Supplementary Figure 1. Graphical model representation of MOFA

Grey-filled nodes denote observed variables whereas white-filled nodes denote unobserved variables and are inferred by the model.

Y represents the observed data. Z represents the latent factors. SW represents the spike-and-slab distributed weights as the product of a Bernoulli variable S and a Gaussian variable W. Theta represents the sparsity parameter of the spike-and-slab prior. Alpha represents the precision of the view and factor-wise Automatic Relevance Determination prior. Tau represents the precision of the normally-distributed noise.

N is the number of samples, M is the number of views, D is the number of features in the mth view and K is the number of latent factors.

Supplementary Figure 2. Technical assessment of MOFA under simulated data.

1. Comparison of the inferred and true dimensionality of the latent space. Boxplots show the distribution across 10 trials.

(b-d) Recovery of the true number of latent factors (k=10) under different number of (b) views, (c) features and (d) fraction of missing values. Each bar is a different trial.

Supplementary Figure 3. Assessment of the spike-and-slab sparsity.

1. Cumulative density function of the weights in a model with a spike-and-slab prior (blue), a model with an automatic relevance determination prior (red) and a PCA solution with the concatenated data (green).
2. …

Supplementary Figure 4. Assessment of the Bernoulli likelihood for binary data.

Sets of 10 models each were fit using a Bernoulli likelihood or Gaussian likelihood.

1. Comparison of the variational evidence lower bound.
2. Comparison of the reconstruction mean squared error.
3. Number of inferred latent variables. The solid line indicates the true number of latent variables (k=10)
4. Distribution of the reconstructed data.

Supplementary Figure 5. Assessment of the Poisson likelihood for count data

Sets of 10 models each were fit with Poisson likelihood and a Gaussian likelihood.

1. Comparison of the variational evidence lower bound.
2. Comparison of the reconstruction mean squared error.
3. Number of inferred latent variables. The line marks the true number of latent variables (k=10).
4. Distribution of the reconstructed data.

Supplementary Figure 6. Scalability of MOFA

Supplementary Figure 7. Comparison of the ability to disentangle sources of variation by iCluster and MOFA under simulated data.

Data was simulated from the generative model using the true activity pattern of factors per view on the left. Recovered factors are sorted according to the number of views, where a factor was called active in a view if it had a R2 greater than 0.05.

1. Number of true underlying factor k=10
2. Number of true underlying factor k=15

Supplementary Figure 8. Robustness across trials in CLL data

1. Training curve on the number of active factors
2. Number of active factors for each trial after convergence, colored by the corresponding evidence lower bound.
3. Correlation between the weights across trials.
4. Correlation between the latent factors across trials.

Supplementary Figure 9. Comparison of iCluster and MOFA in CLL data

As iCluster does not handle missing values, both methods were trained on a reduced data set, which is completely observed, i.e. methylation, RNAseq and drug response data for n=121 samples. The number of factors were fixed to k=10.

1. Absolute Pearson correlation coefficient between the latent variables inferred by MOFA (top) and iCluster (bottom)
2. Variance decomposition in the model inferred by MOFA (top) and iCluster (bottom).

Supplementary Figure 10. Prediction of IGHV label based on MOFA factor

1. Beeswarm of factor 1 with colours denoting agreement between predicted and clinical label
2. Pie plot showing total numbers for agreement of imputed labels with clinical label.
3. Sample-sample correlation based on methylation data
4. Sample-sample correlation based on drug response data

Supplementary Figure 11. Imputation of missing values in the drug response view.

Methods for comparison are SoftImpute, imputation by feature-wise mean (Mean) and k-nearest neighbour (kNN). Curves show the mean value of the mean squared error (MSE) across 10 runs with error bars given by twice the standard error.

1. Values missing at random
2. Full measurements missing

Supplementary Figure 12. Characterisation of latent factor 5 (stress response)

1. Top 5 most statistically significant enriched gene ontology pathways in the mRNA view
2. Distribution of the absolute value of the gene loadings in the mRNA view. The top features are labeled and displayed in black dots
3. Heatmap of the gene expression data for the top features with highest loading, samples are order according to their values on factor 5.
4. Loadings of the top features in the drug response views, colored by the corresponding main target category.
5. Drug response curves for the top three drugs classified as targeting Reactive Oxygen Species. Patients are split in two groups using k-means clustering on the latent factor values.

Supplementary Figure 13. Characterisation of latent factor 7

1. Density of pretreatment
2. Kaplan-Meier plot
3. Heatmap of the hits with top weight in methylation view.
4. Top weights in the mutation view
5. Top weights in the drug response view
6. Drug response curve for rotenone for two groups of patients obtained by hierarchical clustering on the values of the latent factor

Supplementary Figure 14. Comparison of the prediction of time to treatment using the MOFA factors and the original views.

All features of the corresponding view(s) are used as predictors in a L2-penalized Cox model. For comparison the result using a Cox model with the 10 MOFA factors is shown as in figure 4. The y-axis shows Harrell’s C index as a measure of prediction performance. The average value over 5-fold cross-validation is shown with error bars indicating the standard error. Views with missing values were imputed using the feature-wise mean.

Supplementary Table 1. Comparison of MOFA with previous Group Factor Analysis implementations

Supplementary Table 2. Settings for simulations