harrell's C index from 5-fold cross-validation. Error bars denote standard error of the mean.

Author contributions

RA and BV contributed equally and are listed alphabetically.

FB, DA and OS conceived the model.

RA, DA and BV implemented the model.

TZ, SD, WH designed the CLL study and generated the data.

RA and BV performed the analysis.

RA, BV, DA, TZ, SD, WH, OS, FB, JM interpreted the results.

RA, BV, OS, WH, FB conceived the project.

RA, BV, OS, FB, WH wrote the manuscript.

OS, WH, FB, JM supervised the project.

Acknowledgements

We thank everybody involved in the generation and analysis of the original CLL study for sharing their data and analysis ahead of publication, especially M. Oleś for providing the associated data package and to J. Lu, J. Hüllein and A. Mock for discussions on CLL biology. We also thank D. Horta and DJ. McCarthy for advice on the implementation.

Competing financial interest statement

The authors declare no competing financial interests.

References

1. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biol, 2017. **18**(1): p. 83.
2. Cancer Genome Atlas Research Network, *Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma.* Cell, 2017. **169**(7): p. 1327-1341 e23.
3. Mertins, P., et al., *Proteogenomics connects somatic mutations to signalling in breast cancer.* Nature, 2016. **534**(7605): p. 55-62.
4. Gerstung, M., et al., *Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes.* Nat Commun, 2015. **6**: p. 5901.
5. Iorio, F., et al., *A Landscape of Pharmacogenomic Interactions in Cancer.* Cell, 2016. **166**(3): p. 740-754.
6. Kim, M., et al., *Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli.* Nat Commun, 2016. **7**: p. 13090.
7. Soderholm, S., et al., *Multi-Omics Studies towards Novel Modulators of Influenza A Virus-Host Interaction.* Viruses, 2016. **8**(10).

8. Guo, F., et al., *Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells.* Cell Res, 2017. **27**(8): p. 967-988.
9. Angermueller, C., et al., *Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.* Nat Methods, 2016. **13**(3): p. 229-232.
10. Clark, S.J., et al., *Joint Profiling Of Chromatin Accessibility, DNA Methylation And Transcription In Single Cells.* bioRxiv, 2017: p. 138685.
11. Macaulay, I.C., et al., *G&T-seq: parallel sequencing of single-cell genomes and transcriptomes.* Nat Methods, 2015. **12**(6): p. 519-22.
12. Ritchie, M.D., et al., *Methods of integrating data to uncover genotype-phenotype interactions.* Nat Rev Genet, 2015. **16**(2): p. 85-97.
13. Chen, L., et al., *Genetic drivers of epigenetic and transcriptional variation in human immune cells.* Cell, 2016. **167**(5): p. 1398-1414. e24.
14. Lanckriet, G.R., et al., *A statistical framework for genomic data fusion.* Bioinformatics, 2004. **20**(16): p. 2626-2635.
15. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nat Methods, 2014. **11**(3): p. 333-7.
16. Shen, R., A.B. Olshen, and M. Ladanyi, *Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.* Bioinformatics, 2009. **25**(22): p. 2906-12.
17. Mo, Q., et al., *Pattern discovery and cancer gene identification in integrated cancer genomic data.* Proc Natl Acad Sci U S A, 2013. **110**(11): p. 4245-50.
18. Virtanen, S., et al. *Bayesian group factor analysis*. in *Artificial Intelligence and Statistics*. 2012.
19. Klami, A., et al., *Group Factor Analysis.* IEEE Trans Neural Netw Learn Syst, 2015. **26**(9): p. 2136-47.
20. Zhao, S., et al., *Bayesian group factor analysis with structured sparsity.* Journal of Machine Learning Research, 2016. **17**(196): p. 1-47.
21. Khan, S.A., et al., *Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis.* Bioinformatics, 2014. **30**(17): p. i497-504.
22. Bunte, K., et al., *Sparse group factor analysis for biclustering of multiple data sources.* Bioinformatics, 2016. **32**(16): p. 2457-63.
23. Fabbri, G. and R. Dalla-Favera, *The molecular pathogenesis of chronic lymphocytic leukaemia.* Nat Rev Cancer, 2016. **16**(3): p. 145-62.
24. Zenz, T., et al., *From pathogenesis to treatment of chronic lymphocytic leukaemia.* Nature Reviews Cancer, 2010. **10**(1): p. 37-50.
25. Oakes, C.C., et al., *DNA methylation dynamics during B cell maturation underlie a continuum of disease phenotypes in chronic lymphocytic leukemia.* Nat Genet, 2016. **48**(3): p. 253-64.
26. Queiros, A.C., et al., *A B-cell epigenetic signature defines three biologic subgroups of chronic lymphocytic leukemia with clinical impact.* Leukemia, 2015. **29**(3): p. 598-605.
27. Vasconcelos, Y., et al., *Gene expression profiling of chronic lymphocytic leukemia can discriminate cases with stable disease and mutated Ig genes from those with progressive disease and unmutated Ig genes.* Leukemia, 2005. **19**(11): p. 2002-5.
28. Maloum, K., et al., *IGHV gene mutational status and LPL/ADAM29 gene expression as clinical outcome predictors in CLL patients in remission following treatment with oral fludarabine plus cyclophosphamide.* Ann Hematol, 2009. **88**(12): p. 1215-21.
29. Trojani, A., et al., *Gene expression profiling identifies ARSD as a new marker of disease progression and the sphingolipid metabolism as a potential novel metabolism in chronic lymphocytic leukemia.* Cancer Biomarkers, 2012. **11**(1): p. 15-28.

30. Morabito, F., et al., *Surrogate molecular markers for IGHV mutational status in chronic lymphocytic leukemia for predicting time to first treatment.* Leuk Res, 2015. **39**(8): p. 840-5.
31. Plesingerova, H., et al., *COBLL1, LPL and ZAP70 expression defines prognostic subgroups of chronic lymphocytic leukemia patients with high accuracy and correlates with IGHV mutational status.* Leuk Lymphoma, 2017. **58**(1): p. 70-79.
32. Åkerfelt, M., R.I. Morimoto, and L. Sistonen, *Heat shock factors: integrators of cell stress, development and lifespan.* Nature reviews. Molecular cell biology, 2010. **11**(8): p. 545.
33. Srivastava, P., *Roles of heat-shock proteins in innate and adaptive immunity.* Nature reviews. Immunology, 2002. **2**(3): p. 185.
34. Trachootham, D., J. Alexandre, and P. Huang, *Targeting cancer cells by ROS-mediated mechanisms: a radical therapeutic approach?* Nat Rev Drug Discov, 2009. **8**(7): p. 579-91.
35. Westra, H.-J., et al., *MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects.* Bioinformatics, 2011. **27**(15): p. 2104-2111.
36. Mazumder, R., T. Hastie, and R. Tibshirani, *Spectral Regularization Algorithms for Learning Large Incomplete Matrices.* J Mach Learn Res, 2010. **11**: p. 2287-2322.
37. Troyanskaya, O., et al., *Missing value estimation methods for DNA microarrays.* Bioinformatics, 2001. **17**(6): p. 520-5.
38. Akavia, U.D., et al., *An integrated approach to uncover drivers of cancer.* Cell, 2010. **143**(6): p. 1005-17.
39. Buettner, F. and F.J. Theis, *A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst.* Bioinformatics, 2012. **28**(18): p. i626-i632.
40. Damianou, A., N.D. Lawrence, and C.H. Ek, *Multi-view Learning as a Nonparametric Nonlinear Inter-Battery Factor Analysis.* arXiv preprint arXiv:1604.04939, 2016.
41. Buettner, F., et al., *f-scLVM: scalable and versatile factor analysis for single-cell RNA-seq.* Genome Biology (in press), 2017: p. 087775.
42. Hothorn, T. and B. Lausen, *On the exact distribution of maximally selected rank statistics.* Computational Statistics & Data Analysis, 2003. **43**(2): p. 121-137.