Multi-Omics Factor Analysis a framework for unsupervised integration of multi-omics data sets

Speaker: Jeff Ho

Introduction

- Integrative analyses use information across these data modalities promise to deliver more comprehensive insights into the biological systems.
  - Appealing if the relevant information is not known as a priori. → Missed by a single data mode

- Multi-omics profiling is increasingly applied across biological domains
  - Cancer biology
  - Regulatory genomics
  - Microbiology
  - Host-pathogen biology

- Recently, multi-omics analyses at single-cell analysis

- These applications ➔ Characterize heterogeneity between samples.

Introduction

- Basic strategy for the integration of omics data:
  - Testing for marginal associations between different data modalities, such as molecular quantitative trait locus mapping.

    → Useful for variant annotation

    → Cannot provide a coherent global map of the molecular differences between samples.
  - The use of kernel- or graph-based methods to combine different data types into a common similarity network between samples

    → Difficult to pinpoint the molecular determinants of the resulting graph structure.
- The shortage of current methods
  - Cannot reconstruct the underlying factors (such as continuous gradients, discrete clusters, or both) that drive the observed variation across samples. → Interpretability
  - Estimated features lack sparsity → Reduce interpretability
  - Require a substantial number of parameters → computationally demanding
  - Cannot handle missing values and non-Gaussian data modalities (Binary readouts or count-based traits) → Inflexibility for large data sets

Introduction

- Multi-Omics Factor Analysis (MOFA)
  - A statistical method for integrating multiple modalities of omics data in an unsupervised fashion.
  - A versatile and statistically rigorous generalization of principal component analysis (PCA) to multi-omics data.
  - Builds upon the statistical framework of Group Factor Analysis (GFA).

- Results of the MOFA
  - Variance decomposition by factors.
  - Downstream analysis.
    - Visualization, clustering, classification of samples in the low-dimensional spaces.
    - Automated annotation of factors using gene set enrichment analysis
    - The identification of outliers samples.
    - Imputation of missing values.

Introduction

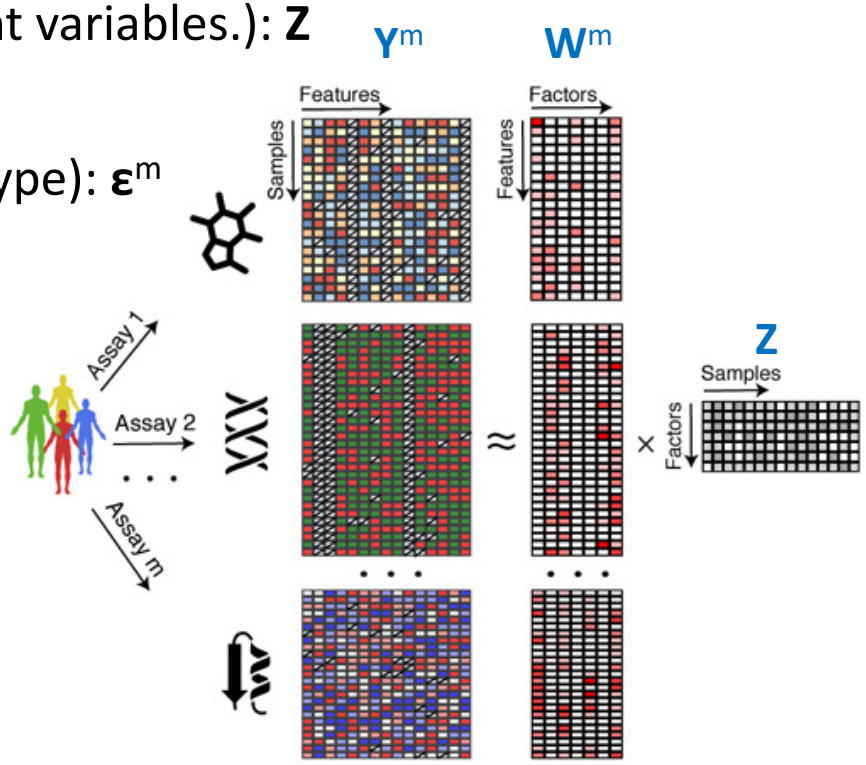
- Benefits for the MOFA
  - Infers an interpretable low-dimensional data representation in terms of (hidden) factors
  - The inferred factor loadings can be sparse. → Facilitating (The factors ⇌ Relevant molecular features)
  - Disentangles to what extent each factor is unique to a single data modality or is manifested in multiple modalities.
  - Fast inference based on a variational approximation.
  - Efficient handling of missing values.
  - Flexible combination of different likelihood models for each data modality

Methods

- MOFA models
  - Factor Analysis: Reduce the dimensionality of a (big) dataset into a small set of variables which are easier to interpret and visualize.
  - Sample size: N (n)
  - Feature numbers: D (d)
  - Data matrices, view: M (m)
  - Factor matrix (matrix that contains the low-dimensional latent variables.): **Z**
  - Loading matrix, weight matrix: $\mathbf{W}^m$
  - Residual noise (its form depends on the specific of the data type): $\boldsymbol{\varepsilon}^m$
  - The expectation of x under the distribution q. ($E_q[x]$): $\langle x \rangle$
  - Model:

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m$$

  - A Bayesian framework is used. (Place prior dist. on **Z**, $\mathbf{W}^m$, $\boldsymbol{\varepsilon}^m$)



Methods

- Models for Gaussian data:
  - Likelihood function:

$$p(y_{nd}^m) = \mathcal{N}\left(y_{nd}^m \,|\, \mathbf{z}_{n,:}\mathbf{w}_{d,:}^{mT}, 1/\tau_d^m\right)$$

    - $\mathbf{w}_d^m$: The d-th row of the loading matrix.
    - $\boldsymbol{\tau}^m$: The precision of the noise. $p(\epsilon_d^m) = \mathcal{N}\left(\epsilon_d^m \,|\, 0, 1/\tau_d^m\right)$
  - Prior dist. for **Z** ,$\boldsymbol{\tau}^m$:

$$p(z_{n,k}) = \mathcal{N}\left(z_{n,k} \,|\, 0, 1\right)$$
$$p(\tau_d^m) = \mathcal{G}\left(\tau_d^m \,|\, a_0^\tau, b_0^\tau\right)$$

  - Prior dist. for $\mathbf{w}_d^m$:
    - Reparameterization: $w_{dk}^m = s_{dk}^m \hat{w}_{dk}^m$
    - The factor- and view-wise sparsity: ARD prior (Identify which factor is active in which view.)
      - Disentangle the sources of variability among different assays.
    - The feature-wise sparsity (second layer): spike-and-slab prior (Result in a small number of features activate.)
      - This relies on the assumption that biological sources of variability are typically sparse.

$$p\left(\hat{w}_{d,k}^m, s_{d,k}^m\right) = N\left(\hat{w}_{d,k}^m \middle| 0, \frac{1}{\alpha_k^m}\right) Ber\left(s_{d,k}^m \middle| \theta_k^m\right)$$

      - Hyper-prior:

$$p(\theta_k^m) = Beta(\theta_k^m | a_0^\theta, b_0^\theta),\ p(\alpha_k^m) = Gamma(\alpha_k^m | a_0^\alpha, b_0^\alpha)$$

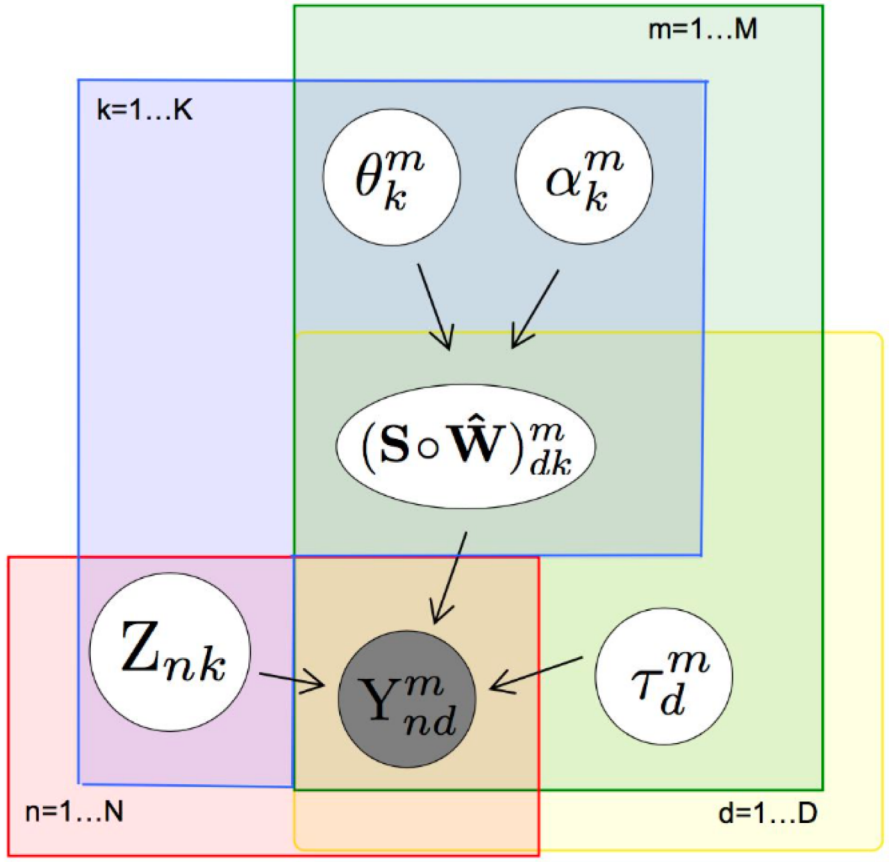
Methods

- Models for Gaussian data:
  - Hyper-parameter: $a_0^\theta, b_0^\theta = 1, \quad a_0^\alpha, b_0^\alpha = 1e^{-14}$ → Uninformative prior → Inactivated factor
  - Meaning of the parameter:
    - θ_k^m close to 0 (1) → Most of the weights of factor k in view m are shrunk to 0. → A (non) sparse factor.
    - α_k^m in $\boldsymbol{\alpha}_{m\times k}$ has four types:
      - Factors that do not explain variation in any data set (inactive factors) :
        - All values in the corresponding columns of α are large. → Rremove factor
      - Factors that explain variation in all data sets (fully shared factors) :
        - All M values in the corresponding columns of α are small.
      - Factors that explain variation in a single data set (unique factors) :
        - All values in the corresponding columns of α are very large, except one.
      - Factors that explain variation in a subset of data sets (partially shared factors) :
        - Some values in the corresponding columns of α are very large whereas others are small.
    - θ_k^m: Feature-wise sparsity level of factor k in view m.
    - α_k^m: The control of the strength of factor k in view m.

Methods

- $Likelihood \times Prior$ (Gaussian data)

$$p(\mathbf{Y}, \hat{\mathbf{W}}, \mathbf{S}, \mathbf{Z}, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \prod_{m=1}^{M} \prod_{n=1}^{N} \prod_{d=1}^{D_m} \mathcal{N}\left(y_{nd}^m \mid \sum_{k=1}^{K} s_{dk}^m \hat{w}_{dk}^m z_{nk}, 1/\tau_d\right)$$



$$\prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}\left(\hat{w}_{dk}^m \mid 0, 1/\alpha_k^m\right) \mathrm{Ber}(s_{d,k}^m \mid \theta_k^m)$$

$$\prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}\left(z_{nk} \mid 0, 1\right)$$

$$\prod_{m=1}^{M} \prod_{k=1}^{K} \mathrm{Beta}\left(\theta_k^m \mid a_0^\theta, b_0^\theta\right)$$

$$\prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}\left(\alpha_k^m \mid a_0^\alpha, b_0^\alpha\right)$$

$$\prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_d^m \mid a_0^\tau, b_0^\tau\right).$$

Methods

- $Posterior$ – model Inference for Gaussian data: Variational Bayes
  - Intention: The posterior distribution has a highly complex form whose information is not analytically tractable.
  - A deterministic approximation for the true posterior.

  - $\boldsymbol{Y}$: The set of all observed data. $\boldsymbol{Y} = \{\mathbf{y_1}, \mathbf{y_2}, \dots, \mathbf{y_N}\}$
  - $\boldsymbol{X}$: The set of all latent variables and parameters. $\boldsymbol{X} = \{\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_N}\}$
  - $\ln p(\boldsymbol{X})$: Model evidence

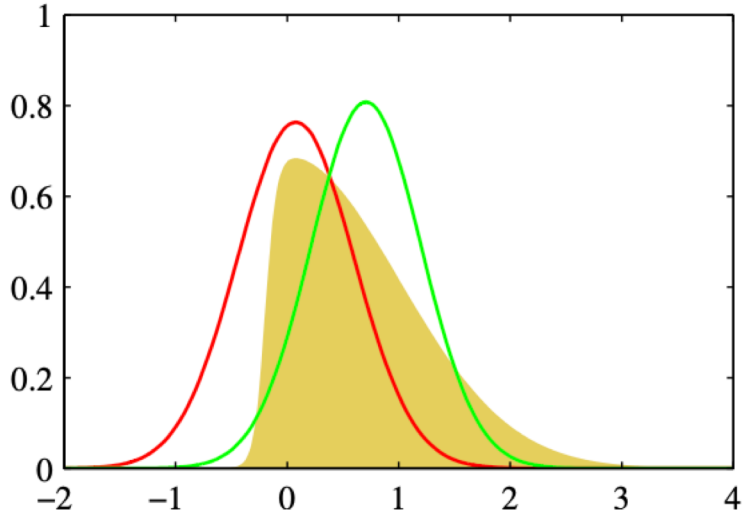
  - ELBO (Evidence Lower Bound, $\mathcal{L}(q)$):
    - $\ln p(\boldsymbol{X}) = \int q(\boldsymbol{X}) \cdot lnp(\boldsymbol{Y})d\boldsymbol{X} = \int q(\boldsymbol{X}) \cdot \ln\left(\frac{p(\boldsymbol{Y},\boldsymbol{X})}{q(\boldsymbol{X})} \cdot \frac{q(\boldsymbol{X})}{p(\boldsymbol{Y},\boldsymbol{X})} \cdot p(\boldsymbol{Y})\right) d\boldsymbol{X}$

    $= \int q(\boldsymbol{X}) \cdot \left(\ln\left(\frac{p(\boldsymbol{Y},\boldsymbol{X})}{q(\boldsymbol{X})}\right) - ln\left(\frac{p(\boldsymbol{Y}|\boldsymbol{X})}{q(\boldsymbol{X})}\right)\right) d\boldsymbol{X} = \mathcal{L}(q) + KL(q||p)$
  - If $q(\boldsymbol{X}) = p(\boldsymbol{Y}|\boldsymbol{X})$, $\mathcal{L}(q)$ attains its maximum $\ln p(\boldsymbol{X})$. (However, $p(\boldsymbol{Y}|\boldsymbol{X})$ is not tractable.)

Methods

- $Posterior$ – model Inference for Gaussian data: Variational Bayes
  - Goal: In the set of the tractable distribution for $q(\boldsymbol{X})$ (or q), maximize $\mathcal{L}(q)$ (or minimize $KL(q||p)$) as possible so that $q(\boldsymbol{X})$ (or q) can provide a good approximation to $p(\boldsymbol{Y}|\boldsymbol{X})$.
  - Find the general form for the set of the tractable distribution for $q(\boldsymbol{X})$ (or q).
  - Mean-field theory assumption
    - Partition the elements of $\boldsymbol{X}$ into M disjoint groups, then

$$q(\boldsymbol{X}) = \prod_{i=1}^{M} q_i(\boldsymbol{x}_i)$$



  - Maximize $\mathcal{L}(q)$ with respect to each of the latent variables and parameters in turn.
    - Dissect out the dependence on one of the latent variables and parameters $q_j(\boldsymbol{X}_j)$. (Keep the $\{q_{i\neq j}\}$ fixed, maximize $\mathcal{L}(q)$ respect to $q_j(\boldsymbol{X}_j)$).

$$\mathcal{L}(q) = \int \prod_{i=1}^{M} q_i(\boldsymbol{x}_i)\,\{lnp(\boldsymbol{Y},\boldsymbol{X}) - \sum_{i=1}^{M} lnq_i(\boldsymbol{x}_i)\}d\boldsymbol{X} = \int q_j \left\{ \int lnp(\boldsymbol{Y},\boldsymbol{X}) \prod_{i\neq j} q_i\, d\boldsymbol{x}_i \right\} d\boldsymbol{x}_j - \int q_j lnq_j d\boldsymbol{x}_j + const$$

$$= \int q_j \ln\tilde{p}(\boldsymbol{Y},\boldsymbol{x}_j)d\boldsymbol{x}_j - \int q_j lnq_j d\boldsymbol{x}_j + const = -KL(q_j(\boldsymbol{x}_j)||\tilde{p}(\boldsymbol{Y},\boldsymbol{x}_j)) \text{ , where } \ln\tilde{p}(\boldsymbol{Y},\boldsymbol{x}_j) = E_{q_{i\neq j}}[\ln p(\boldsymbol{Y},\boldsymbol{X})] + const\,.$$

Methods

- $Posterior$ – model Inference for Gaussian data: Variational Bayes
  - Maximize $\mathcal{L}(q) \Leftrightarrow$ Minimize $KL(q_j(\boldsymbol{x}_j)||\tilde{p}(\boldsymbol{Y}, \boldsymbol{x}_j))$, → $q_j(\boldsymbol{x}_j) = \tilde{p}(\boldsymbol{Y}, \boldsymbol{x}_j)$
  - The optimal solution $\hat{q}_j(\boldsymbol{x}_j)$ satisfy

$$\ln\left(\hat{q}_j(\boldsymbol{x}_j)\right) = E_{q_{i\neq j}}[\ln p(\boldsymbol{Y}, \boldsymbol{X})] + const$$

$$\hat{q}_j(\boldsymbol{x}_j) = \frac{\exp\left\{E_{q_{i\neq j}}[\ln p(\boldsymbol{Y},\boldsymbol{X})]\right\}}{\int \exp\left\{E_{q_{i\neq j}}[\ln p(\boldsymbol{Y},\boldsymbol{X})]\right\} d\boldsymbol{x}_j}$$

  - The optimum $\hat{q}_j(\boldsymbol{x}_j)$ depends on expectations computed with respect to the other factors $q_i(\boldsymbol{x}_i)$ for $i \neq j$. → An iterative algorithm until the change in the ELBO is small enough.

- A MOFA simplified example for Variational Bayes Algorithm.
  - Denote the observation $\mathbf{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{N}\} \sim \mathrm{N}\left(\mu, \frac{1}{\tau}\right)$, $\quad \mu \sim N\left(0, \frac{1}{\lambda_0 \tau}\right)$, $\quad \tau \sim Gamma(a_0^\tau,\ b_0^\tau)$
  - Goal: Infer the posterior for μ and τ.
  - $p(\boldsymbol{Y}|\boldsymbol{X}) = p(Y|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{\frac{N}{2}} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(y_n - \mu)^2\right\}$
  - By mean-filed theory, $q(\boldsymbol{X}) = \prod_i q_i(\boldsymbol{x}_i) = q_\mu(\mu) q_\tau(\tau)$

Methods

- A simplified example for Variational Bayes Algorithm.
  - $ln\hat{q}_\mu(\mu) = E_\tau[lnp(Y|\mu,\tau) + lnp(\mu|\tau) + lnp(\tau)] + const.$

    $= -\frac{E(\tau)}{2}\{\sum_{n=1}^{N}(y_n - \mu)^2 + \lambda_0\mu^2\} + const.$

    $\rightarrow N\left(\mu\middle|\mu_N, \frac{1}{\lambda_N}\right), \mu_N = \frac{N\bar{Y}}{\lambda_0+N}, \lambda_N = (\lambda_0 + N)E(\tau)$

  - $ln\hat{q}_\tau(\tau) = E_\mu[lnp(Y|\mu,\tau) + lnp(\mu|\tau)] + lnp(\tau) + const.$

    $= (a_0^\tau - 1)ln\tau - b_0^\tau + \frac{N}{2}ln\tau - \frac{\tau}{2}E_\mu\{\sum_{n=1}^{N}(y_n - \mu)^2 + \lambda_0\mu^2\} + const.$

    $\rightarrow Gamma(\tau|a_N, b_N), a_N = a_0^\tau + \frac{N}{2}, b_N = b_0^\tau + \frac{1}{2}E_\mu\{\sum_{n=1}^{N}(y_n - \mu)^2 + \lambda_0\mu^2\}$

  - Continuous iteration until the change in the ELBO is small enough.

    $\mathcal{L}(q) = \int q(\boldsymbol{X})\left(\frac{\ln p(\boldsymbol{Y},\boldsymbol{X})}{q(\boldsymbol{X})}\right)d\boldsymbol{X} = E_q \ln p(\boldsymbol{Y}|\boldsymbol{X}) + \sum_i \left(E_q lnp(\boldsymbol{X}_i) - E_q lnq(\boldsymbol{X}_i)\right),$

    where the expectation is under the variational distribution of the current step.

  - Since $q(\boldsymbol{X}_i)$ is known, $\mathcal{L}(q)$ is known, and it can be used to evaluate the convergence result.

Methods

- Posterior Inference for MOFA model
  - By using the mean-filed theory (except for the re-parametrization part $w_{dk}^m = s_{dk}^m \hat{w}_{dk}^m$),

$$q(\mathbf{Z}, \mathbf{S}, \hat{\mathbf{W}}, \boldsymbol{\alpha}, \boldsymbol{\tau}, \boldsymbol{\theta}) = q(\mathbf{Z}) q(\boldsymbol{\alpha}) q(\boldsymbol{\theta}) q(\boldsymbol{\tau}) q(\mathbf{S}, \hat{\mathbf{W}})$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} q(z_{n,k}) \prod_{m=1}^{M} \prod_{k=1}^{K} q(\alpha_k^m) q(\theta_k^m) \prod_{m=1}^{M} \prod_{d=1}^{D_m} q(\tau_d^m) \prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} q(\hat{w}_{d,k}^m, s_{d,k}^m)$$

  - Update equations $q_i(x_i)$ for Gaussian data:
    - Latent Variable

$$q(\mathbf{Z}) = \prod_{k=1}^{K} \prod_{n=1}^{N} q(z_{nk}) = \prod_{k=1}^{K} \prod_{n=1}^{N} \mathcal{N}\left(z_{nk} \,|\, \mu_{z_{nk}}, \sigma_{z_{nk}}\right)$$

$$\sigma_{z_{nk}}^2 = \Big(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau_d^m \rangle \langle (s_{dk}^m \hat{w}_{dk}^m)^2 \rangle + 1 \Big)^{-1}$$

$$\mu_{z_{nk}} = \sigma_{z_{nk}}^2 \sum_{m=1}^{M} \sum_{d=1}^{D_m} \langle \tau_d^m \rangle \langle s_{dk}^m \hat{w}_{dk}^m \rangle \Big(y_{nd}^m - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \langle z_{nj} \rangle \Big)$$

    - Noise precision

$$q(\boldsymbol{\tau}) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} q(\tau_d^m) = \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{G}(\tau_d^m | \hat{a}_{md}^\tau, \hat{b}_{md}^\tau)$$

$$\hat{a}_{md}^\tau = a_0^\tau + \frac{N}{2}$$

$$\hat{b}_{md}^\tau = b_0^\tau + \frac{1}{2} \sum_{n=1}^{N} \langle (y_{nd}^m - \sum_{k}^{K} \hat{w}_{dk}^m s_{dk}^m z_{n,k})^2 \rangle$$

Methods

- Posterior Inference for MOFA model
  - Update equations $q_i(x_i)$ for Gaussian data:
    - ARD precision

$$q(\boldsymbol{\alpha}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}(\alpha_k^m | \hat{a}_{mk}^{\alpha}, \hat{b}_{mk}^{\alpha})$$

$$\hat{a}_{mk}^{\alpha} = a_0^{\alpha} + \frac{D_m}{2}$$

$$\hat{b}_{mk}^{\alpha} = b_0^{\alpha} + \frac{\sum_{d=1}^{D_m} \langle (\hat{w}_{d,k}^m)^2 \rangle}{2}$$

    - Spike and Slab sparsity parameter

$$q(\theta) = \prod_{m=1}^{M} \prod_{k=1}^{K} \text{Beta}(\theta_k^m | \hat{a}_{mk}^{\theta}, \hat{b}_{mk}^{\theta})$$

$$\hat{a}_{mk}^{\theta} = \sum_{d=1}^{D_m} \langle s_{dk}^m \rangle + a_0^{\theta}$$

$$\hat{b}_{mk}^{\theta} = b_0^{\theta} - \sum_{d=1}^{D_m} \langle s_{dk}^m \rangle + D_m$$

    - Spike and Slab weights

$$q(\hat{w}_{dk}^m | s_{dk}^m) q(s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \mid s_{dk}^m \mu_{w_{dk}^m}, s_{dk}^m \sigma_{w_{dk}^m}^2 + (1 - s_{dk}^m)/\alpha_k^m\right) (\gamma_{dk}^m)^{s_{dk}^m} (1 - \gamma_{dk}^m)^{1 - s_{dk}}$$

$$\gamma_{dk^m} = q(s_{dk} = 1) = \frac{1}{1 + \exp(-\lambda_{dk^m})}$$

$$\mu_{w_{dk}^m} = \frac{\sum_{n=1}^{N} y_{nd}^m \langle z_{nk} \rangle - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \sum_{n=1}^{N} \langle z_{nk} \rangle \langle z_{nj} \rangle}{\sum_{n=1}^{N} \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle}}$$

$$\sigma_{w_{dk}^m} = \frac{\langle \tau_d^m \rangle^{-1}}{\sum_{n=1}^{N} \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle}}$$

$$\lambda_{dk}^m = \langle \log \frac{\theta}{1 - \theta} \rangle + 0.5 \log \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle} - 0.5 \log \left(\sum_{n=1}^{N} \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle} \right) + \frac{\langle \tau_d^m \rangle}{2} \frac{\left(\sum_{n=1}^{N} y_{nd}^m \langle z_{nk} \rangle - \sum_{j \neq k} \langle s_{dj}^m \hat{w}_{dj}^m \rangle \sum_{n=1}^{N} \langle z_{nk} \rangle \langle z_{nj} \rangle \right)^2}{\sum_{n=1}^{N} \langle z_{nk}^2 \rangle + \frac{\langle \alpha_k^m \rangle}{\langle \tau_d^m \rangle}}$$

Methods

- $Posterior$ – model Inference for non-Gaussian data: Variational Lower Bounds
  - Noise model
    - For Poisson model:
      - $y_{nd}^{m} \sim Poi\left(\lambda\left(Z_{n:}W_{d:}^{T}\right)\right)$, $\lambda(x) = \log(1 + e^{x})$
    - For Bernoulli model:
      - $y_{nd}^{m} \sim \mathrm{Ber}\left(\sigma\left(Z_{n:}W_{d:}^{T}\right)\right)$, $\sigma(x) = \frac{1}{(1+e^{-x})^{-1}}$
  - Parameter inference
    - Variational lower bounds on the likelihood.
    - To be continued….

Methods

- Some Practical Considerations
  - Convergence for ELBO:
    - ELBO increase monotonically.
    - Threshold: An iterative change in ELBO smaller than 0.1%
  - Handling for missing values
    - No prior imputation.
    - Non-observed data do not intervene in the likelihood.
    - Non-observed data is ignored in the update equations.
    - In practice, a binary mask is used: $\mathcal{O}^m \in \mathbb{R}^{N \times D_m}$ for each view m.
      ($\mathcal{O}_{n,d} = 1$, feature d is observed for sample n, $\mathcal{O}_{n,d} = 0$ otherwise.)
  - Data Preprocessing (Normalization)
    - MOFA does not require the data to be centered or scaled.
    - Not centered: A constant factor of ones that will capture any feature-wise intercept effect.

      → Ensures the rest of the factors capture variation independent of the feature-wise means.
    - Not scaled: ARD prior α_k^m allows different scales of the weights for each view (m).
    - For Gaussian noise model, normalization and variance stabilization are recommended.

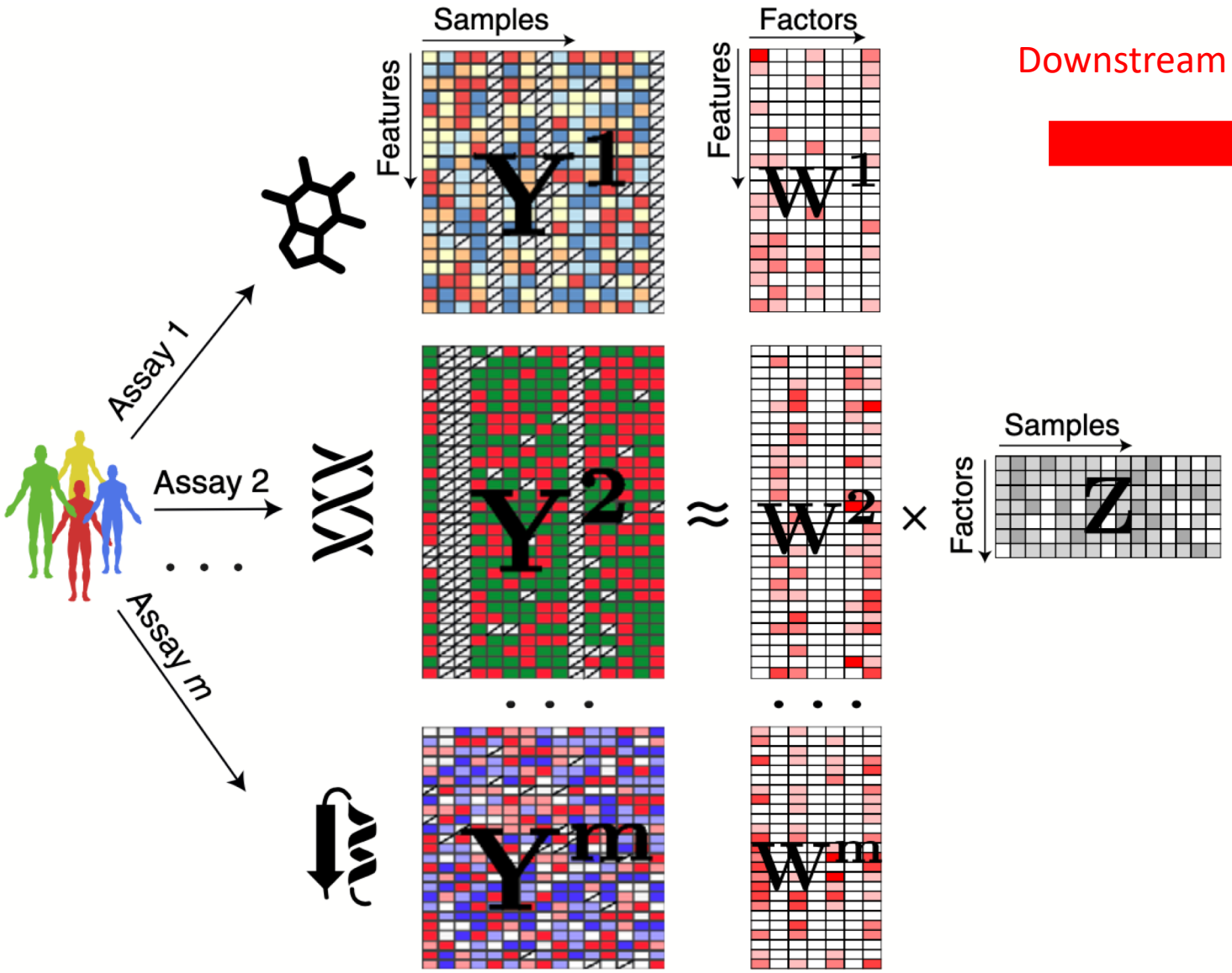
Methods

- Some Practical Considerations
  - Consistency across random parameter initializations
    - The variational Bayes is not guaranteed to find the optimal solution.
    - The estimates will depend on the parameter initialization.
    - Assess the consistency by running 10 trials under different initialization.
    - Model selection: The model with the highest ELBO.
  - Determining the number of factors

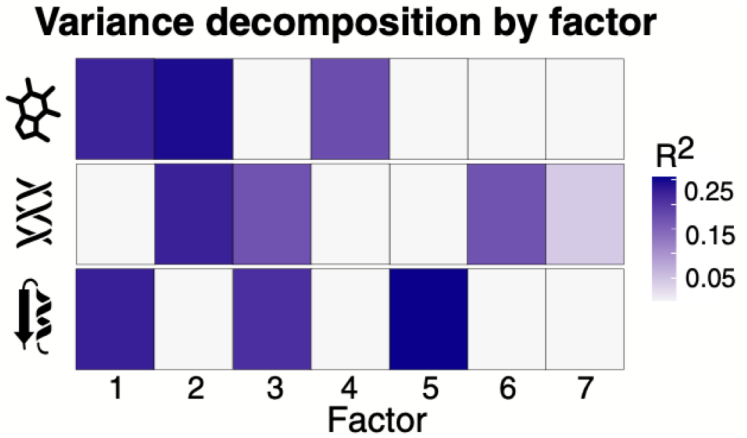
    A. User defined.

    B. Automatically pruning by an user-defined minimum fraction of variance explained threshold. (achieved by the view- and factor-wise ARD prior α_k^m.)
    - Initialized factor number: K = 25, variance explained threshold: 2%

Methods

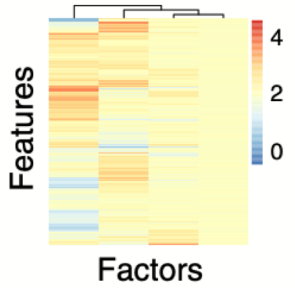


Downstream analysis

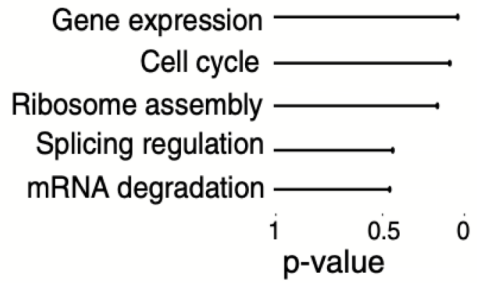


Annotation of factors

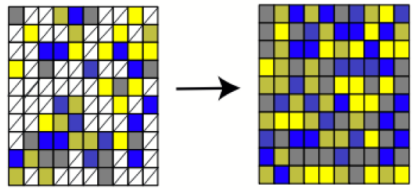
Inspection of loadings



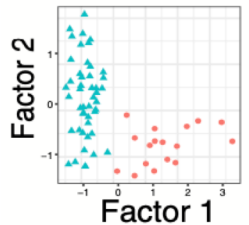
Feature set enrichment analysis



Imputation of missing values



Inspection of factors



Methods

- Downstream analysis for factor interpretation and annotation
  - Use the expectations of the model variables under the inferred posterior dist..
  - To see the the variation explained by each factor in each view,

    Define the fraction of the variance explained (R^2) by factor k in view m as

$$R^2_{m,k} = 1 - \left(\sum_{n,d} y^m_{nd} - z_{nk} w^m_{kd} - \mu^m_d\right)^2 / \left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2$$

    Define the fraction of variance explained per view taking into account all factors as

$$R^2_m = 1 - \left(\sum_{n,d} y^m_{nd} - \sum_k z_{nk} w^m_{kd} - \mu^m_d\right)^2 / \left(\sum_{n,d} y^m_{nd} - \mu^m_d\right)^2$$

    μ^m_d denotes the feature-wise mean

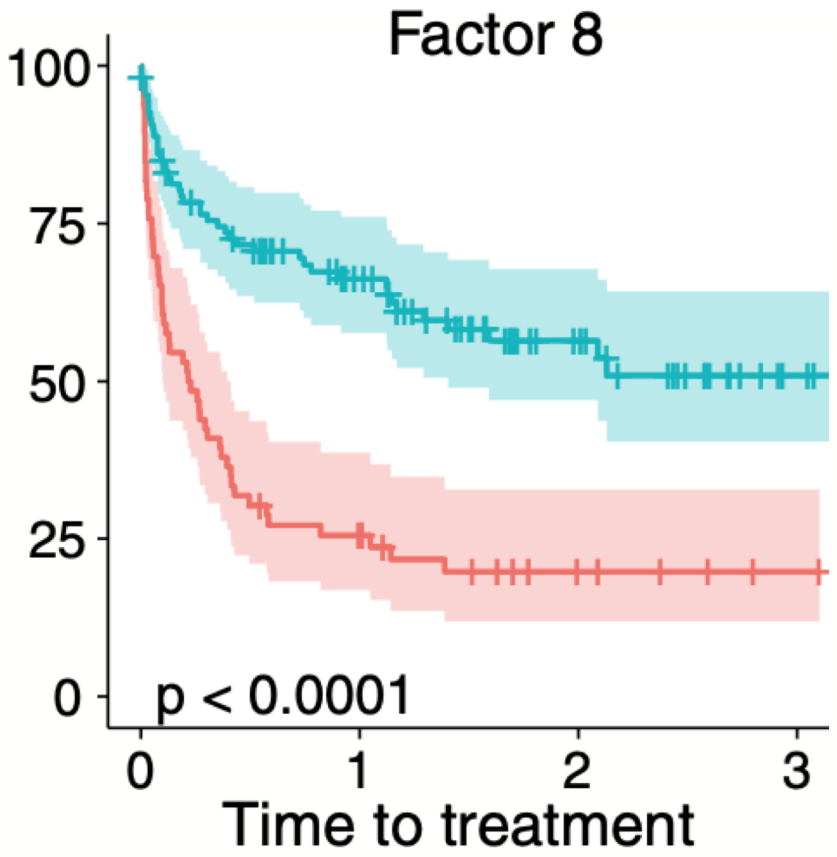
  - Each factor is characterized by following
    - Ordination of the samples in factor space → Sample heterogeneity
    - Inspection of top features with largest weight (Scale each weight vector by its absolute value)
    - Feature set enrichment analysis (t-test: test the mean for the weights of features that belong to and not belong to a set G.)

Methods

- Downstream analysis for imputation
  - In the CLL data,
    - Use the subset of samples with all measurements (N=121).
    - Masked data at random either single values or all measurements for randomly selected samples in the drug response.
    - Trained MOFA on the masked data.
    - Accuracy: MSE: masked values vs. true values.
    - Fix the number of factors for MOFA K = 10.
    - To study the dependence on K for imputation and compare MOFA to GFA, varying K = 1, …, 20 and re-ran the same masking experiments.

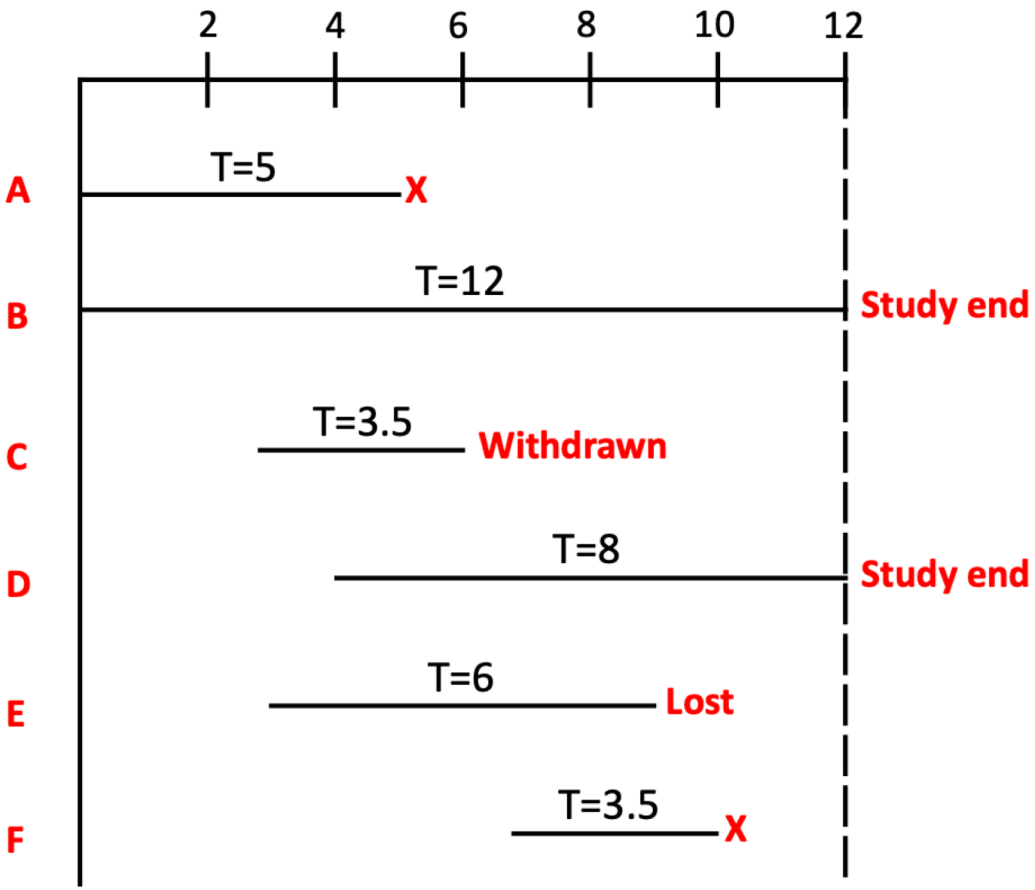
Methods

- Downstream analysis for Survival Analysis (Kaplan–Meier plots)
  - Cut-point for each factor: maximally selected rank statistics with P-values based on a log-rank test between the resulting groups.



Methods

- Censoring data



Methods

- Downstream analysis for Survival Analysis (Cox Model)
  - After the dimensional reduction (MOFA), using these new factors as covariates in the Cox model.
  - Scaled all predictors to ensure comparability of the hazard ratios.
  - N = 174, censored data: 78 samples.
  - Model Comparison:
    - PCA: 10 principal components for each single view.
    - MOFA: 10 MOFA factors.
    - The complete set of all features in a view with a ridge penalty in the Cox model
  - A stratified fivefold cross-validation scheme.
  - Evaluation: Harrel's C-index

Methods

- Methods Comparison
  - iCluster
    - Primary application: Clustering
    - It cannot drive variation in distinct subsets of views.
    - It cannot handle missing values.
    - Computationally demanding.
  - Grouped Factor Analysis (GFA)
    - Only deal with Gaussian observation data.
    - Use sparsity constraints throughout training → Maintain factors that have near-zero relevance.
    - Use Gibbs sampling.
    - Still infer for deactivate factors.
    - Time-consuming when the data has missing values.

Simulation

- Model Validation (Only MOFA)
  - Simulate data from the generative model

1. Simulation settings to validate the ability to learn the number of active factors (Appendix Figure S1)

likelihood	# factors	# features	# views	# samples	Missingness (%)
gaussian	(5,10,...,60)	5000	3	100	0
gaussian	10	(100,500,...,10000)	3	100	0
gaussian	10	5000	(1,3,...,21)	100	0
gaussian	10	5000	3	100	(0,5,10,...,90)

$K_{initial} = 100\ factors$

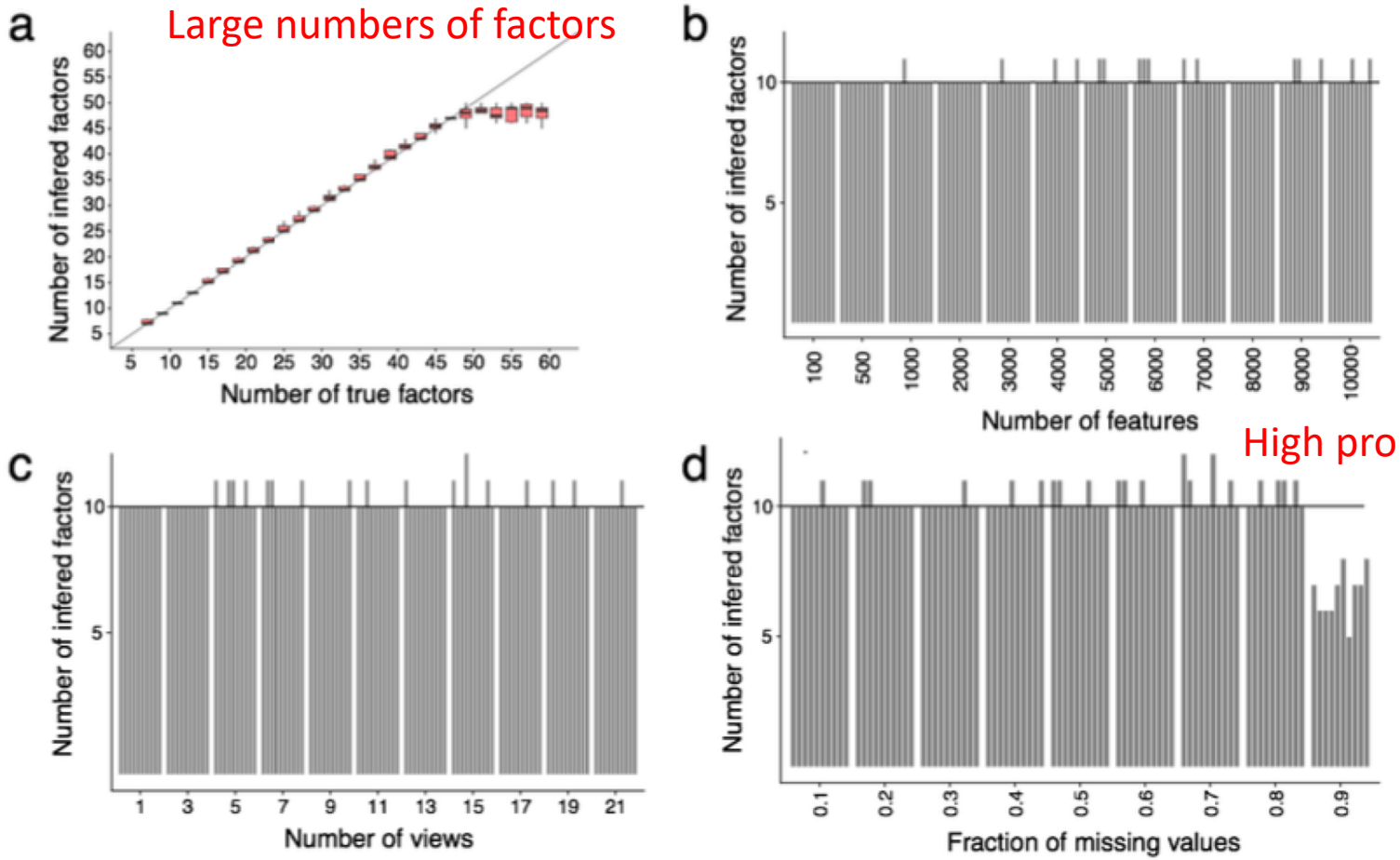
2. Simulation settings to validate non-gaussian likelihoods (Appendix Figure S2-3)

likelihood	# factors	# features	# views	# samples	Missingness (%)
bernoulli	10	5000	3	100	0
poisson	10	5000	3	100	0

- 10 repeats for every config.
- Evaluation: The estimated number of factors.

Simulation

- Results for model validation (Only MOFA)
  - Accurately reconstruct the latent dimension, except



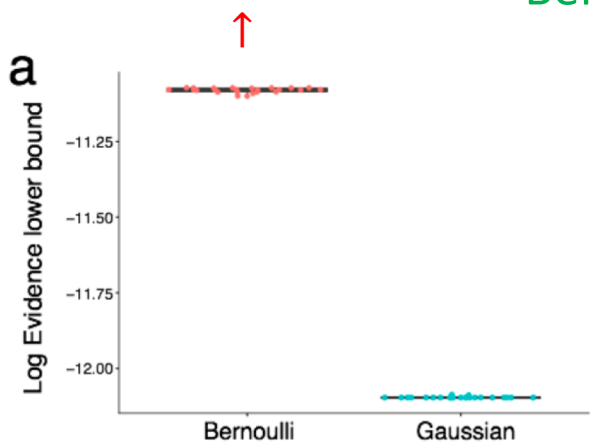
$K_{initial} = 100\ factors$

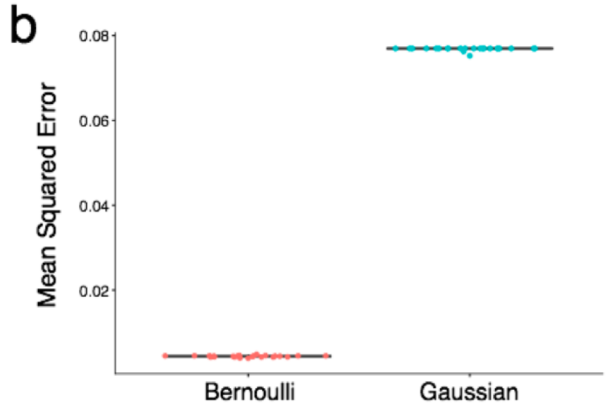
High proportions of missing values

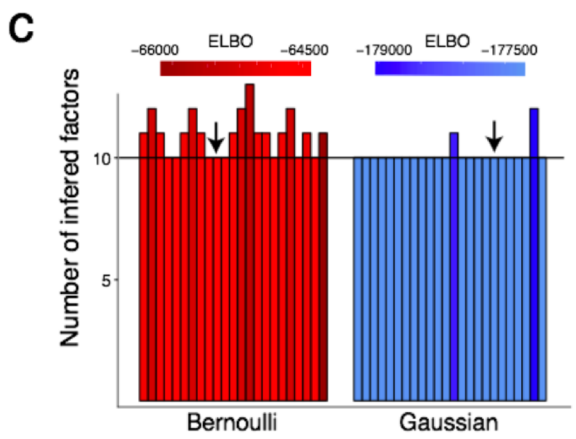
Simulation

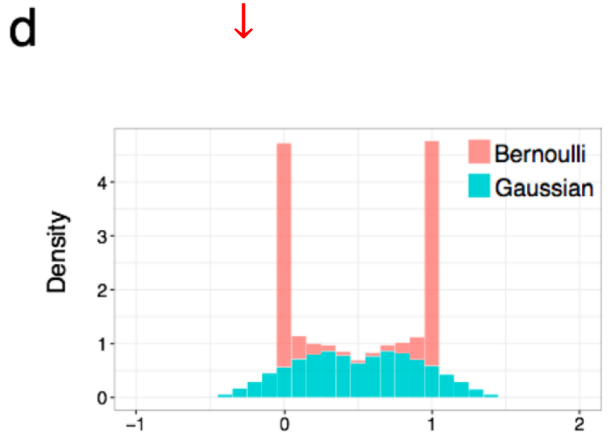
- Results for model validation (Only MOFA)
  - Models that account for non- Gaussian observations improved the fit when simulating binary or count data.

Bernoulli

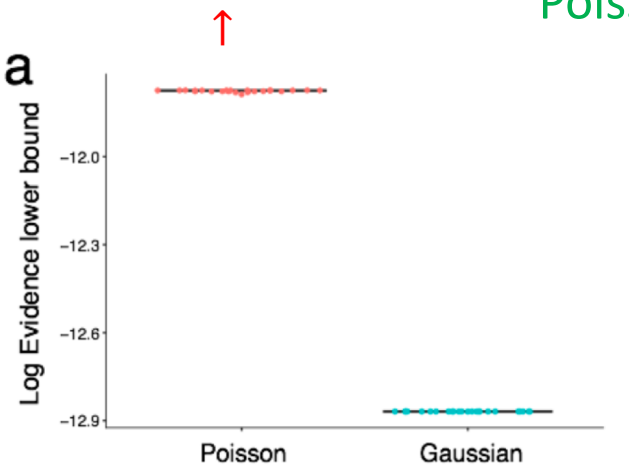


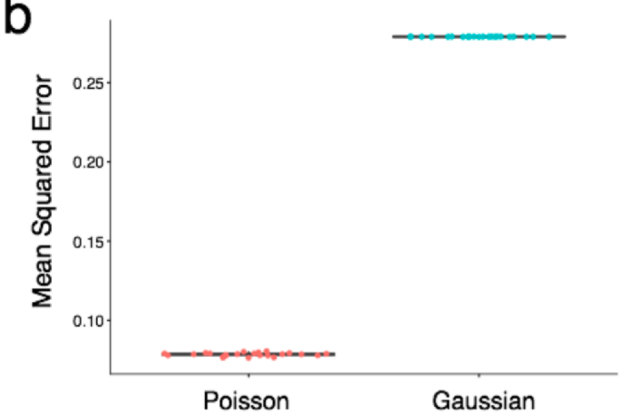


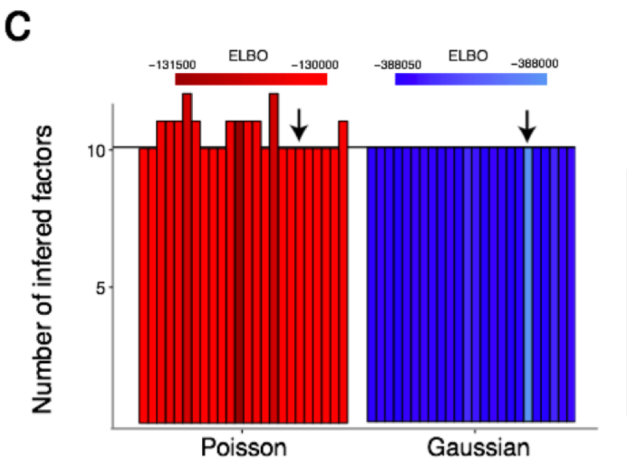


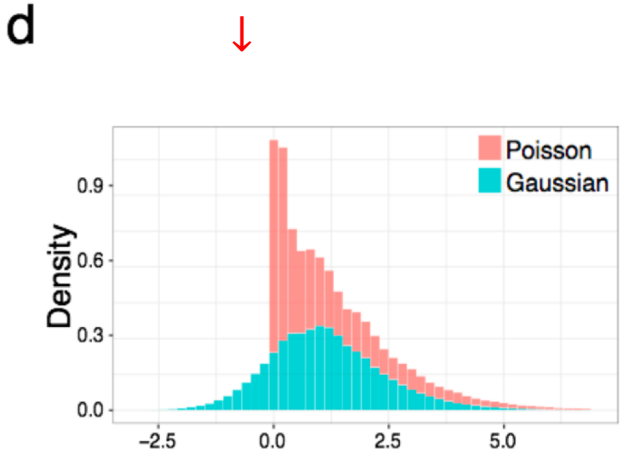


Poisson









Simulation

- Model Comparison (MOFA, GFA, iCluster)
  - Simulate data from the generative model

3. Simulations settings for the GFA and iCluster comparison (Appendix Figure S4-5)

Likelihoods	# factors	# features	# views	# samples	Missingness (%)
gaussian, bernoulli and poisson	10	5000	3	100	5

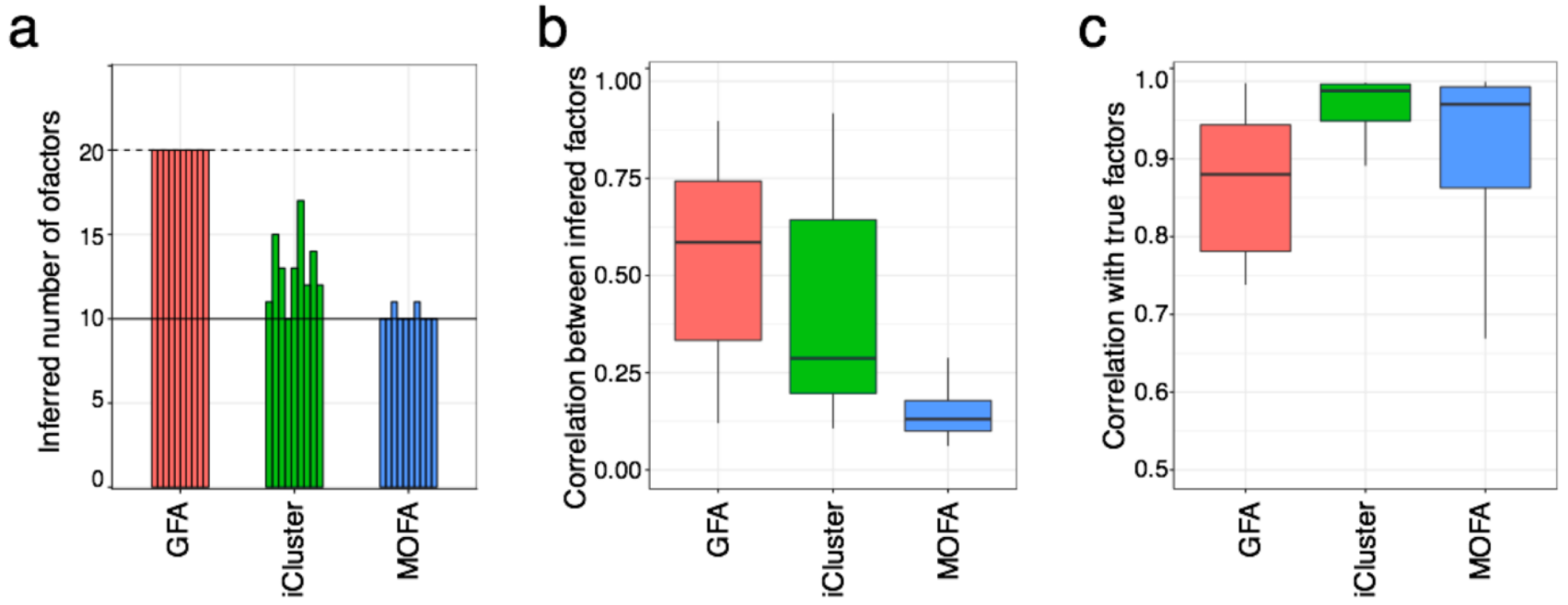
$K_{initial} = 20\ factors$

  - Evaluation
    - For the reconstruction of factor activity patterns:
      - Generate data from $K_{true} = 10\ and\ 15\ factors$.

Simulation

- Results for model comparison (MOFA, GFA, iCluster)
  - GFA and iCluster tends to infer redundant factors.

ρ(inferred factors, inferred factors) ↓



$K_{initial} = 20\ factors$

$K_{true} = 10\ factors$

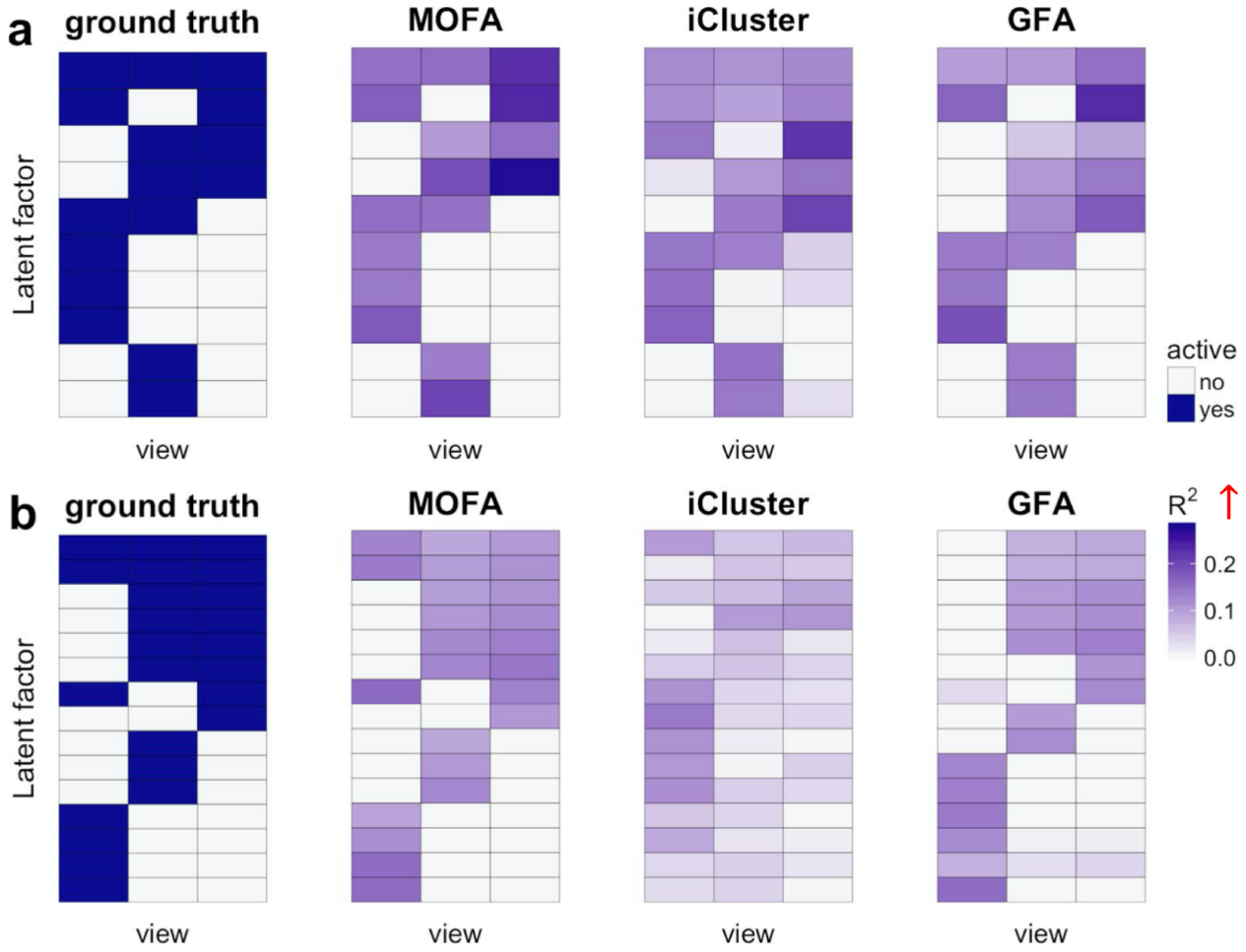
ρ(true factors, inferred factors) ↑

Simulation

- Results for model comparison (MOFA, GFA, iCluster)
  - GFA and iCluster were less accurate in recovering patterns of shared factor activity across views.

$K_{true} = 10\ factors$

$K_{true} = 15\ factors$

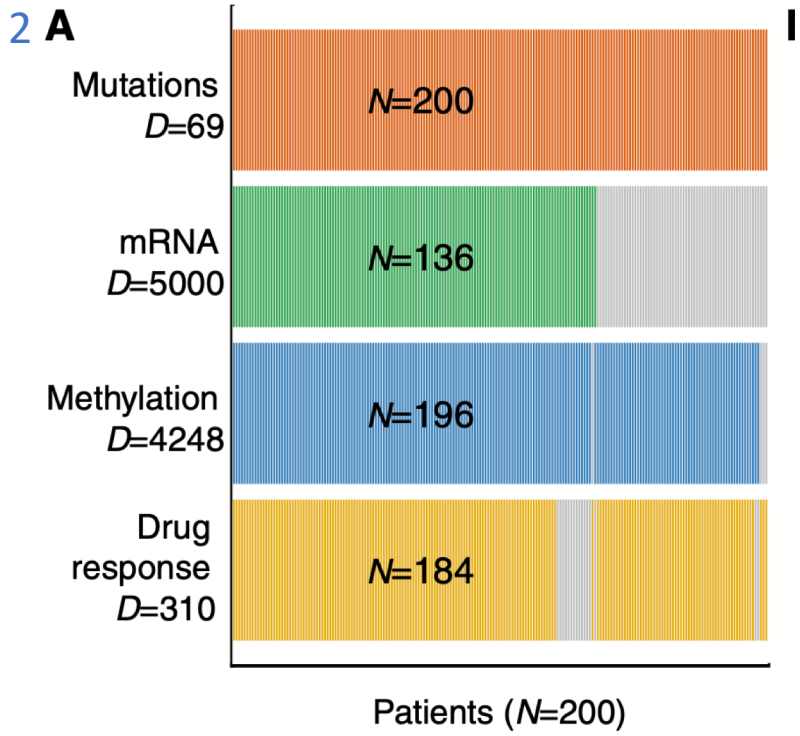


$K_{initial} = 25\ factors$

For iCluster, the true number of factors was used

Real Data Analysis – CLL

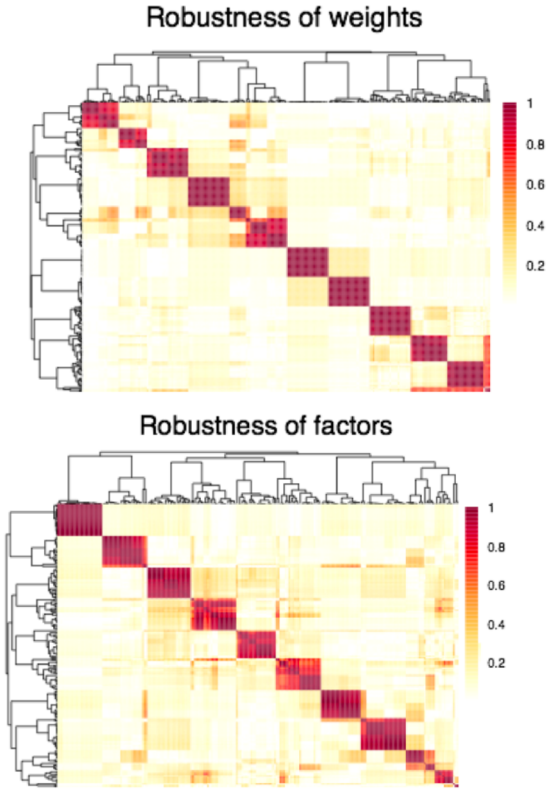
- Chronic lymphocytic leukemia (CLL)
- Data types: (Choosed)
  - Somatic mutation status. (Mutations)
  - Transcriptome profiling. (mRNA)
  - DNA methylation. (Methylation)
  - Ex vivo drug response measurements. (Drug response)
- Nearly 40% of 200 samples data is partially missing.
- Consuming Time:
  - MOFA: 25mins
  - GFA: 34hrs
  - iCluster: 5-6 days



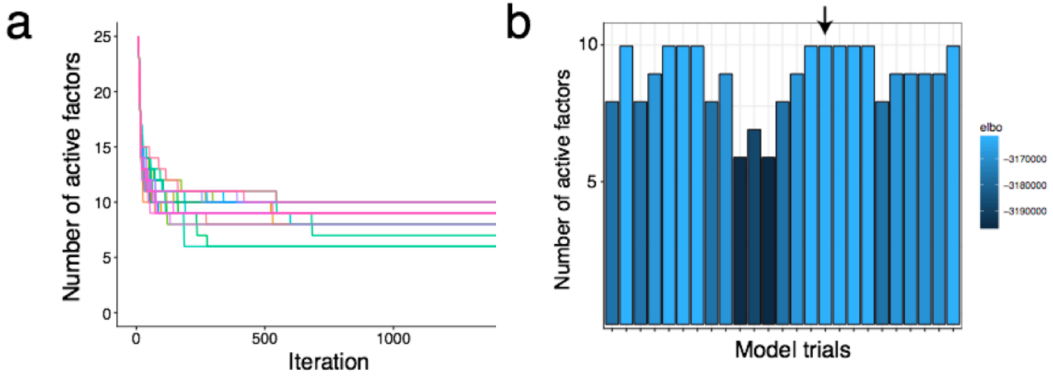
Real Data Analysis – CLL

- 10 factors are detected. (variance threshold: 2%)
- 25 trials (random initializations) are conducted.
- Factors are orthogonal. (S6: c, d)
- Robustness in algorithm initialization. (S6: a, b, c, d)
- Robustness in subsampling of the data. (S7: a, b)

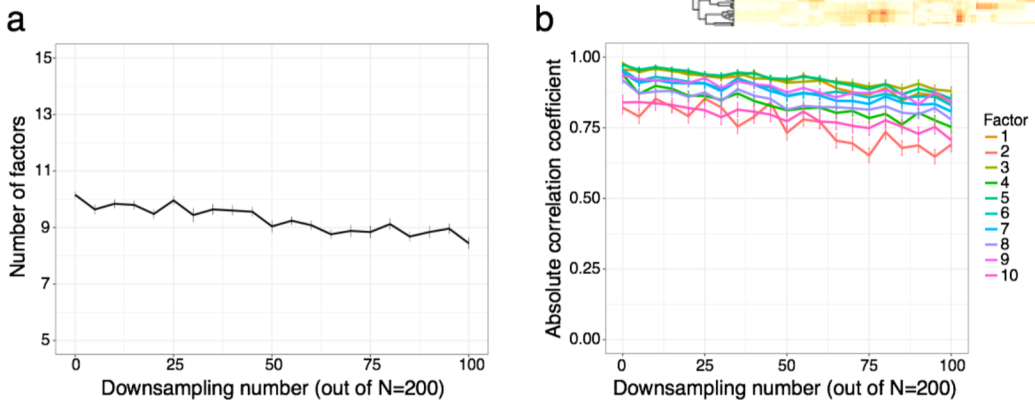
S6



S6

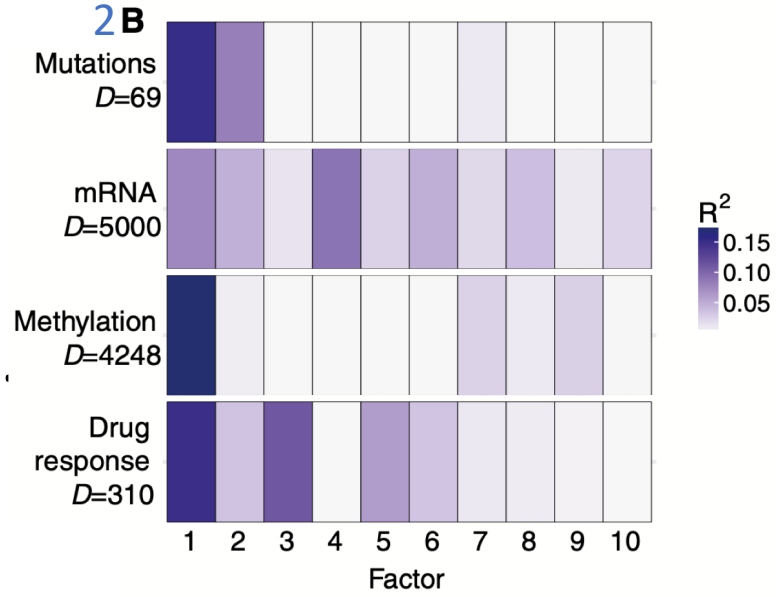


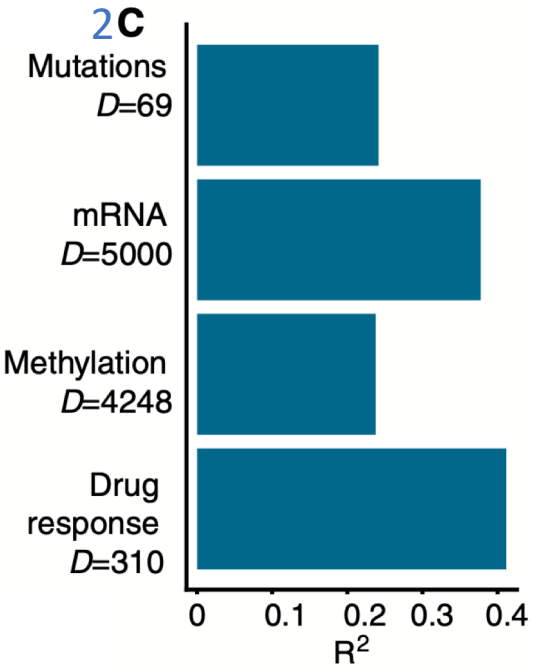
S7



Real Data Analysis – CLL

- Variance decomposition by factor (2B)
  - Multiple molecular data modalities: Factor1 and factor2.
  - Two data modalities: Factor 3 and factor 5.
  - Single data modality: Factor 4.
- Total variation explained (2C)
  - Drug response: 41%
  - mRNA: 38%
  - Methylation: 24%
  - Mutations: 24%

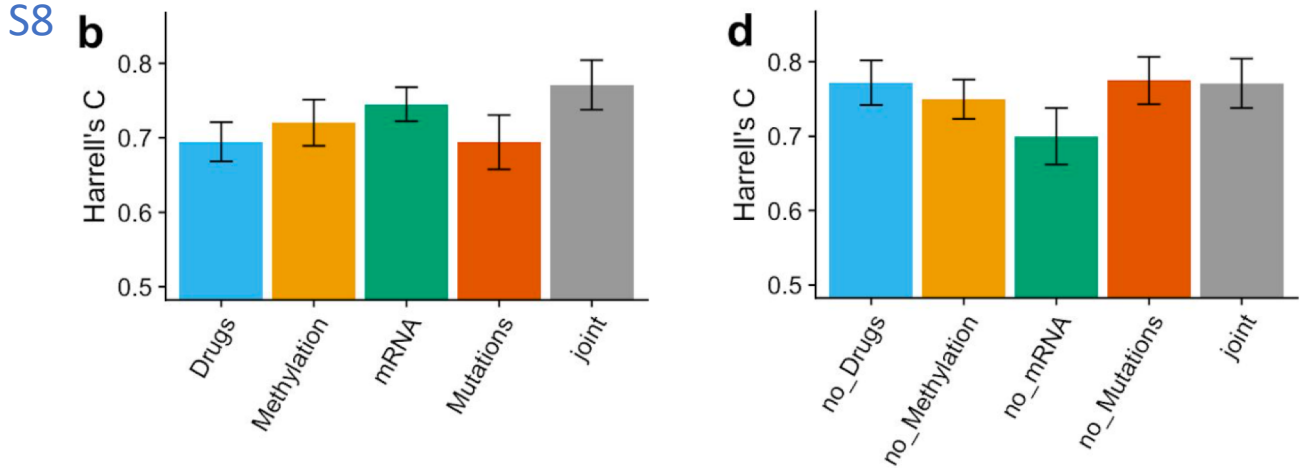




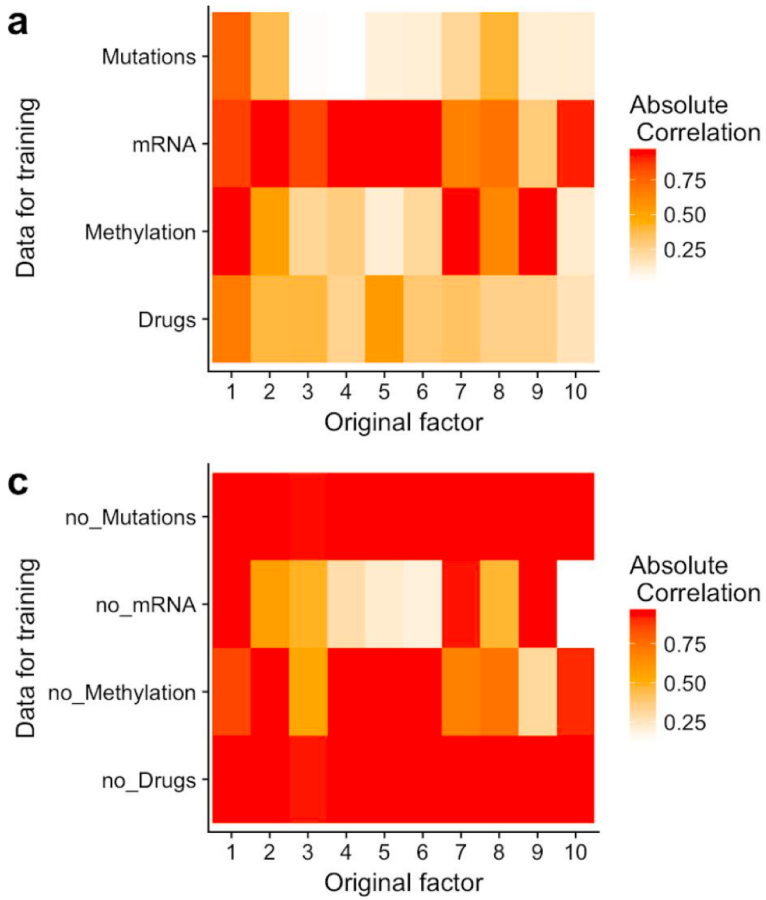
Real Data Analysis – CLL

- Factor redundancy (S8: a, b, c, d)
  - Factors activate in multiple data modalities could still be detected when excluding specific data modality.
  - S8 a, c: correlation on 121 samples (X-axis: the MOFA factors discovered on the full datasets.)
  - S8 b, d: prediction for 121 samples based on 10 factors.

S8



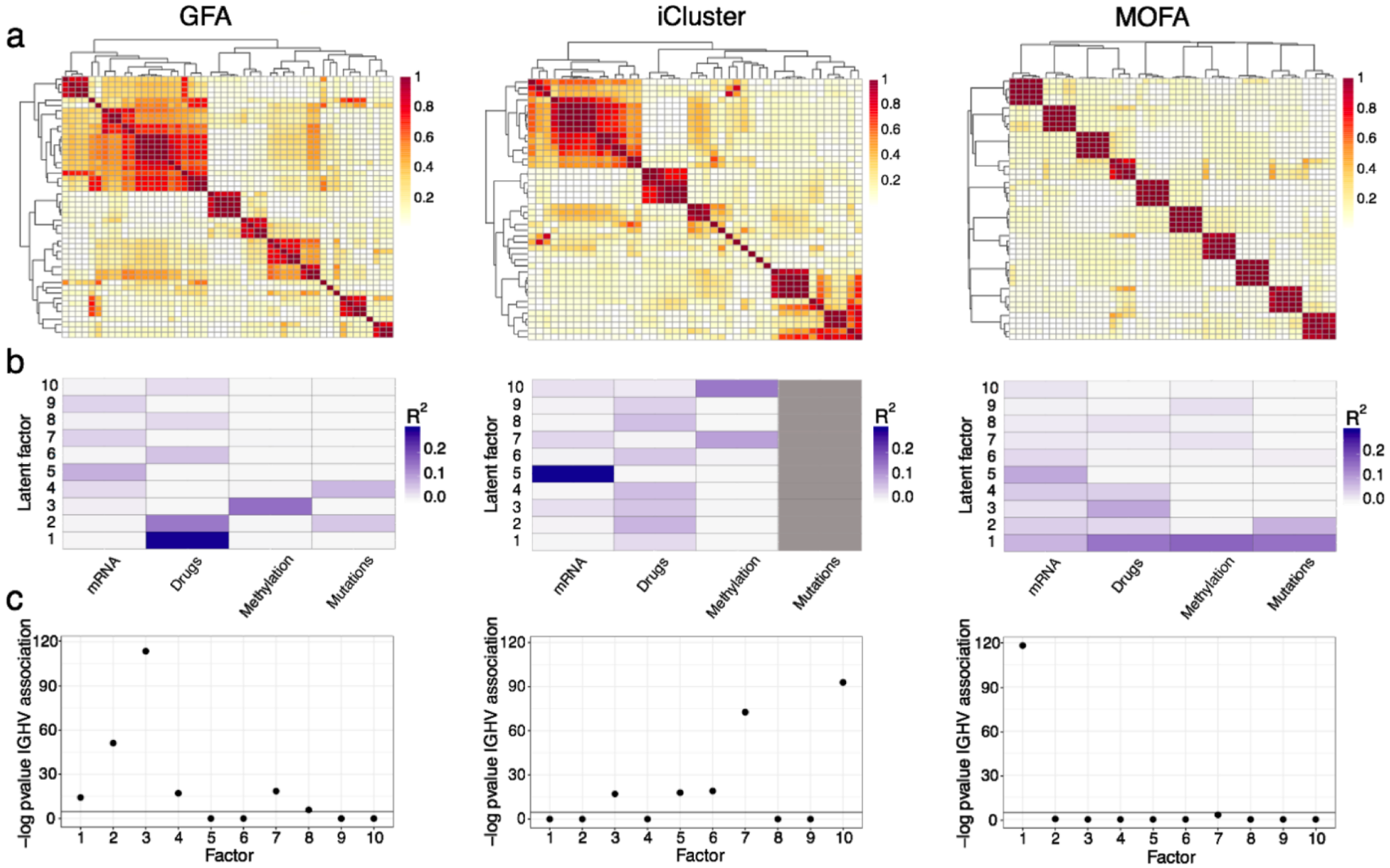
S8



Real Data Analysis – CLL

- Comparison with GFA and iCluster.
  - MOFA's result is more consistent in identifying factors. (S9)

S9



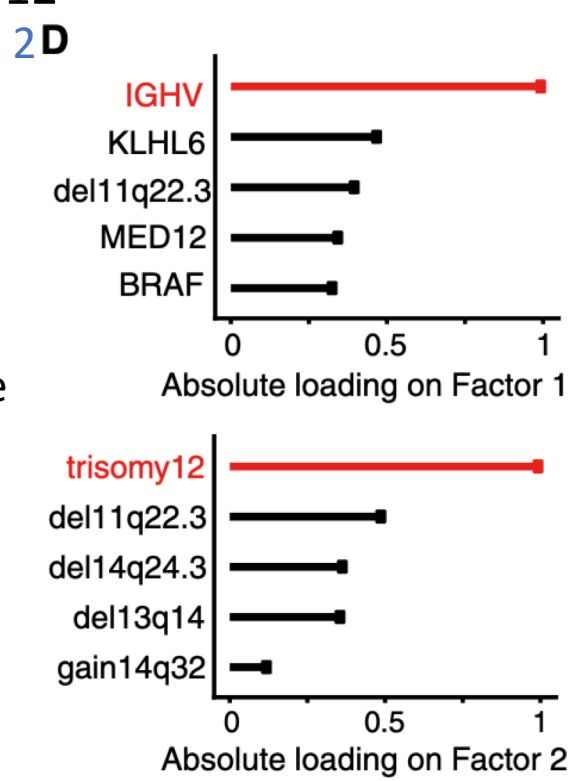
For the first trial of each model (S9 b)

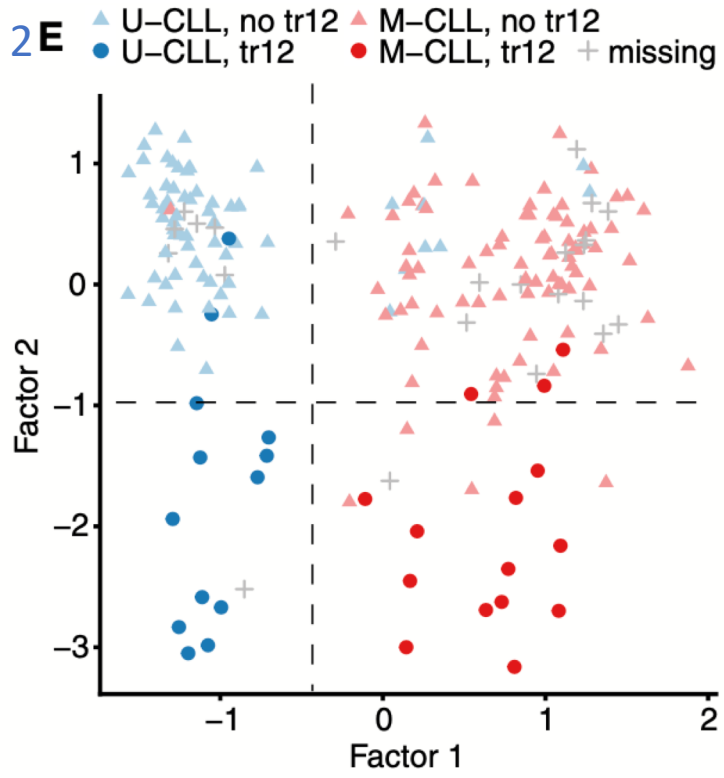
Line in S9 c: 1% FDR

Real Data Analysis – CLL

- Downstream Analysis – Annotation of factors
  - Based on the all data result, (2D, 2E)
    - Factor 1 ⇔ somatic mutation of the immunoglobulin heavy-chain variable region gene (IGHV)
    - Factor 2 ⇔ trisomy of chromosome 12
    - Both factors are aligned with two of the most important clinical markers in CLL.

- IGHV: a surrogate of the differentiation state of the tumor's cell of origin and the level of activation of the B-cell receptor.





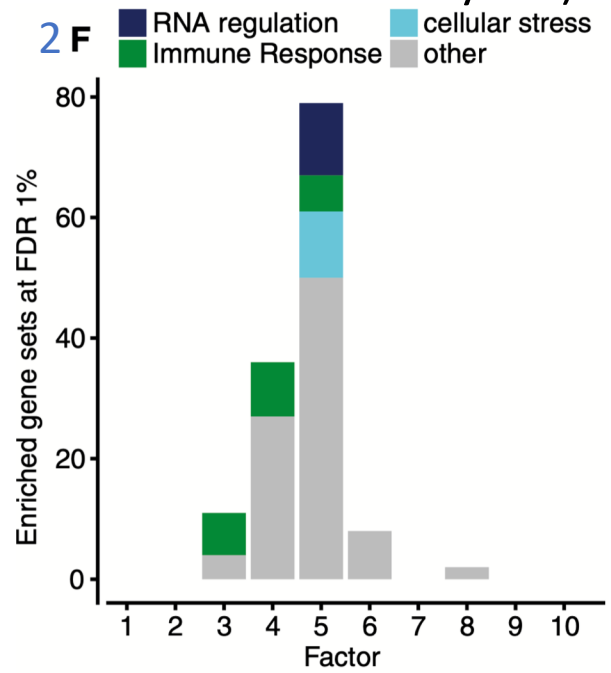
Base on the mutation data

Real Data Analysis – CLL

- Downstream Analysis – Annotation of factors
  - Based on the drug response and mRNA data result,
    - Factor 5 ⇔ A set of genes enriched for oxidative stress and senescence pathways.
  - Based on the mRNA data result only,
    - Factor 4 ⇔ Immune response pathways and T-cell receptor signaling.

      → Due to differences in cell type composition among samples.

      (Mainly all B cells, some are contaminated with T cells and monocytes.)
  - Based on the drug response data only,
    - Factor 3 ⇔ Samples' drug sensitivity.

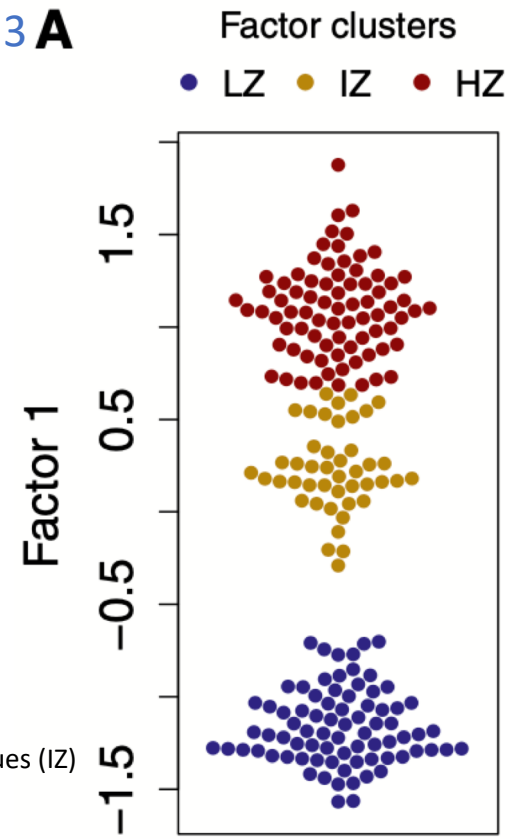


Real Data Analysis – CLL

- Downstream Analysis – Annotation of factors
  - More on the factor 1.
    - In the clinical data, IGHV status is binary.

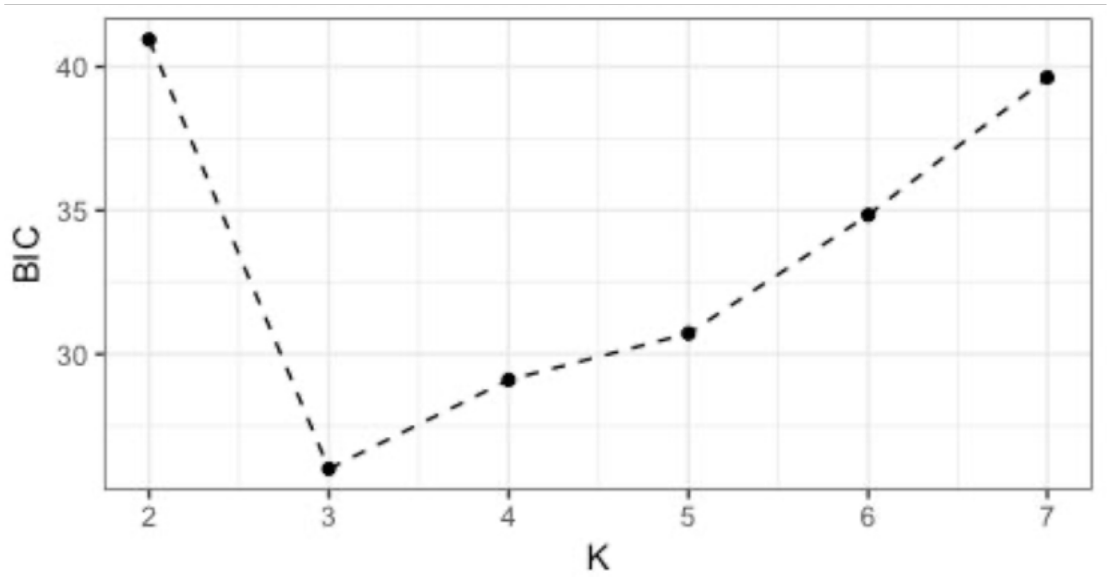
      → The value of factor 1 has a more complex substructure. (3A, S10)

3 **A**



low factor values (LZ)
Intermediate factor values (IZ)
high factor values (HZ).

S10 K-means clustering

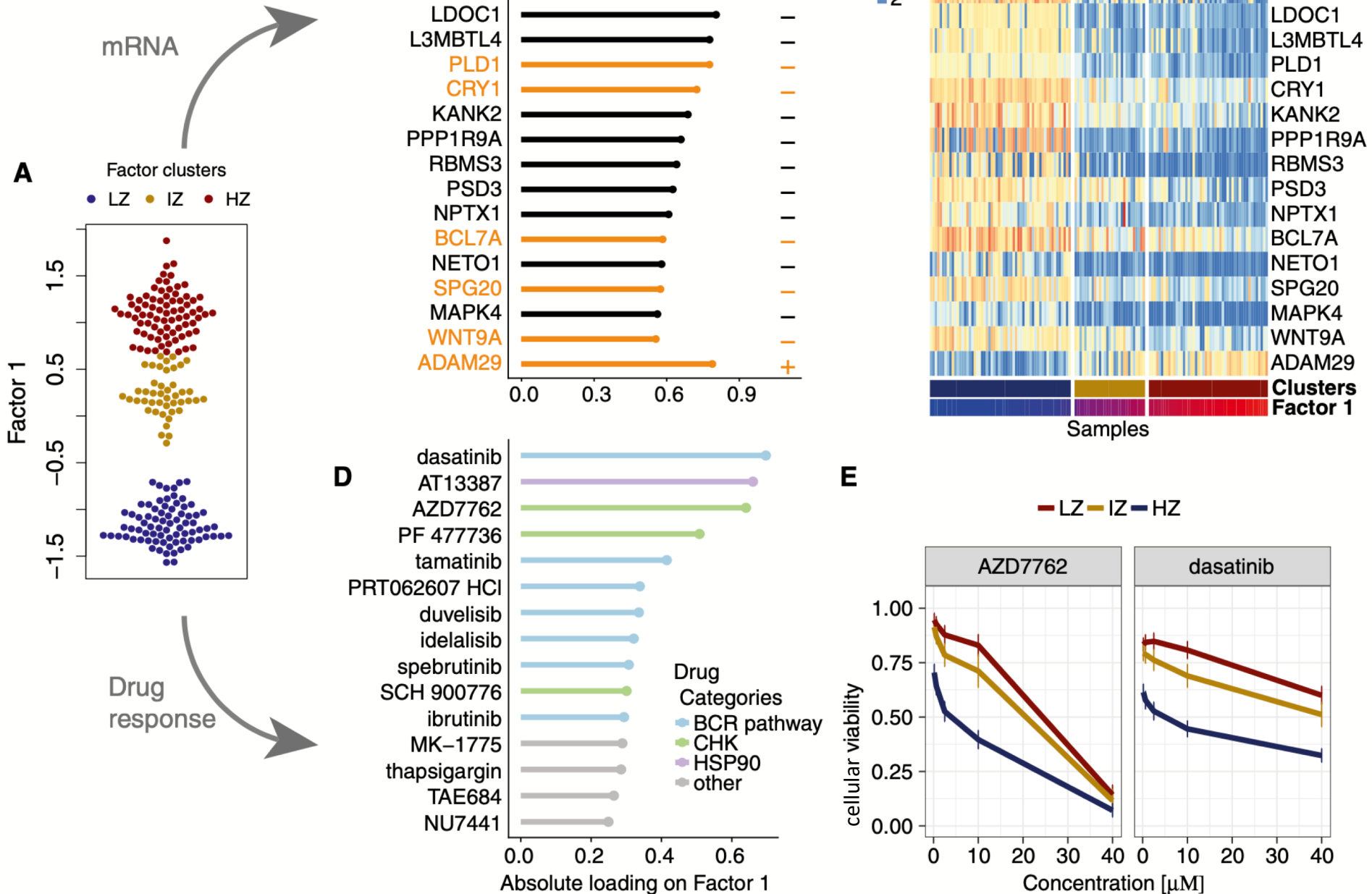


→ 3 subgroups
Consistent with other studies

Real Data Analysis – CLL

- Downstream Analysis – Annotation of factors
  - More on the factor 1.
    - MOFA connect this factor to multiple molecular layers. (S12, S13)
    - Related to the changes in the expression of genes linked to IGHV status. (3B, 3C)
    - Related to the drugs that target kinases or the downstream of the B-cell receptor pathway (BCR pathway). (3D, 3E)

Figure 3


mRNA
A
Factor clusters
LZ
IZ
HZ
Factor 1
1.5
0.5
−0.5
−1.5
Drug response
B
known IGHV-association
ZNF667
SEPT10
SOWAHC
ZNF471
LPL
LDOC1
L3MBTL4
PLD1
CRY1
KANK2
PPP1R9A
RBMS3
PSD3
NPTX1
BCL7A
NETO1
SPG20
MAPK4
WNT9A
ADAM29
0.0
0.3
0.6
0.9
C
normalized gene expression
14
12
10
8
6
4
2
Clusters
Factor 1
Samples
D
dasatinib
AT13387
AZD7762
PF 477736
tamatinib
PRT062607 HCl
duvelisib
idelalisib
spebrutinib
SCH 900776
ibrutinib
MK−1775
thapsigargin
TAE684
NU7441
Drug Categories
BCR pathway
CHK
HSP90
other
0.0
0.2
0.4
0.6
Absolute loading on Factor 1
E
LZ
IZ
HZ
AZD7762
dasatinib
cellular viability
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
Concentration [μM]

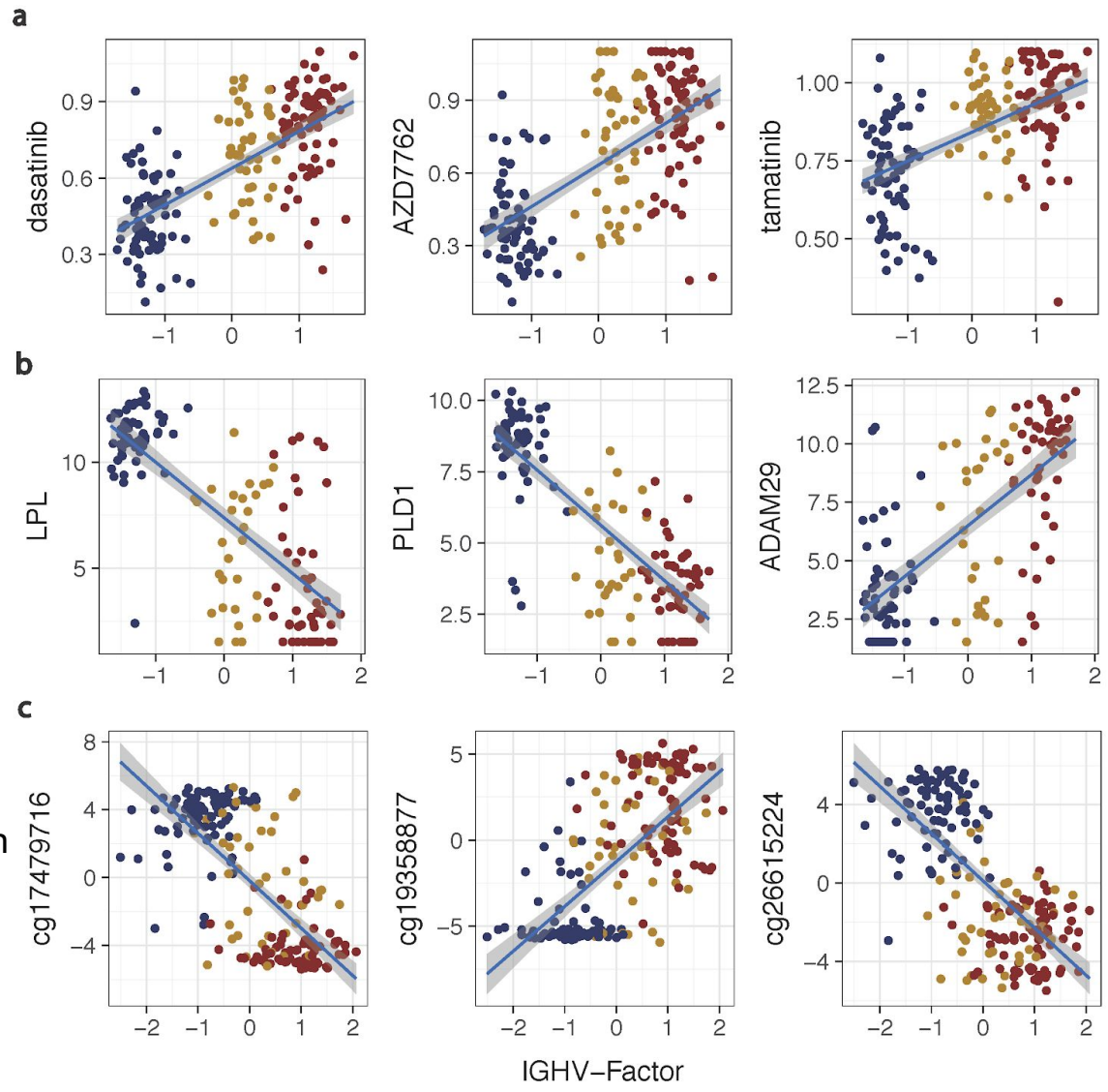
S12

$\rho(F1, individual\ molecular\ features)$

drug response

gene expression

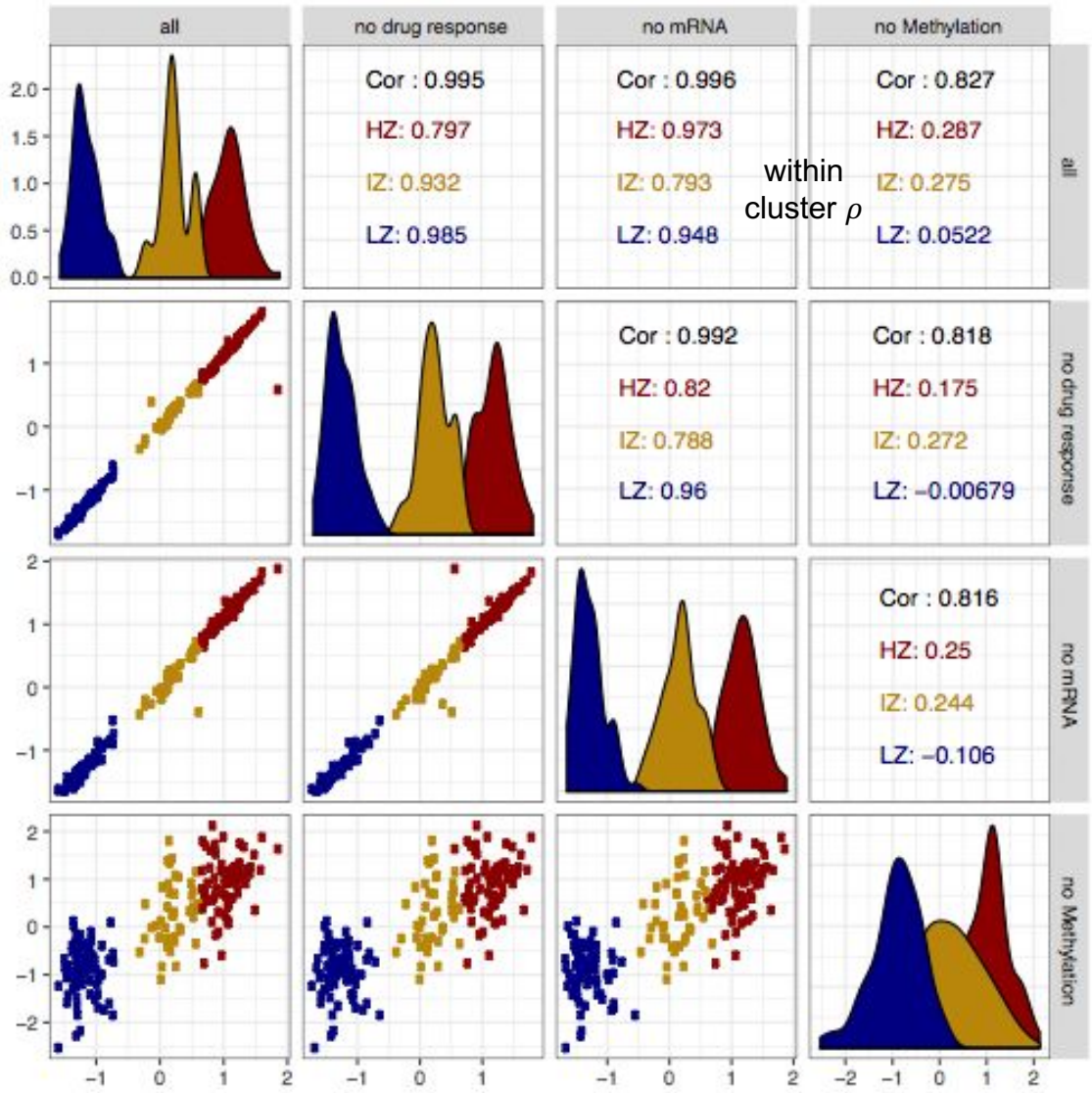
methylation



Displayed are representative features with high absolute loading on the Factor 1 from the full model.

S13

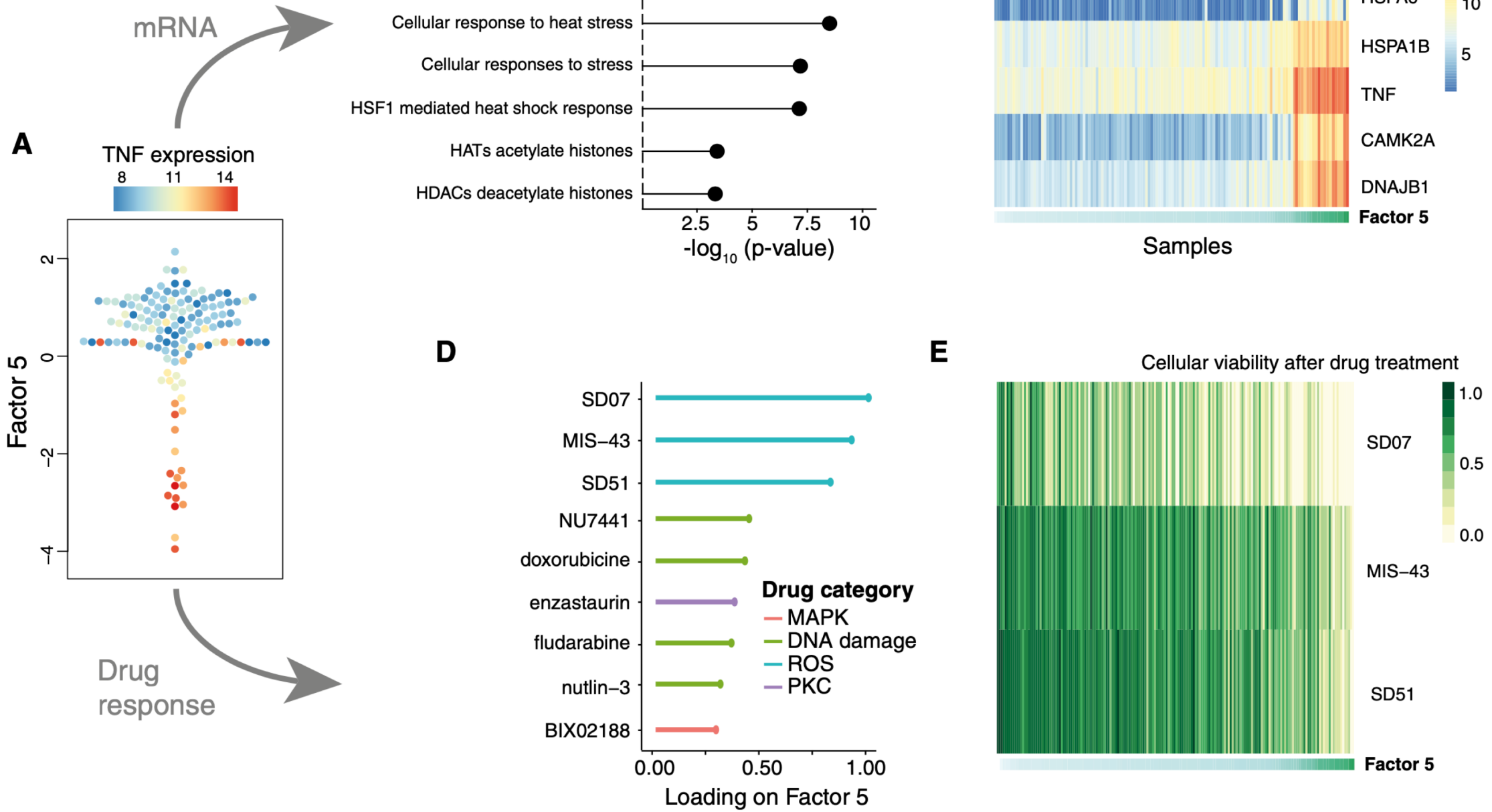
$\rho(F1\ on\ full\ data, F1\ on\ partial\ data\ modality)$



Real Data Analysis – CLL

- Downstream Analysis – Annotation of factors
  - More on the factor 5.
    - Based on the mRNA data (EV2B, EV2C)
      - The top weight pathway: heat-shock proteins (HSPs)
        - HSPs contains genes that are essential for protein folding and are up-regulated upon stress conditions.
        - Genes in HSP pathways are up-regulated in some cancers, but receive little attention in the context of CLL.
    - Based on the drug response data (EV2D, EV2E)
      - The top weight drug: oxidative stress (e.g. reactive oxygen species (ROS), DNA damage response and apoptosis).

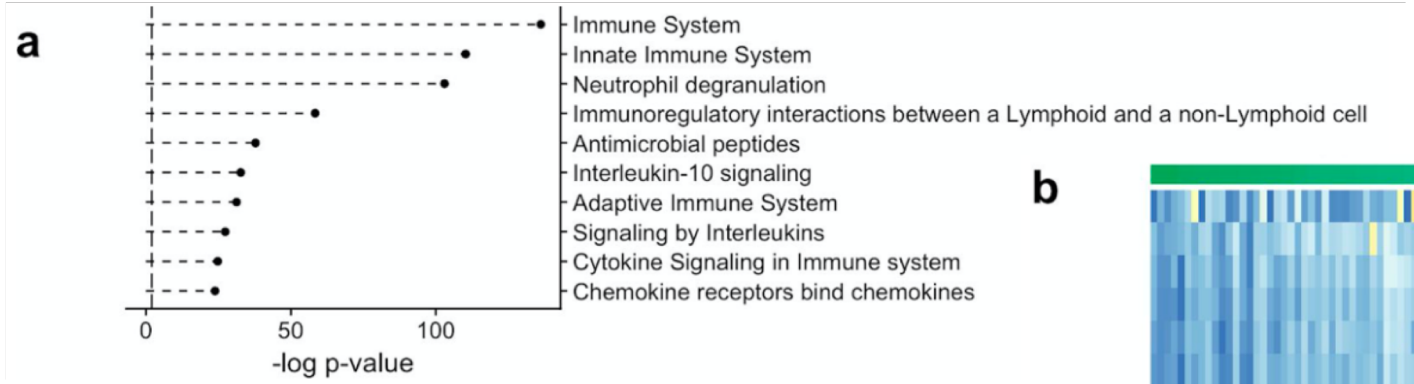
Figure EV2

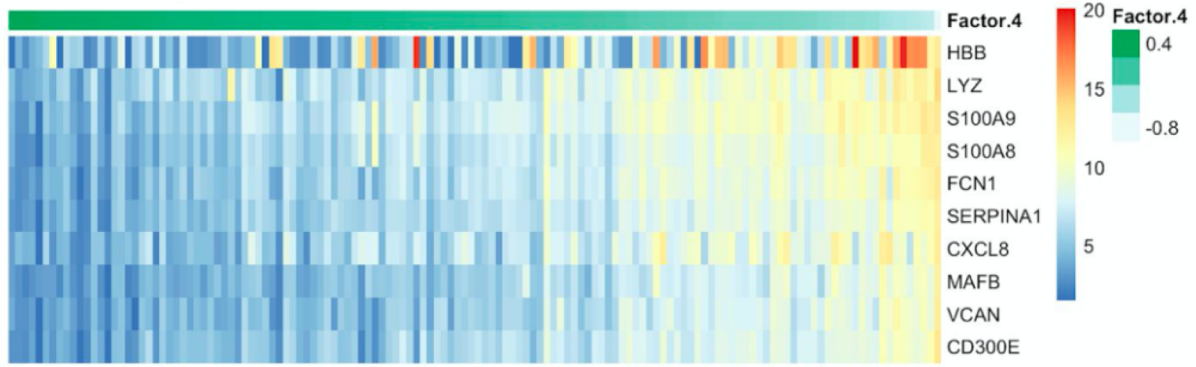

A
mRNA
TNF expression
8
11
14
Factor 5
2
0
−2
−4
Drug response
B
HSF1-dependent transactivation
Attenuation phase
Cellular response to heat stress
Cellular responses to stress
HSF1 mediated heat shock response
HATs acetylate histones
HDACs deacetylate histones
2.5
5
7.5
10
-log10 (p-value)
C
Normalized gene expression
HSPA1A
HSPA6
HSPA1B
TNF
CAMK2A
DNAJB1
15
10
5
Factor 5
Samples
D
SD07
MIS−43
SD51
NU7441
doxorubicine
enzastaurin
fludarabine
nutlin−3
BIX02188
Drug category
MAPK
DNA damage
ROS
PKC
0.00
0.50
1.00
Loading on Factor 5
E
Cellular viability after drug treatment
SD07
MIS−43
SD51
1.0
0.5
0.0
Factor 5

Real Data Analysis – CLL

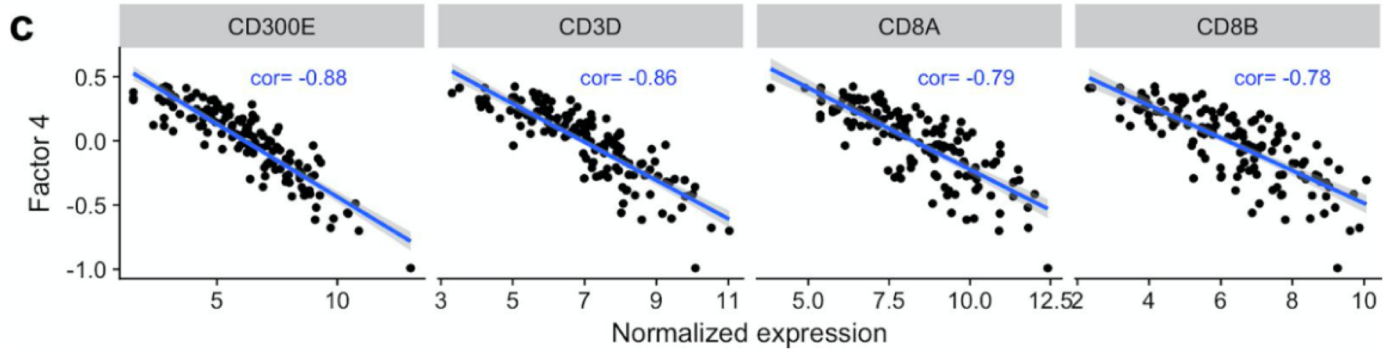
- Downstream Analysis – Annotation of factors
  - More on the factor 4.

S14





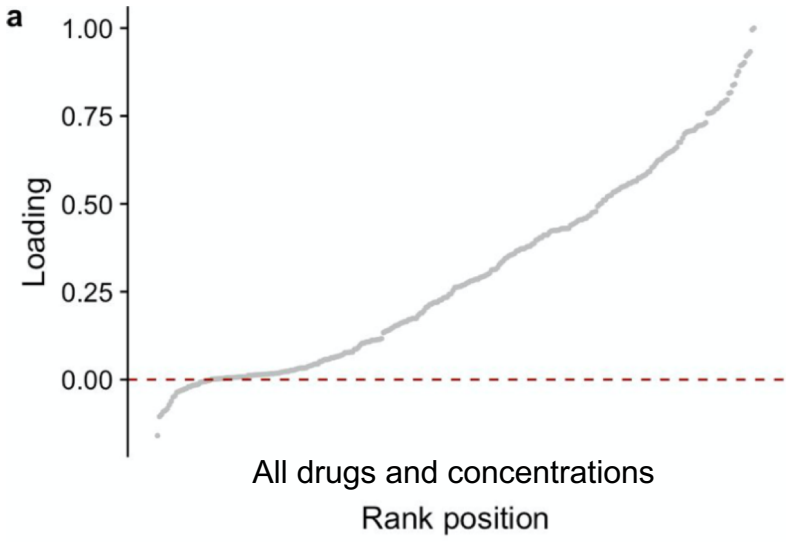
Important surface markers of T-cells

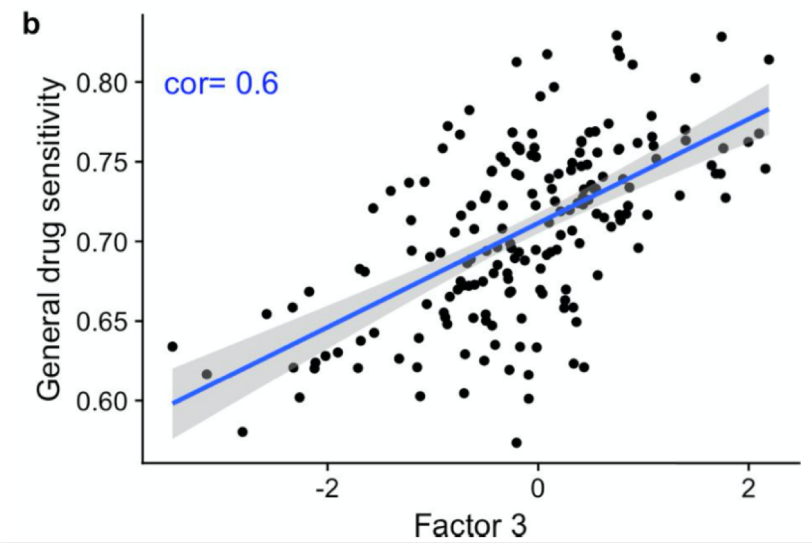


Real Data Analysis – CLL

- Downstream Analysis – Annotation of factors
  - More on the factor 3.

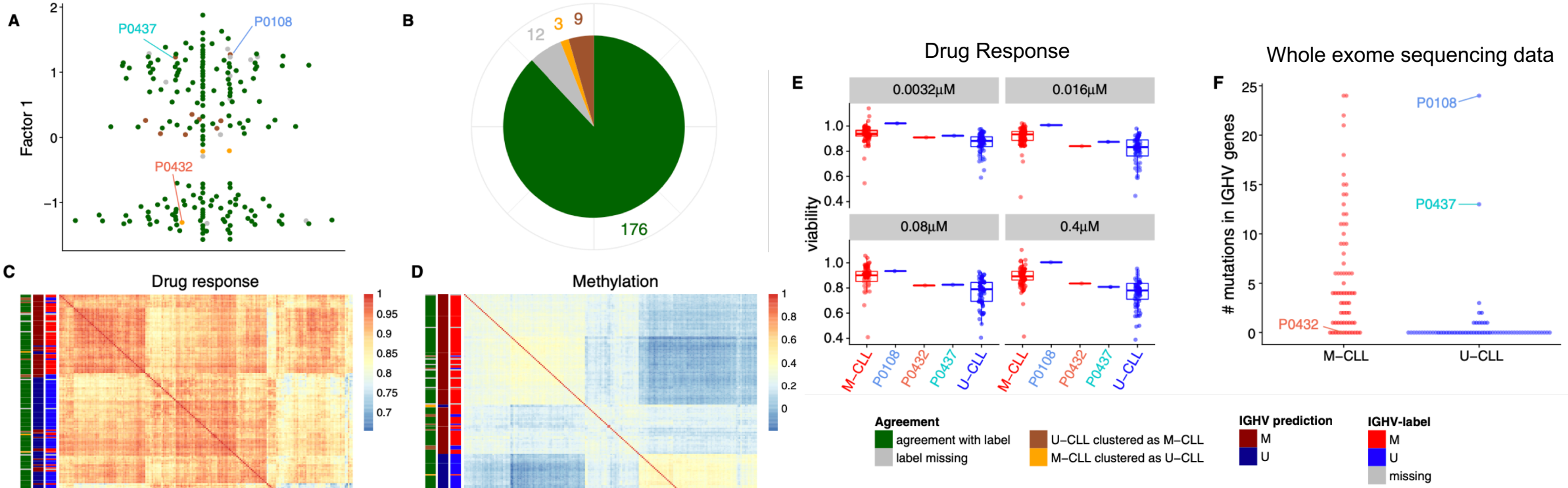
S15





Real Data Analysis – CLL

- Downstream Analysis – Outlier identification
  - Infer IGHV status by using factor 1. Among 200 samples (12 missing),
    - 9 samples: Intermediate molecular signatures => borderline cases.
    - 3 samples: molecular signatures discordant. => (EV3C, EV3D, EV3E, EV3F)



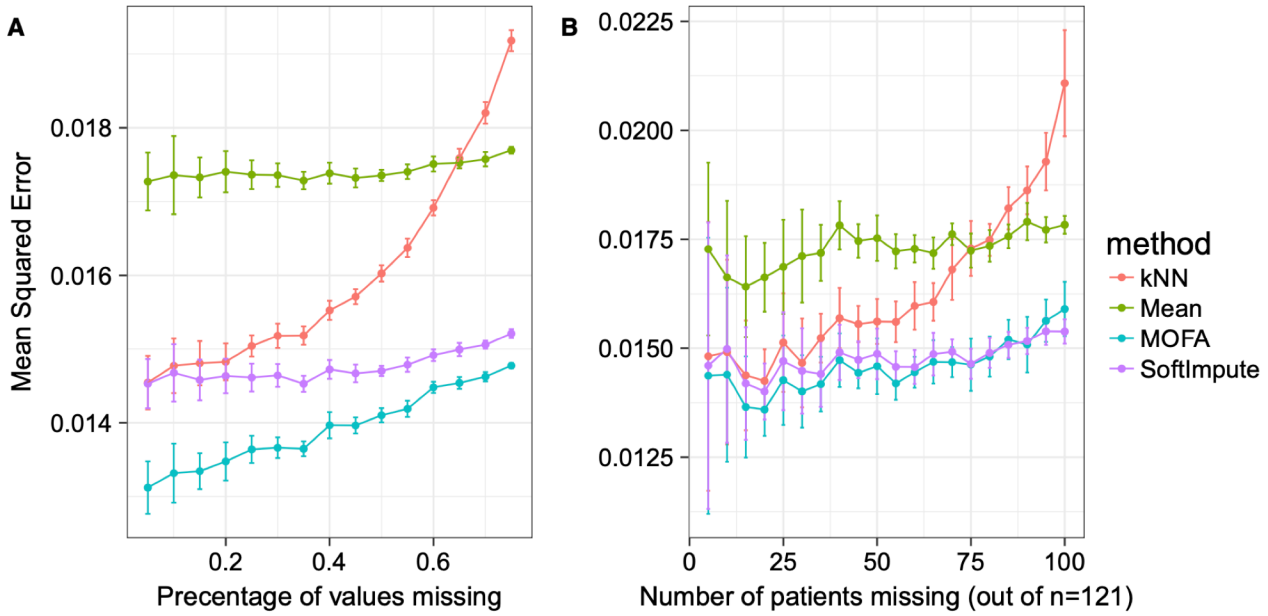
EV3

Real Data Analysis – CLL

- Downstream Analysis – Missing data imputation
  - MOFA yield a more accurate predictions than other established imputation strategies. (EV4)
  - The imputation accuracy of MOFA is more robust than GSA (S17)

EV4

drug response

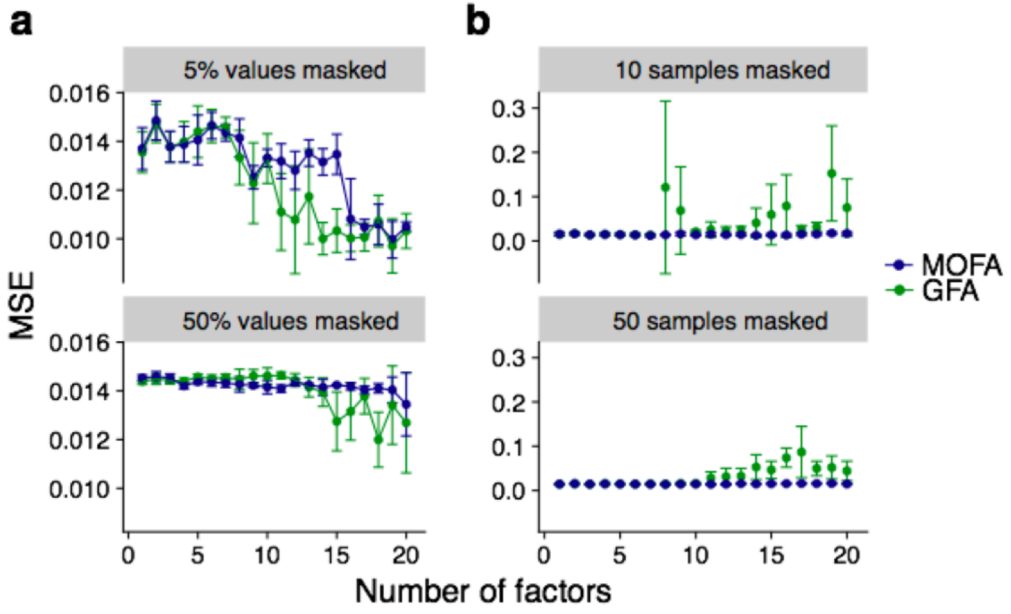


values missing at random

entire assay missing for samples at random.

S17

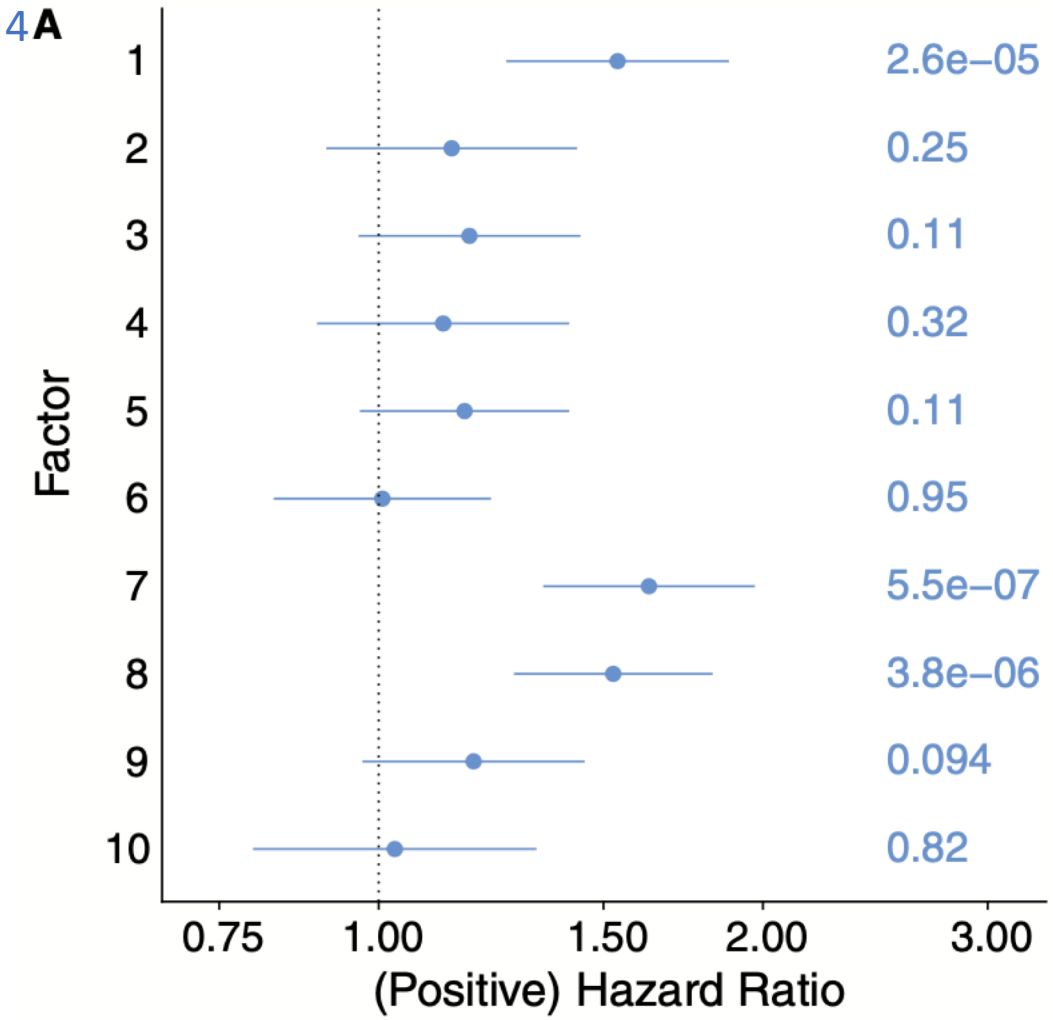
drug response



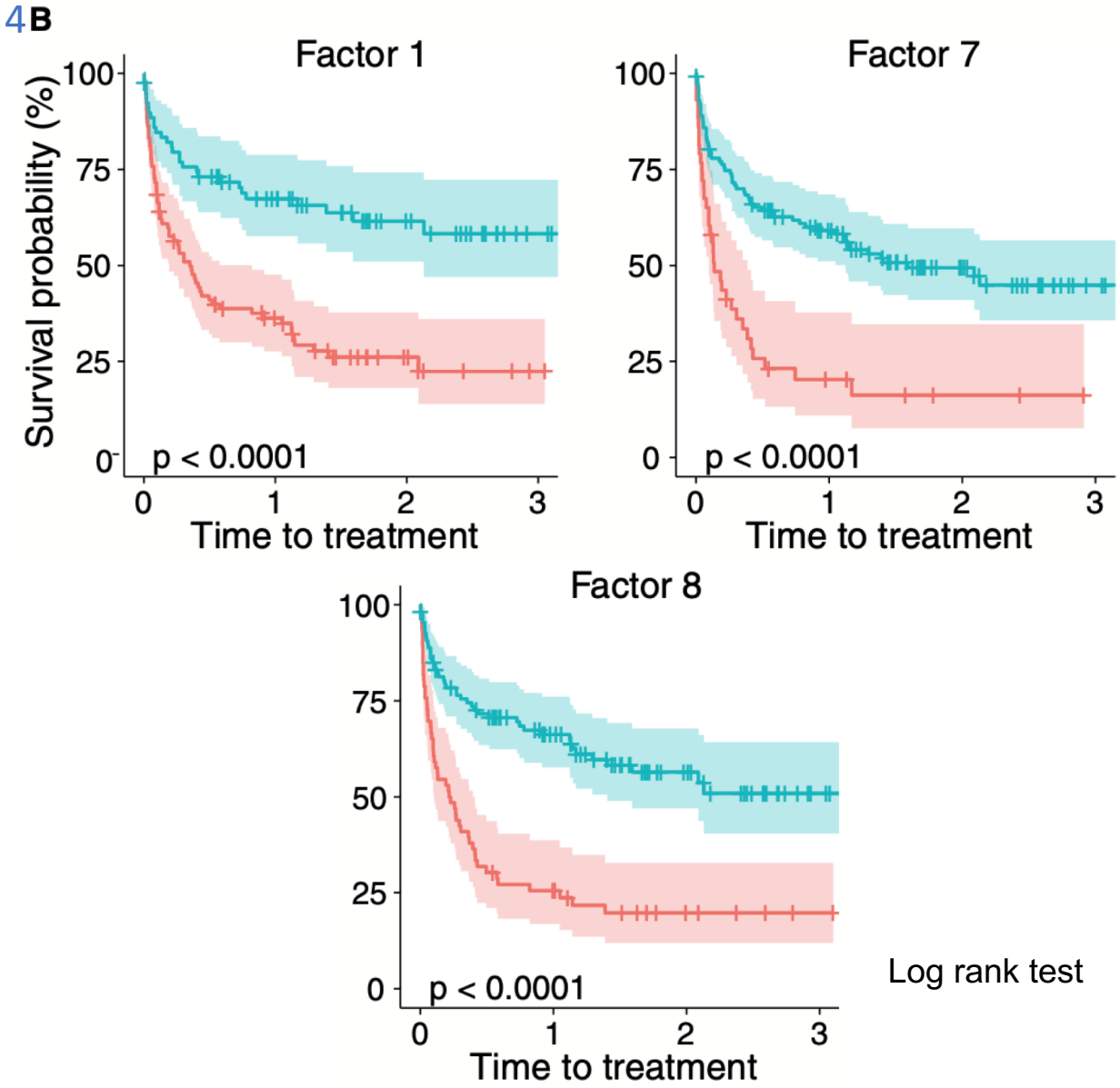
Real Data Analysis – CLL

- Prediction of clinical outcomes (time to next treatment)
  - 3 of 10 factors (1, 7, 8) are significantly associated. (Cox regression; 4A, 4B)
    - Factor 7, 8 are related to chemo-immunotherapy treatment (prior to sample collection)
      - Factor 7: captures del17p and TP53 mutations as well as differences in methylation patterns of oncogenes. (S18)
      - Factor 8: associated with WNT signaling. (S19)
  - Combining the 10 MOFA factors in a multivariate Cox regression model to predict the time.
    - Higher prediction accuracy than other models, e.g.
      a. Models using conventional PCA components. (4C)
      b. Models using individual molecular features. (S20)
      c. Models using MOFA factors derived from only a subset of the available data modalities. (S8 B, D)
    - The predictive value is similar to clinical covariates that are used to guide treatment decisions (S21)

Real Data Analysis – CLL



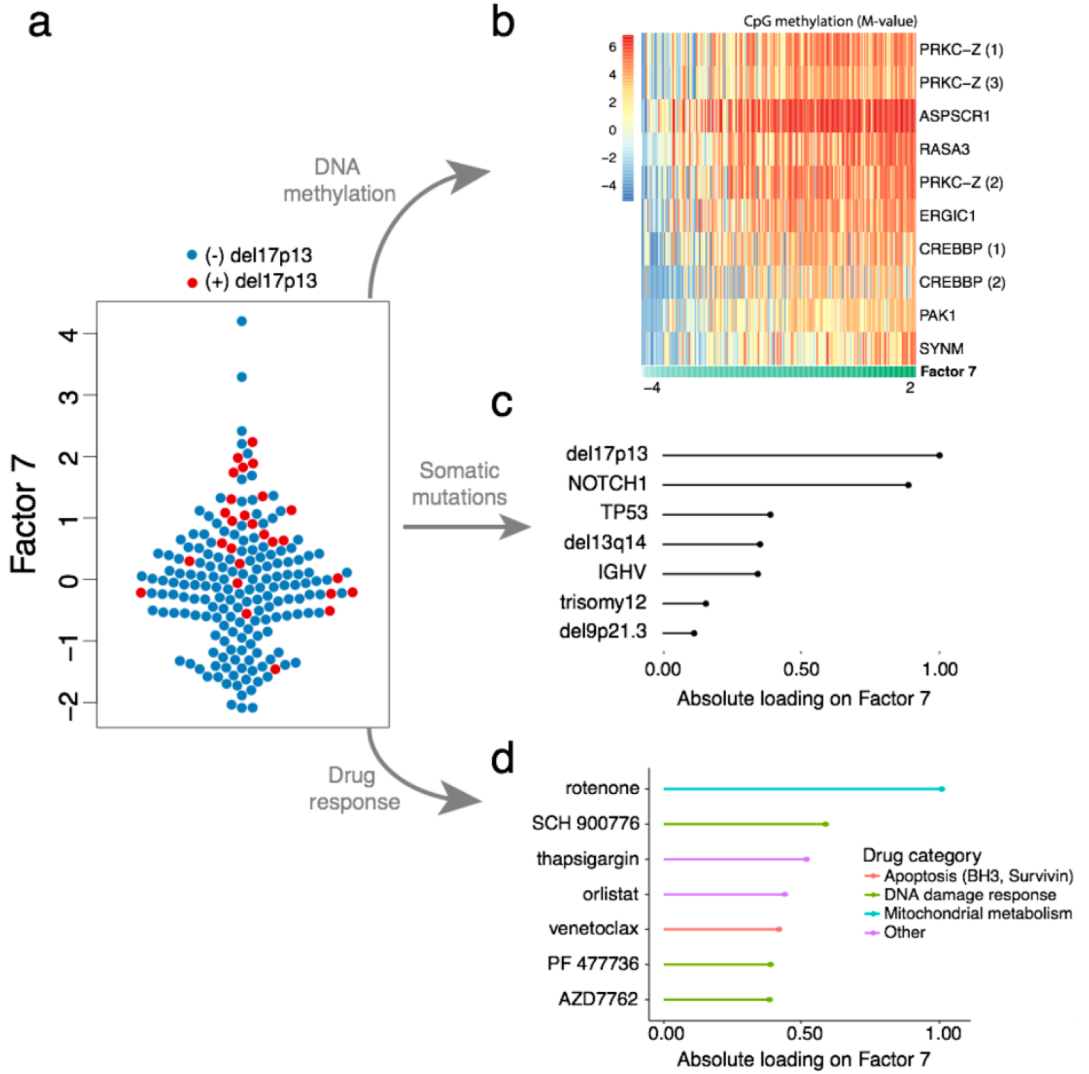
Univariate Cox models



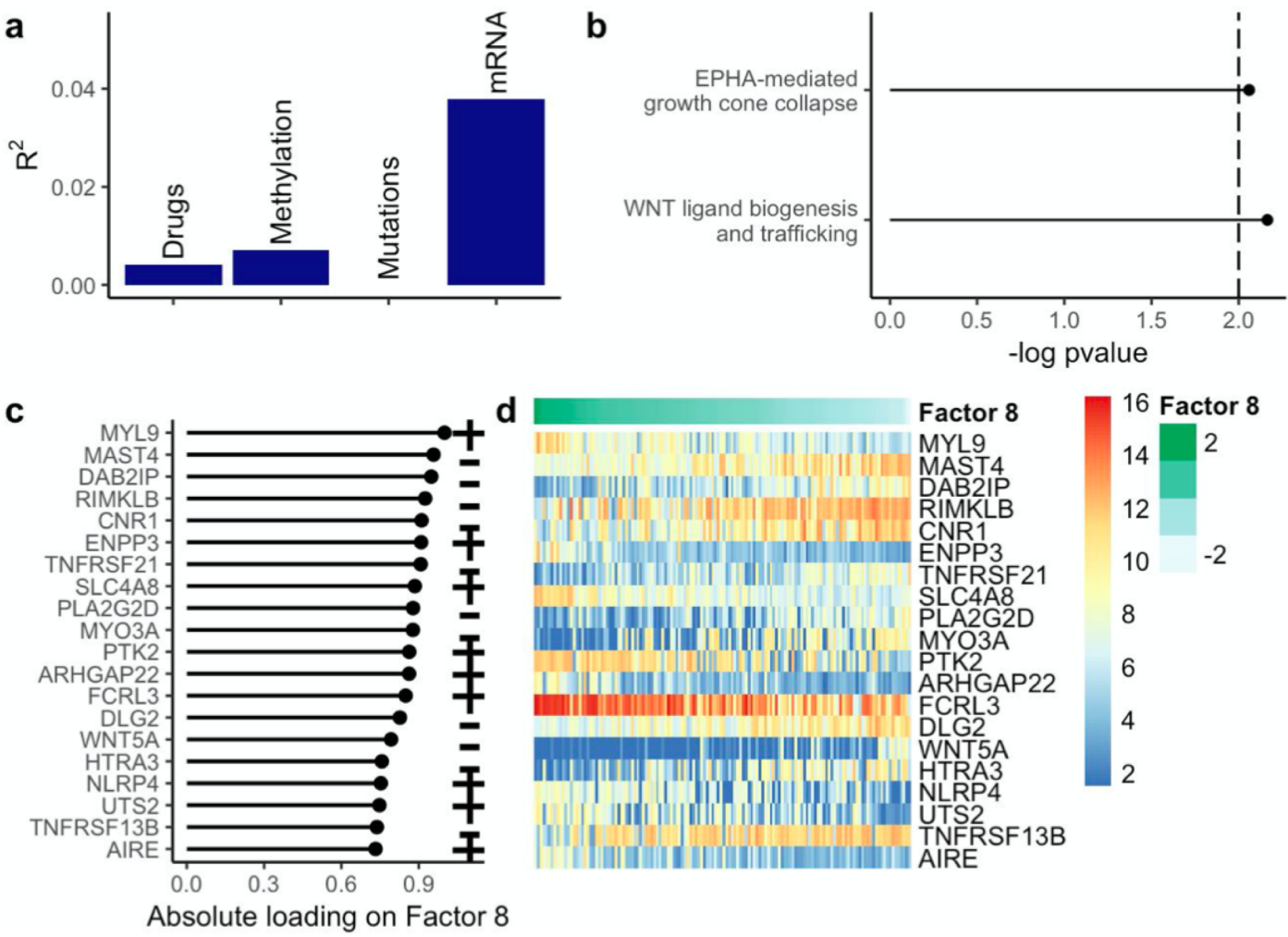
Log rank test

Real Data Analysis – CLL

S18 Factor 7 annotation

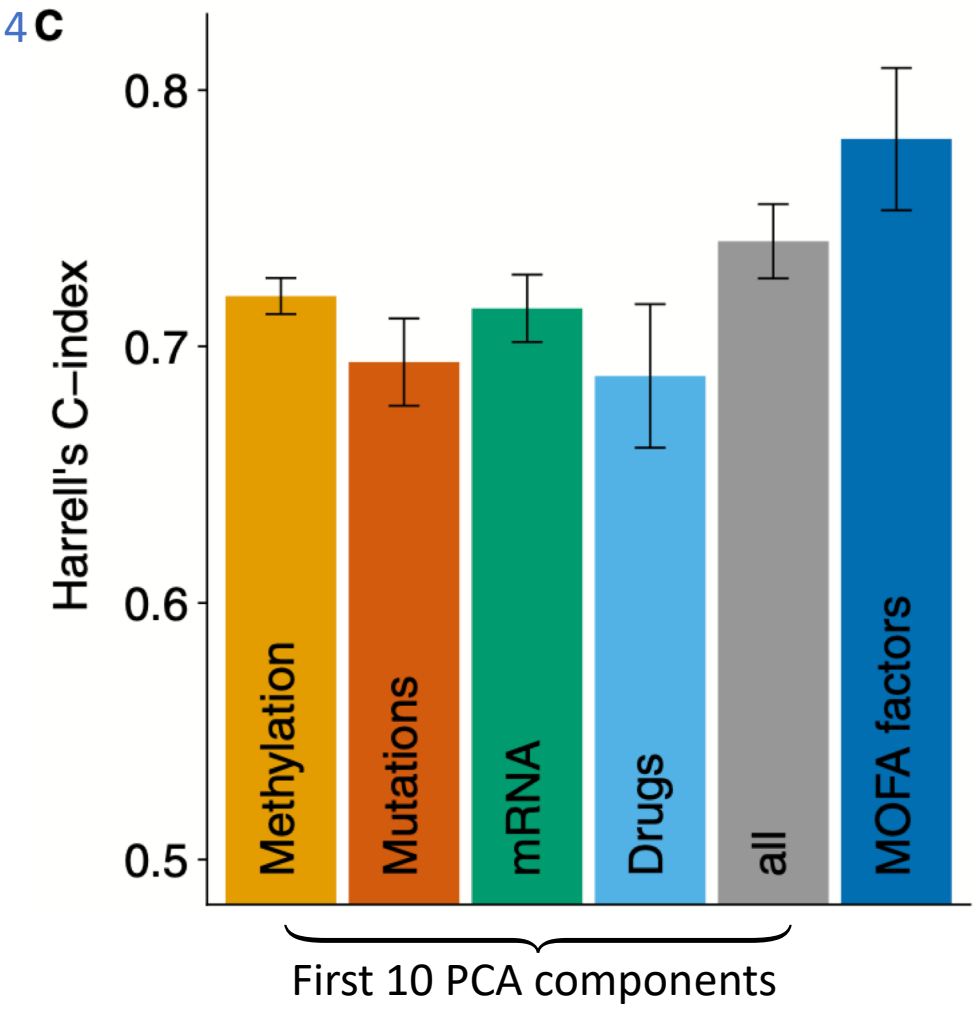


S19 Factor 8 annotation

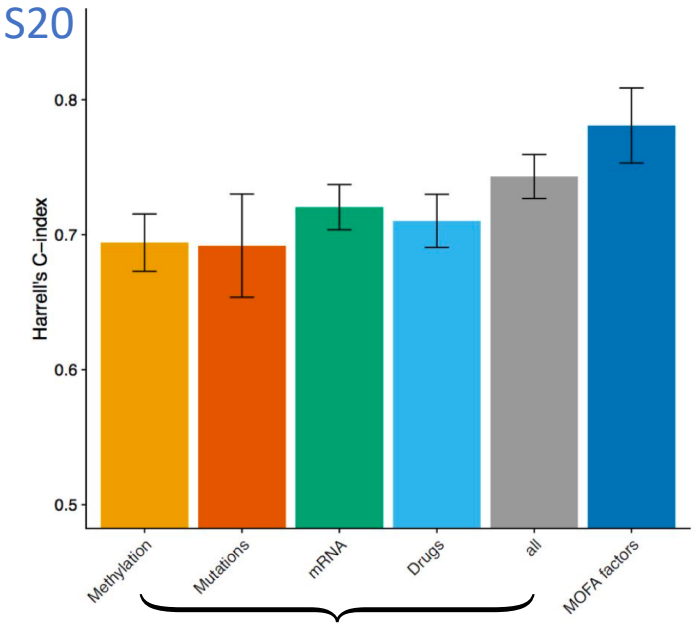


Real Data Analysis – CLL

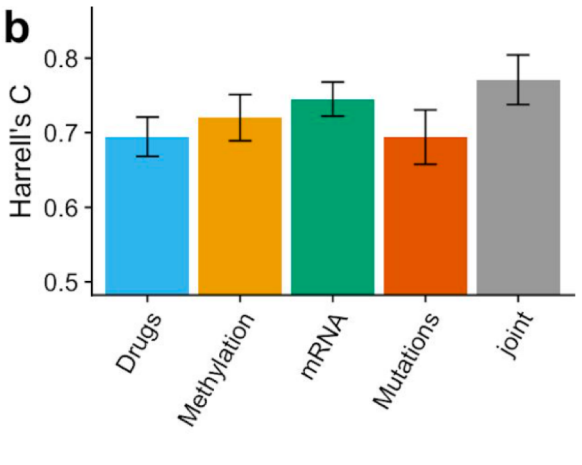
4

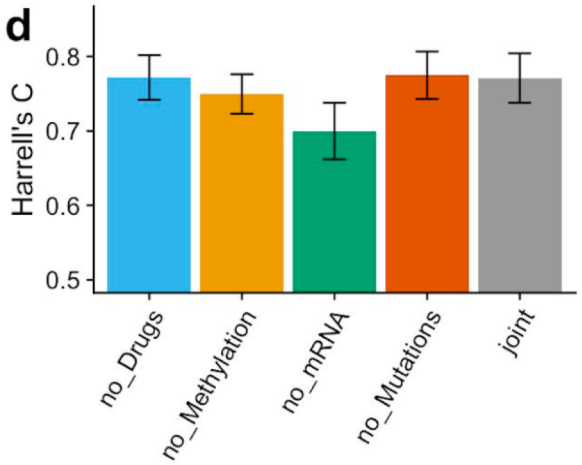


S20



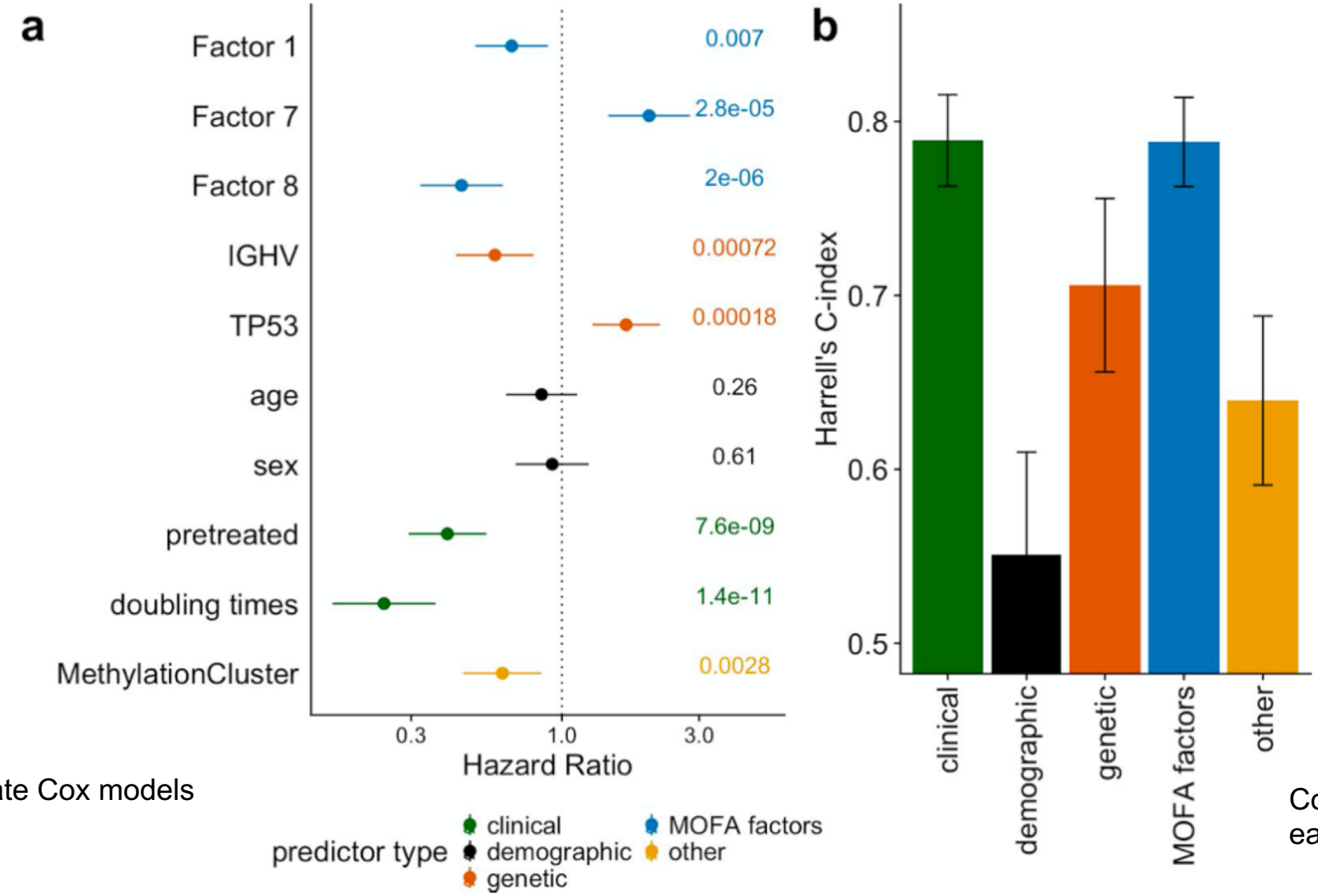
S8





Real Data Analysis – CLL

S21



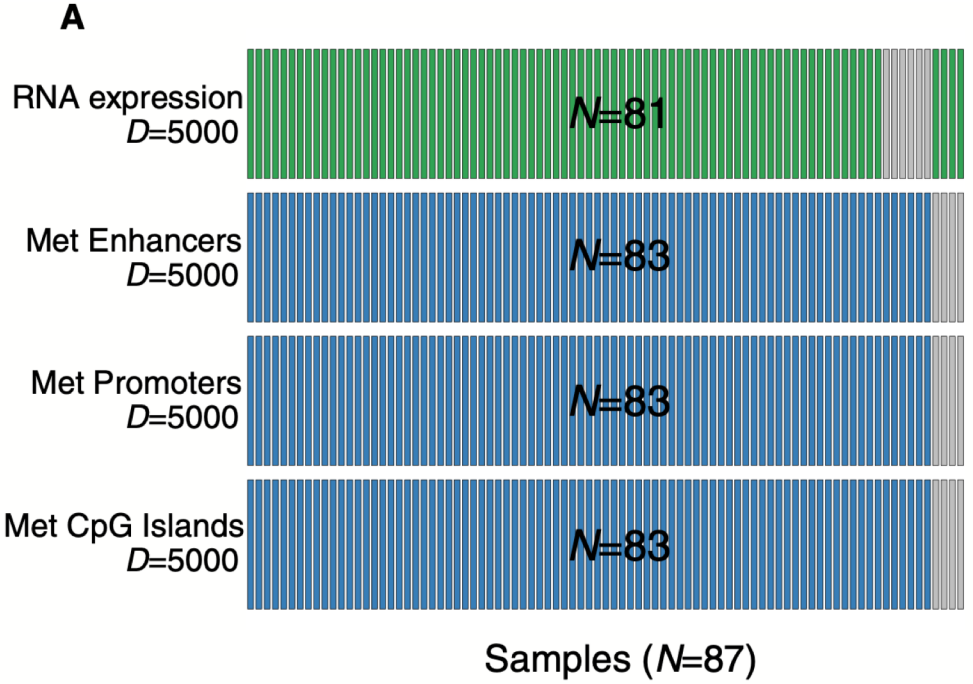
Univariate Cox models

Covariates components for each color are shown in a.

Comparison of MOFA factors with clinical covariates

Real Data Analysis – Single Cell Data

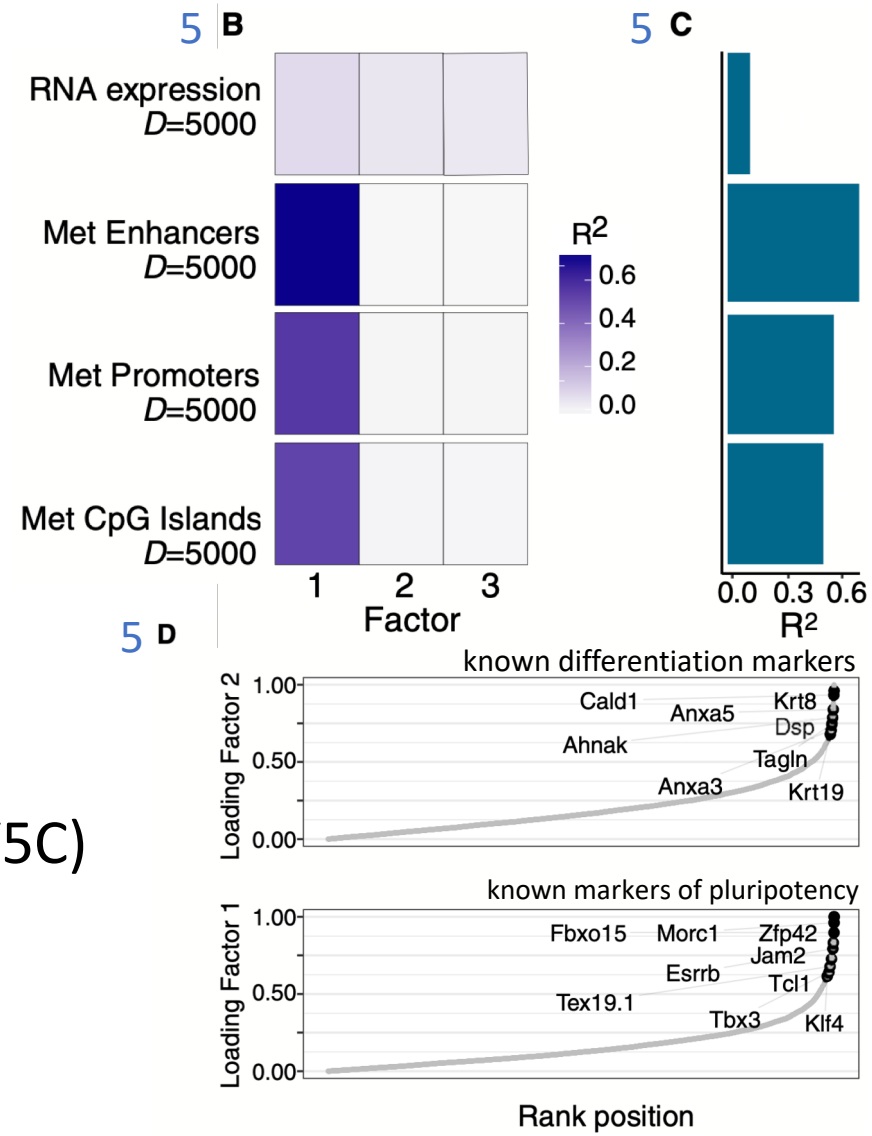
- 87 mouse embryonic stem cells (mESCs)
  - 16 “2i” media, a naïve pluripotency state.
  - 71 serum-grown cells, a primed pluripotency state.
- Data types:
  - Single-cell methylation (CpG methylation at three different genomic contexts)
    - Promoters
    - CpG islands
    - Enhancers
  - Transcriptome sequencing



Real Data Analysis – Single Cell Data

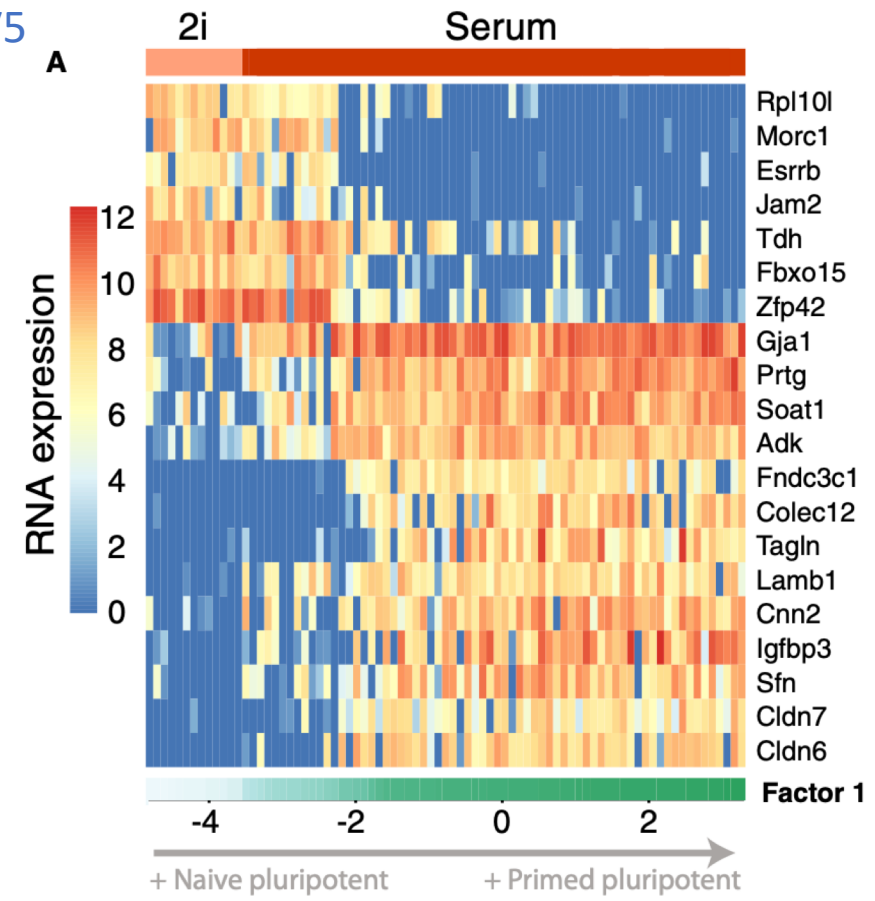
- Variance decomposition by factors (5B, C)
  - All data modality: Factor 1
  - RNA Data: Factor 2 and Factor 3

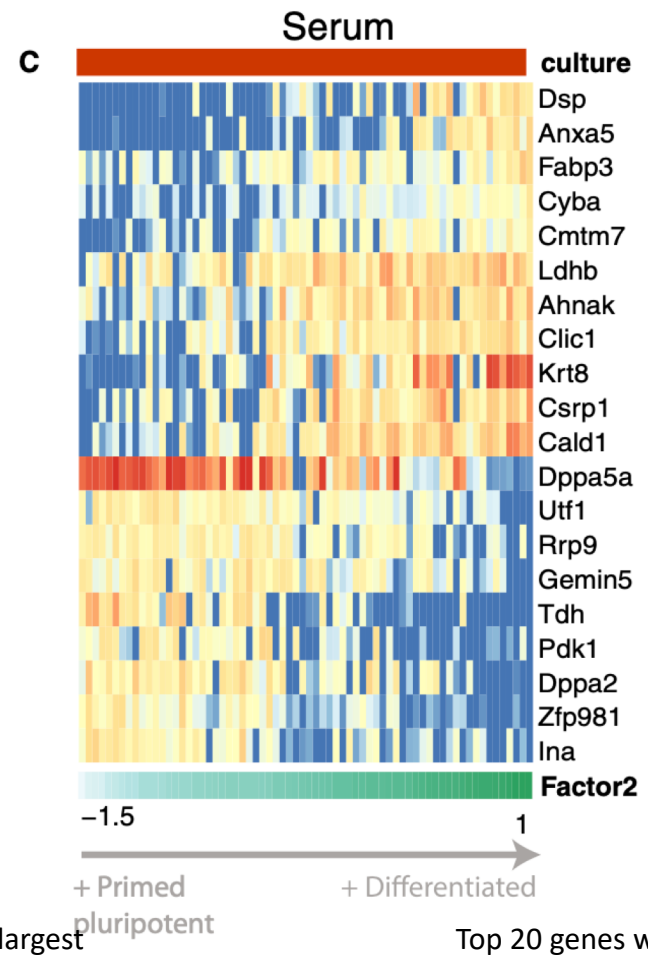
- Downstream Analysis – Annotation of factors
  - Factor 1 ⇔ The cell transition from naïve to primed pluripotent states. (5D, EV5A)
  - Factor 2 ⇔ The cell transition from primed pluripotent states to differentiated state. (5D, EV5C)
  - Factor 3 ⇔ The cell detection rate, associated with cell quality and mRNA content. (S22)



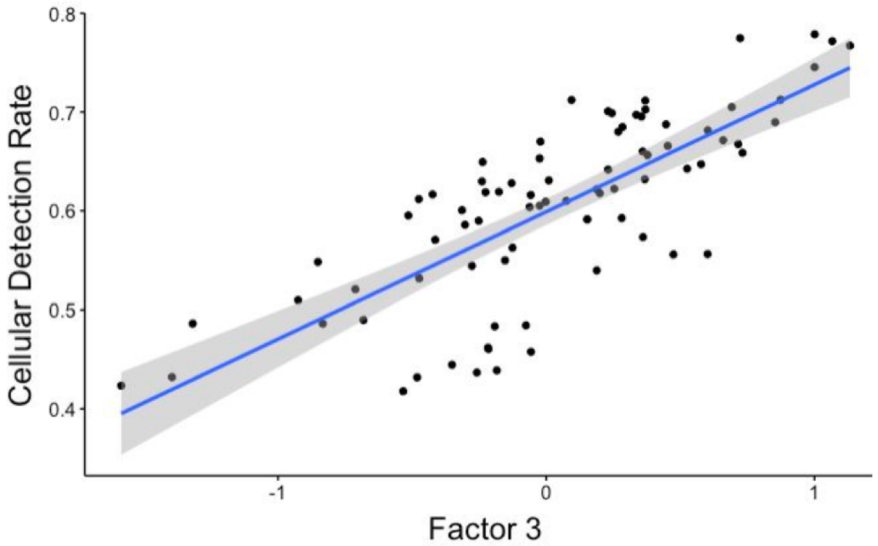
Real Data Analysis – Single Cell Data

EV5





S22



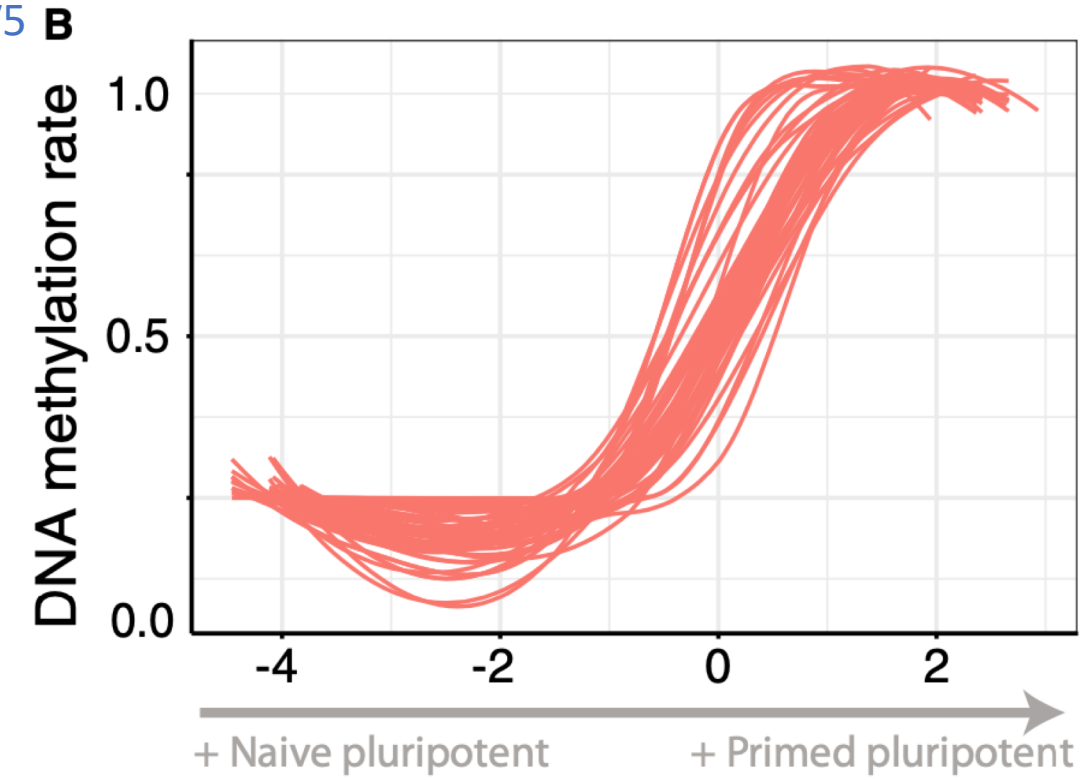
Top 20 genes with largest weight on Factor 1.

Top 20 genes with largest weight on Factor 2.

Real Data Analysis – Single Cell Data

- More on the Factor 1.
  - MOFA connect these transcriptomic changes to coordinated changes in the genome-wide DNA methylation rate

EV5

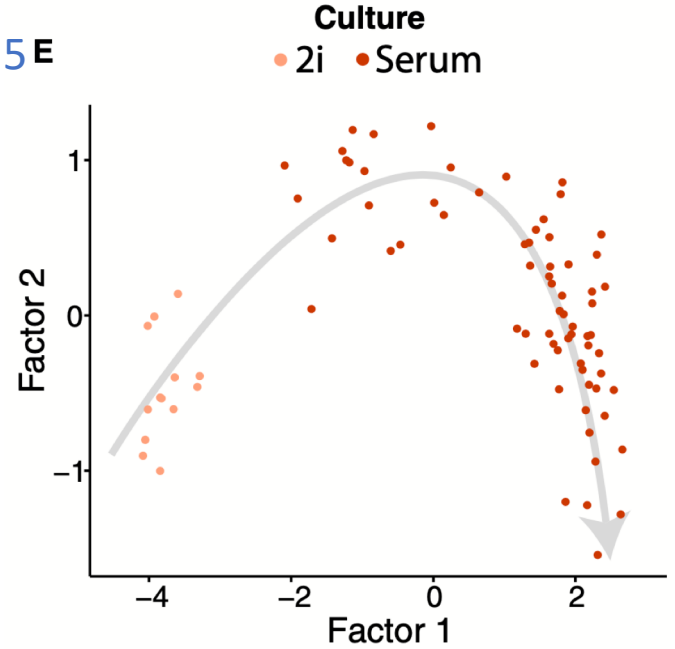


Real Data Analysis – Single Cell Data

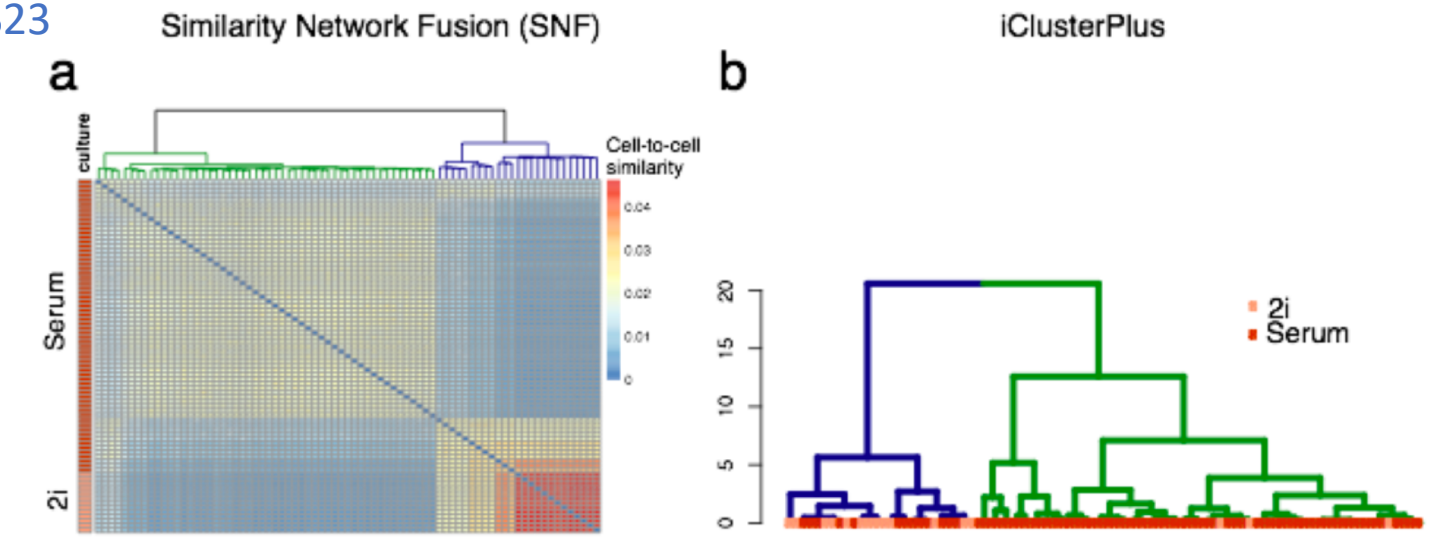
- Downstream Analysis – Annotation of factors
  - Factor 1 and factor 2 captured the entire differentiation trajectory. (5E)

    → Learning continuous latent factors rather than discrete sample assignments.
  - Other Multi-omics clustering algorithms (e.g. SNF, icluster) can only distinguishing cellular subpopulation (discrete). (S23)

5 E



S23



Discussion and Summary

- MOFA is an unsupervised method for decomposing the heterogeneity in multi-omics data.
- MOFA is applied in the patient-derived tumor samples and a single-cell study of mESCs in this study.
- In the CLL study, MOFA
  - Capture variations of multiple features and data modalities.
  - Increase the sensitivity for identifying molecular signatures compare to using individual feature or data modality.
  - Use information from multiple omics layers to accurately impute missing values.
  - Guide the detection of outliers.
- In the mESCs study, MOFA learn continuous factors instead of discrete factors.

  → Find a differentiation trajectory to predict pluripotent states.

Discussion and Summary

- Different from other models for integrating different data types, MOFA
  - Provide interpretable reconstruction of the underlying factors.
  - Accommodate different data type and different patterns of missing data.
- Challenges for MOFA:
  - Miss strong non-linear relationships between features and across data types.

    → Non-linear extensions of MOFA. (Model complexity, computational efficiency and interpretability decrease)

    → New likelihood and noise models needed.

    → Bayesian treatment could be considered.

Thank You