V10 Multi-omics data integration

Program for today:

- Data integration methods
    - multi-staged approaches
    - meta-dimensional analysis
- Multiomics factor analysis (MOFA) + example
- Similarity network fusion (SNF) + example
- Outlook: rethink data analysis

Today, we will consider the issue of analyzing multiple types of omics-data and correlating this with e.g. clinical data.

We will first classify the existing approaches on how the analysis is done, step-wise or simultaneously.

Then we will discuss 2 methods in more detail, SNF and MOFA.

You will implement and apply the SNF method in assignment #5.

Benefits of multi-omics data from a biological viewpoint

Main motivation behind combining different data sources:
Identify genomic factors and their interactions
that **explain** or **predict disease risk.**

(1) Additional data dimensions may compensate for **missing** or **unreliable information** in any single data type

(2) If multiple sources of evidence point to the same gene or pathway, one can expect that the likelihood of **false positives** is reduced.

(3) It is likely that one can uncover the **complete biological model** only by considering different levels of genetic, genomic and proteomic regulation.

Ritchie et al.

Nature Rev Genet **16**, 85 (2015)

Link to this review paper: https://www.nature.com/articles/nrg3868

Regulation and deregulation take place on different layers.

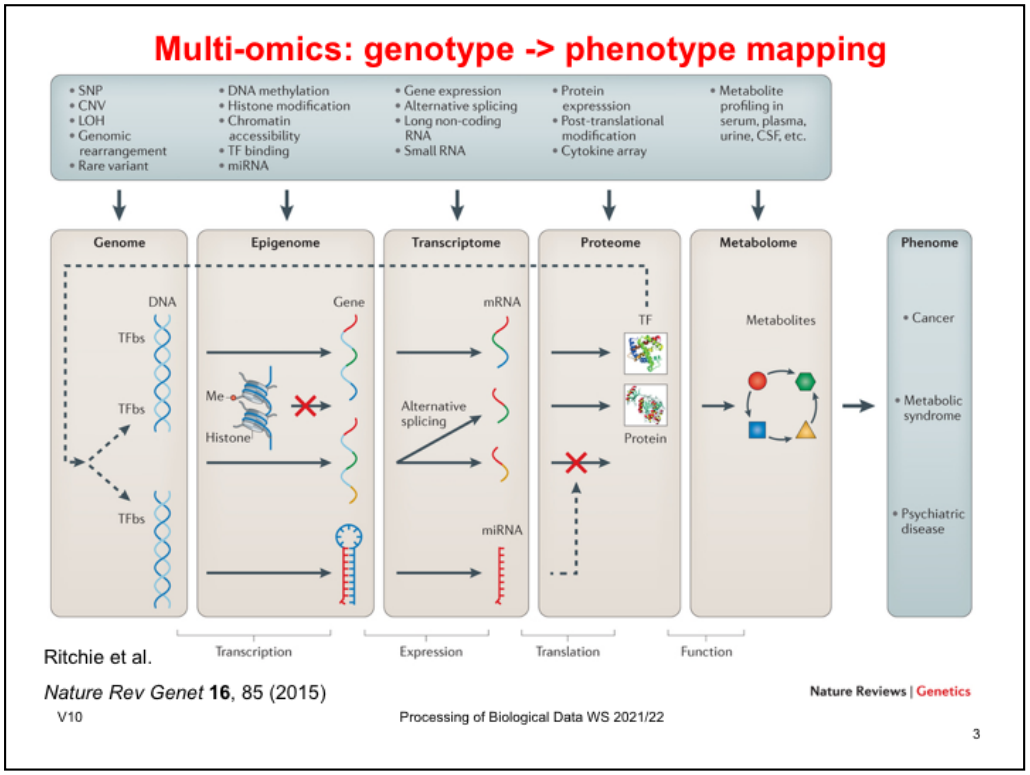
By including multiple data types, we hope to capture more aspects e.g. why and how deregulation may lead to disease processes.

The first point – circumvent missing data in one data dimension – may sound a bit childish. As if the available methods work so poorly that we need „tricks“ to overcome this.

This may be partly true. But there are many reasons why data points are missing, not only imperfect omics methods. Some points simply cannot be measured.

The second point is always true. If one has independent evidence from multiple directions, the confidence about a finding increases.

The third point is also true. Eventually we like to understand basically every aspect about cell biology. However, we are still far away from this stage.



Shown here are different levels of molecular omics data: genome, epigenome, transcriptome, proteome and metabolome

Within each level and also between different levels, there exist heterogeneous data.

Arrows indicate the flow of genetic information from the genome level to the metabolome level and, ultimately, to the phenome level. The red crosses indicate inactivation of transcription or translation.

Abbreviations:

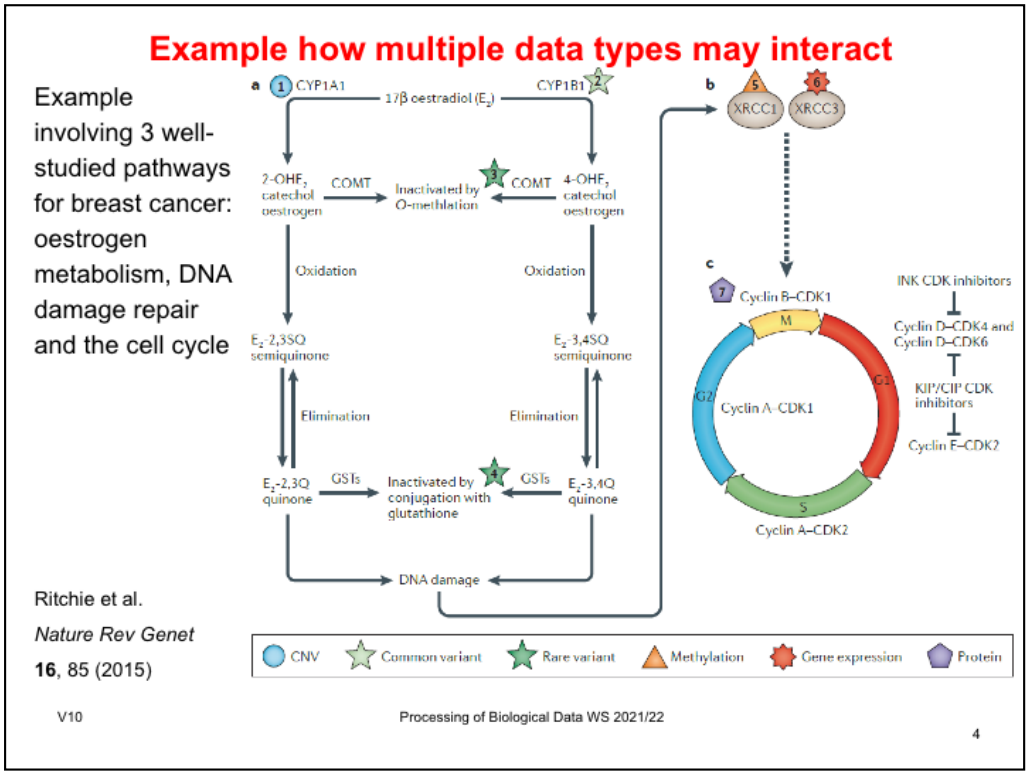
LOH stands for „loss of heterozygosity“ = one parental copy of a gene is lost due to a chromosomal (mutational) event.

CNV stands for „copy number variation“ = a type of duplication or deletion event that affects a considerable number of base pairs (kb up to Mb).

CSF, cerebrospinal fluid;

Me, methylation;

TFBS, transcription factor-binding site.



In this purely hypothetical example taken from Ritchie et al., we illustrate that an analysis that assesses variation of only a single omic data type can miss complex models that require variation across multiple levels of biological regulation.

It is now established that oestrogen can cause DNA damage if it is not properly metabolized. Two genes, cytochrome P450 1A1 (*CYP1A1)* and *CYP1B1*, participate in the first step of oestrogen breakdown. The metabolite created by CYP1B1 (4-OHE_2 catechol oestrogen) creates a more carcinogenic form of oestrogen by-product than that metabolized by CYP1A1.

In this hypothetical scenario, a copy number variation (CNV) in *CYP1A1* (label 1 in the left figure) reduces activity, and single-nucleotide polymorphisms (SNPs) in *CYP1B1* (label 2) increase activity, resulting in higher levels of carcinogenic by-products. Additionally, multiple rare variants in the gene coding for the enzymes caffeic acid 3-*O*-methyltransferase (*COMT*; label 3), glutathione *S*-transferase µ1 (*GSTM1*) and glutathione *S*-transferase θ1 (*GSTT1*; label 4) reduce the metabolism (i.e. degradation) of carcinogenic by-products, resulting in a higher level of DNA damage. Even so, these variations may not increase the risk of cancer if the DNA damage repair pathway can offset the increase in carcinogenic metabolites. However, hypermethylation of X-ray repair cross-complementing 1 (*XRCC1*; label 5) and variation in the gene expression of *XRCC3* (label 6) result in reduced

transcription levels, and this repair pathway may no longer be able to adequately keep DNA repair at necessary levels (see right figure). Finally, dysregulated protein expression of genes in the cell cycle pathway — for example, in cyclin-dependent kinase 1 (CDK1; label 7) — may result in a rate of cell replication that is higher than average and therefore DNA damage (right figure). The end result can lead to an abundance of damaged cells (that is, breast cancer cells). In this hypothetical model, all of the variation mentioned above is required to pass the threshold into cancer development. Therefore, only an analytical approach that integrates data from the genome, transcriptome and proteome would identify the full model.

Methods for data integration

Ritchie et al. classify multi-omics data integration methods into these 2 classes:

(1) Multi-staged approaches consider different data types in a stepwise / linear / hierarchical manner.

(2) **Meta-dimensional approaches** consider different data types simultaneously.

Ritchie et al.

Nature Rev Genet **16**, 85 (2015)

Multi-staged analysis is conceptually much simpler than meta-dimensional analysis.

Multi-staged analysis: eQTL analysis

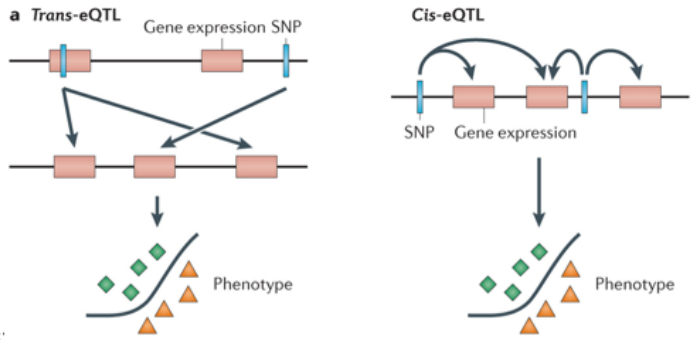
Steps: (1) associate SNPs with phenotype; filter by significance threshold

(2) Test the SNPs that are associated with phenotype with other omic data. E.g. check for the association with gene expression data -> eQTL (expression quantitative trait loci). Also: methylation QTLs, metabolite QTLs, protein QTLs …

(3) Test omic data used in step 2 for correlation with phenotype of interest.

Trans-eQTL: effect on remote gene

Cis-eQTL: effect on nearby gene



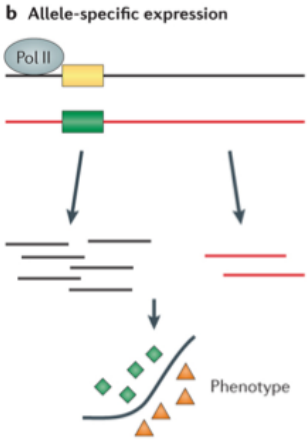
Ritchie et al.

Nature Rev Genet **16**, 85 (2015)

The idea of this stepwise approach is to not only identify a biomarker-SNP, but also understand how the SNP leads to the phenotypic change.

For example, analysis of expression quantitative trait loci (eQTLs) tries to identify elements of genetic variation associated with measures of quantitative gene expression.

Multi-staged analysis: allele specific expression (ASE)



In diploid organisms, some genes show differential expression of the two alleles.

Similar to the analysis of eQTL SNPs, ASE analysis tries to correlate single alleles with phenotypes.

ASE analysis tests whether the maternal or paternal allele is preferentially expressed.

Then, one associates this allele with *cis*-element variations and epigenetic modifications.

Ritchie et al.

Nature Rev Genet **16**, 85 (2015)

This is another example of a multi-staged analysis.

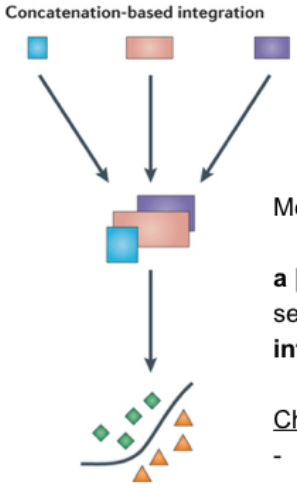
In the top figure, the black and orange lines symbolize the two copies of the chromosome inherited from father and mother.

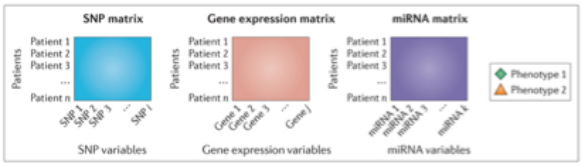
In the example, the RNA polymerase would preferentially bind to the promoter of the paternal copy of a gene (yellow) and hence produce more mRNA transcripts from it (short black lines, middle) than from the orange allele.

In the first step of this multi-stage analysis, one checks for allele specific expression.

In the second step, one tries to link the obtained results (which genes show allele specific expression?) to variations in promoter/enhancer elements or to epigenetic variations.

Meta-dimensional analysis: concatenation-based integration





Meta-dimensional analysis can be divided into 3 categories.

a | Concatenation-based integration involves **combining** data sets from different data types at the raw or processed data level **into one matrix** before modelling and analysis.

Challenges:

- what is the best approach to combine multiple matrices that include data from different scales in a meaningful way?
- the high-dimensionality of the data is inflated further (number of samples < number of measurements per sample)

Ritchie et al.

Nature Rev Genet **16**, 85 (2015)

V10 Processing of Biological Data WS 2021/22 8

Now, we turn to the case when multiple data types are analyzed at once. This is called **meta-dimensional analysis**.

In this area, we can distinguish 3 types of approaches: concatenation-based integration, transformation-based integration, and model-based integration.

We will start with data concatenation where multiple data types are available in individual data matrices.

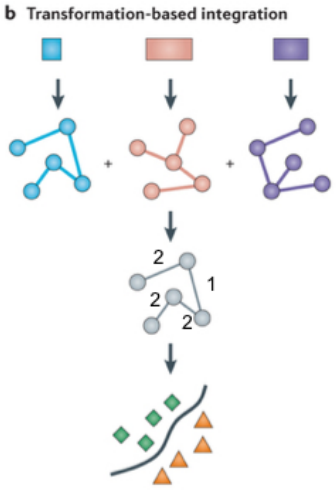
This is illustrated in the top line. The blue square contains SNP data – what nucleotide does each patient have at each SNP position?

The red square contains transcriptomics data – what are the expression levels of all genes for each patient?

The purple square contains miRNA data – e.g. what is the expression level of all miRNAs for each patient?

If one concatentates all this data into one matrix, this matrix may become pretty large. Also, the solution space may become severly underdetermined because there are typically many more variables than samples (patients).

Meta-dimensional analysis: transformation-based integration



b | Transformation-based integration involves performing mapping or data transformation of the underlying data sets before analysis.

In this example, the 3 initial graphs are all spanning trees.
Then, one of them is selected as representative.
It has most "support" from the 3 initial trees.

The modelling approach is then applied at the level of transformed matrices.

Ritchie et al.

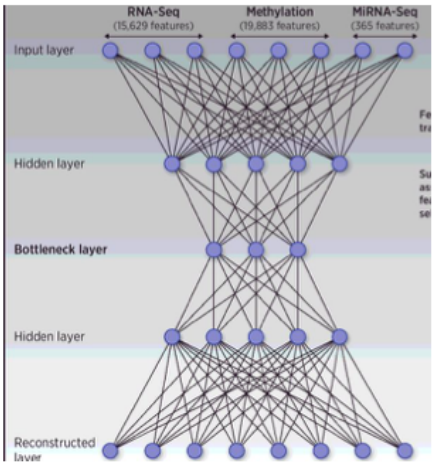
Nature Rev Genet **16**, 85 (2015)

The second type of approaches involves independent mapping or data transformation of the separate data types prior to integrating them.

Meta-dimensional analysis: transformation-based integration

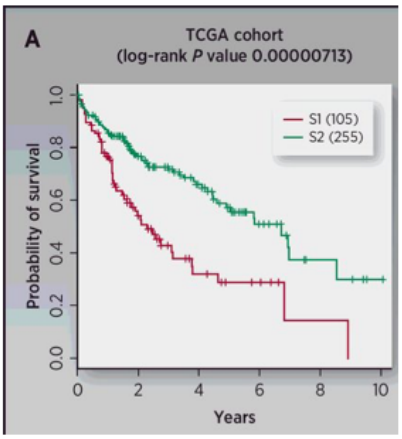
Typical examples of transformation-based integration are deep learning studies.

Here, the authors combined RNA-seq data, miRNA-seq data, and DNA methylation data for HCC patients from TCGA.



Chaudhary et al. Clin Cancer Res. 24: 1248–1259 (2018)

All data were preprocessed appropriately.



Link to the Chaudhary paper: https://pubmed.ncbi.nlm.nih.gov/28982688/

From the TCGA HCC project, the authors obtained 360 tumor samples with coupled RNA-seq (15,629 genes after preprocessing), miRNA-seq (365 miRNAs) and DNA methylation data (19,883 genes).

From the DNA methylation data, they considered CpG islands within 1500 bp ahead of transcription start sites (TSS) of genes and averaged their methylation values.

Missing values were processed in the following way: First, the biological features (e.g. genes/miRNAs) were removed if having zero value in more than 20% of patients. The samples were removed if missing across more than 20% features. The other missing values were imputed with the *impute* function from R impute package. Lastly, input features with zero values across all samples were removed. (Comment: such features contain no information -> are not useful for the deep learning approach.)

The 3 types of omics features (contained in 3 matrices that are unit-norm scaled by sample) were then stacked into a unique matrix

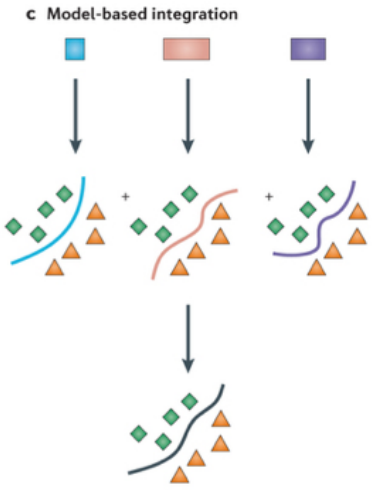
Then, an autoencoder, a deep learning framework, was trained. Its topology is shown in the top figure. The authors used the activity of the 100 nodes from the bottleneck hidden layer as new features. They then conducted univariate Cox-PH regression on each of the 100 features, and identified 37 features

significantly (log-rank p-value <0.05) associated with survival. These 37 features were subjected to K-means clustering, with cluster number K ranging from 2 to 6. Using silhouette index and the Calinski-Harabasz criterion, they found that K=2 was the optimum with the best scores for both metrics.

Survival analysis on the full TCGA HCC data showed that the survivals in the two sub-clusters are drastically different (log-rank p-value =7.13e-6, right figure).

In association with clinical characteristics, the more aggressive subtype (S1) has consistent trends of association with higher *TP53* inactivation mutation frequencies. Association of stemness markers (*KRT19, EPCAM*) with S1 subtype is also in congruence with the literature. Moreover, S1 subtype is enriched with activated Wnt signaling pathway.

Meta-dimensional analysis: model-based integration



c | Model-based integration is the process of performing analysis on each data type independently.

This is followed by integration of the resultant models to generate knowledge about the trait of interest.

Ritchie et al.

Nature Rev Genet **16**, 85 (2015)

This is the third type of meta-dimensional analysis.

Example of model-based integration: icluster

The main idea behind iCluster is that **tumor subtypes** can be modeled as unobserved (latent) variables that can be simultaneously estimated from copy number data, mRNA expression data and other available data types.

Let's assume we have one only data type (expression data) available and the input data is already correctly clustered into K clusters (or appropriately labeled e.g. by tumor subtype)..

Then, we can formulate a Gaussian latent variable model:

$$\mathbf{X} = \mathbf{W}\,\mathbf{Z} + \boldsymbol{\varepsilon}$$

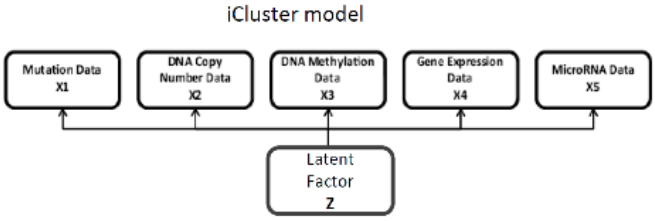
where **X** is the mean-centered expression matrix of dimension $p \times n$ (no intercept),
$\mathbf{Z} = (z_1, \ldots, z_{K-1})'$ is the cluster indicator matrix of dimension $(K-1) \times n$,
W is the coefficient matrix of dimension $p \times (K-1)$, and
$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$ is a set of independent error terms with zero mean.

Shen et al. Bioinformatics 25: 2906 (2009)

The **icluster** method is presented in this paper:
https://academic.oup.com/bioinformatics/article/25/22/2906/180866

Example of model-based integration: icluster2

The basic concept of iCluster is to jointly estimate $Z = (z_1,\ldots, z_{K-1})'$, the latent tumor subtypes, from, say, DNA copy number data (denoted by X_1, a matrix of dimension $p_1 \times n$), DNA methylation data (denoted by X_2, a matrix of dimension $p_2 \times n$), mRNA expression data (denoted by X_3, a matrix of dimension $p_3 \times n$) and so forth.



The mathematical form of the integrative model is

$$X_1 = W_1 Z + \varepsilon_1$$
$$X_2 = W_2 Z + \varepsilon_2$$
$$\ldots$$
$$X_m = W_m Z + \varepsilon_m$$

where *m* is the number of genomic data types available for the same set of samples.

V10 Processing of Biological Data WS 2021/22 13

Shen et al. Bioinformatics 25: 2906 (2009)

Actually, the matrix Z (cluster indicator e.g. for the latent tumor subtypes) is not known. This is what we want to derive.

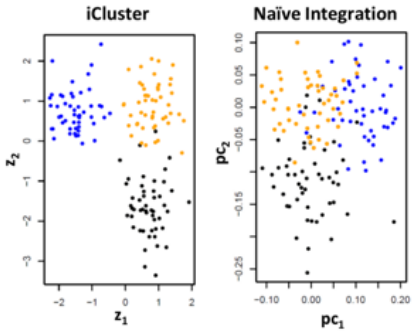
Here, we set up separate latent models for each data type. Each of them contains the same Z matrix.

W and Z are then obtained by an expectation maximization (EM) approach.

Example of model-based integration: icluster2

Application of icluster2 to data from glioblastoma patients from TCGA:
- copy number
- gene expression data
- DNA methylation data

iCluster clearly identified 3 subtypes, PCA was less successful in separating the clusters.



.

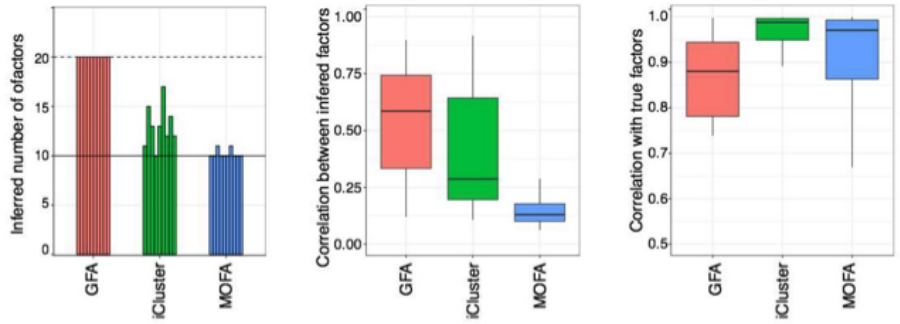
Shen et al. PLoS One 7: e35236 (2012)

Shen et al. Bioinformatics 25: 2906 (2009)

This application is presented in: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0035236

This example shows that iCluster yielded better separated clusters than standard PCA (termed „naive integration“ here).

Comparison of MOFA, GFA and iCluster on simulated data



(a) Estimated number of factors. The solid horizontal line denotes the true number of simulated factors (*K* =10). and the dashed horizontal line indicates the initial number of factors (*K* =20). Each bar represents a different model realization of the simulated data. **(b)** Pearson correlation coefficient between pairs of inferred latent factors for individual trials. For each factor, shown is the maximum correlation coefficient with any of the remaining factors. Factors were simulated to be uncorrelated. **(c)** Pearson correlation coefficient between true and inferred factors (for the top ten factors in each fit). For each factor, shown is the maximum correlation coefficient with any of the true factors.

V10 Argelaguet et al Mol Syst Biol. 14, e8124 (2018) Processing of Biological Data WS 2021/22 15

There exist different tools to identify latent factor models.

They use different methods to identify the factors.

The authors of the lastest methed termed MOFA compared their tool to the earlier tools GFA and iCluster using simulated data.

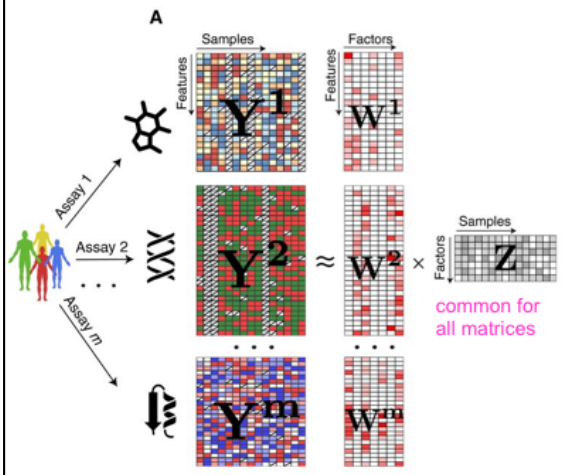
Presumably, the simulation of data is done in the reverse way from how these methods work.

In the left plot, the authors showed that MOFA identified the correct number of factors.

In the middle plot, MOFA gave the smallest correlation between the factors (which were constructed to be uncorrelated).

In the right plot, both iCluster and MOFA factors were well correlated to the correct factors.

Method 1: Multiomics Factor Analysis (concatenation-based)



Model overview: MOFA takes *M* data matrices as input ($\mathbf{Y}^1,\ldots,\mathbf{Y}^M$), one or more from each data modality, with co-occurrent samples but features that are not necessarily related and that can differ in numbers.

MOFA decomposes these matrices into a matrix of factors (**Z**) for each sample and *M* weight matrices, one for each data modality ($\mathbf{W}^1,..,\mathbf{W}^M$).

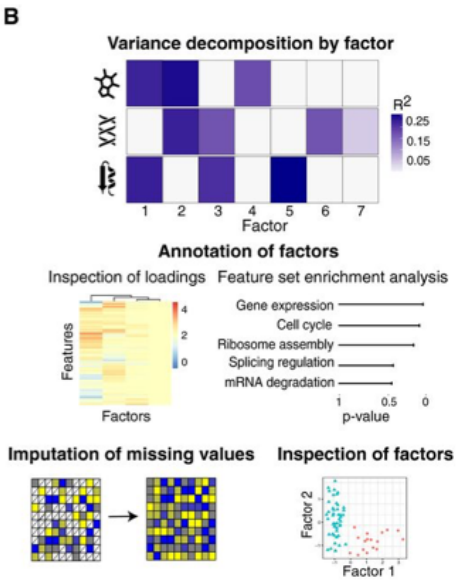
White cells in the weight matrices correspond to zeros, i.e. inactive features. Cross symbol in the data matrices denotes missing values.

Q: What are the underlying factors that drive the observed variation across samples?

V10 Argelaguet et al Mol Syst Biol. 14, e8124 (2018) Processing of Biological Data WS 2021/22 16

Link to this paper: https://www.embopress.org/doi/10.15252/msb.20178124

Multiomics Factor Analysis

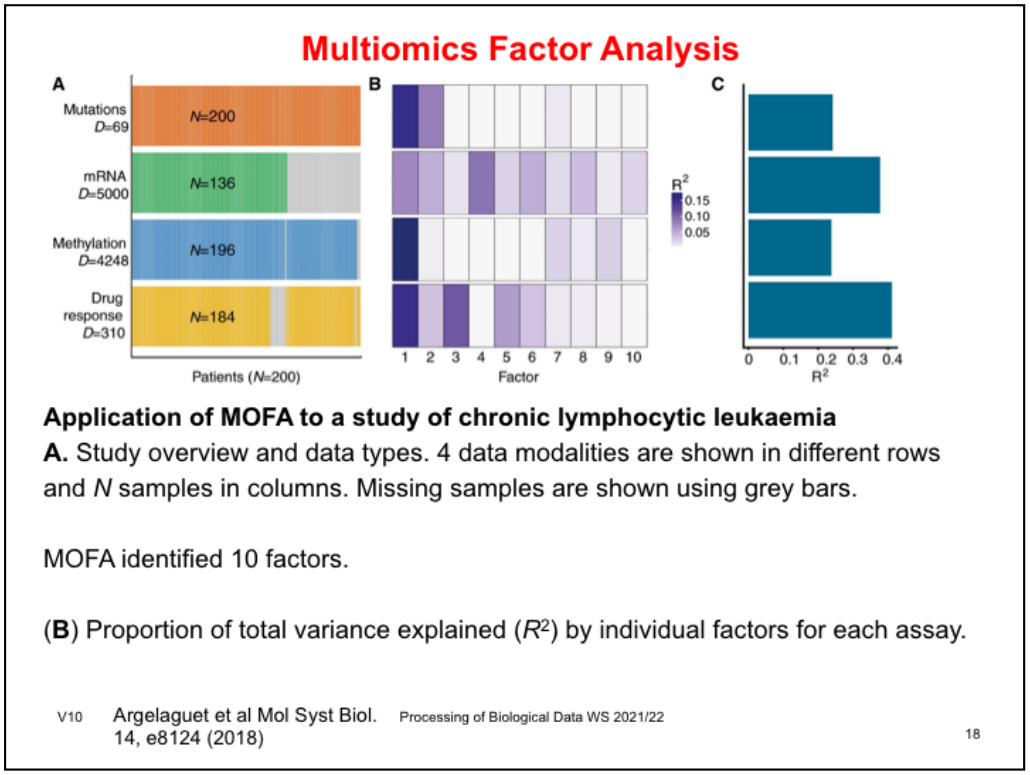


MOFA can be viewed as a generalization of principal component analysis (PCA) to multi-omics data.

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 data modality,
(ii) semi-automated factor annotation based on the inspection of loadings (coeffs in the weight matrices) and gene set enrichment analysis,
(iii) visualization of the samples in the factor space and
(iv) imputation of missing values, including missing assays.

V10 Argelaguet et al Mol Syst Biol. 14, e8124 (2018) Processing of Biological Data WS 2021/22 17

No comments

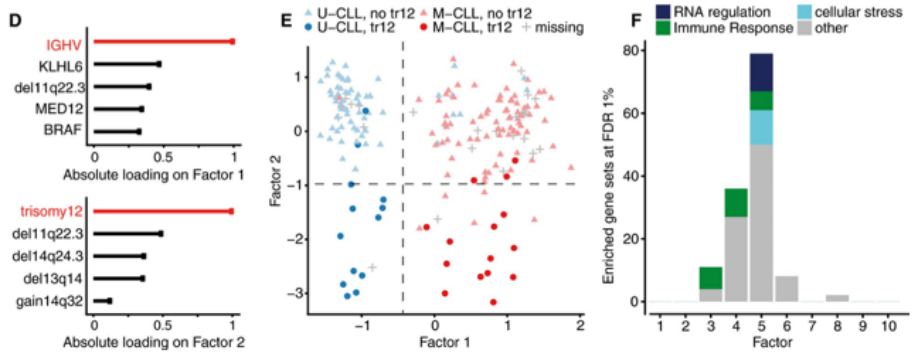


We applied MOFA to a study of chronic lymphocytic leukaemia (CLL), which combined *ex vivo* drug response measurements with somatic mutation status, transcriptome profiling and DNA methylation assays.

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

Among the 10 identified factors, factors 1 and 2 were active in most assays, indicating broad roles in multiple molecular layers (B). In contrast, 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 10 factors explained 41% of variation in the drug response data, 38% in the mRNA data, 24% in the DNA methylation data and 24% in the mutation data

Multiomics Factor Analysis



D. Absolute loadings of the top features of Factors 1 and 2 in the Mutations data.
E. Visualization of samples using Factors 1 and 2. The colors denote the IGHV status of the tumors; symbol shape and color tone indicate chromosome 12 trisomy status.
F. Number of enriched Reactome gene sets per factor based on the gene expression data (FDR < 1%). The colors denote categories of related pathways.

V10 Argelaguet et al Mol Syst Biol. 14, e8124 (2018) Processing of Biological Data WS 2021/22 19

The **loadings** describe the contributions of the features to each factor. For example, based on the top weights in the mutation data, 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. Thus, MOFA correctly identified two major axes of molecular disease heterogeneity and aligned them with two of the most important clinical markers in CLL.

Multiomics Factor Analysis

The latent factors can be used for several purposes, such as:
(1) Non-linear **dimensionality reduction**: the latent factors can be fed into non-linear dimensionality reduction techniques such as UMAP or t-SNE. This is very powerful because you can detect variability or stratifications beyond the RNA expression!
(2) **Imputation**: factors can be used to predict missing values, including entire missing assays.
(3) **Predicting clinical response**: factors can be fed into Cox models to predict patient survival.
(4) **Regressing out technical variability**: if a factor is capturing an undesired technical effect, its effect can be regressed out from your original data matrix.
(5) **Clustering**: clustering in the latent space is much more robust than in the high-dimensional space.
(6) **factor-QTL mapping**: factors are a compressed and denoised representation of your samples. This is a much better proxy for the phenotype than the expression of individual genes. Hence, a very promising area is to do eQTL's with the factors themselves!

V10 Argelaguet et al Mol Syst Biol. 14, e8124 (2018) Processing of Biological Data WS 2021/22 20

UMAP and t-SNE are modern tools to visualize e.g. the results of single cell transcriptomics. Here, the authors argue that the identified latent factors contain additional information over the transcriptomics alone.

If one is able to parametrize perfect latent factors, these factors contain basically „every information“ that can be of interest.

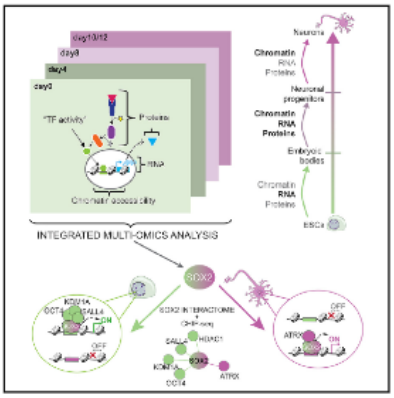
Application of MOFA

Article

Cell Systems

Genomic Rewiring of SOX2 Chromatin Interaction Network during Differentiation of ESCs to Postmitotic Neurons

Graphical Abstract



Authors

Daria Bunina, Nade Abazova, Nichole Diaz, Kyung-Min Noh, Jeroen Krijgsveld, Judith B. Zaugg

Correspondence

noh@embl.de (K.-M.N.), j.krijgsveld@dkfz.de (J.K.), zaugg@embl.de (J.B.Z.)

In Brief

In this work, Bunina et al. profiled transcriptome, proteome, and chromatin accessibility dynamics during differentiation of mouse embryonic stem cells to postmitotic neurons. They revealed extensive associations between molecular layers within and across differentiation time points and uncovered an unexpected interaction of two chromatin-bound proteins, SOX2 and ATRX, in neurons. Their role in regulating

Application of MOFA:
https://www.sciencedirect.com/science/article/pii/S2405471220301885

Cellular differentiation requires dramatic changes in chromatin organization, transcriptional regulation, and protein production. To understand the regulatory connections between these processes, Bunina et al. generated proteomic, transcriptomic, and chromatin accessibility data during differentiation of mouse embryonic stem cells (ESCs) into postmitotic neurons

Input to MOFA

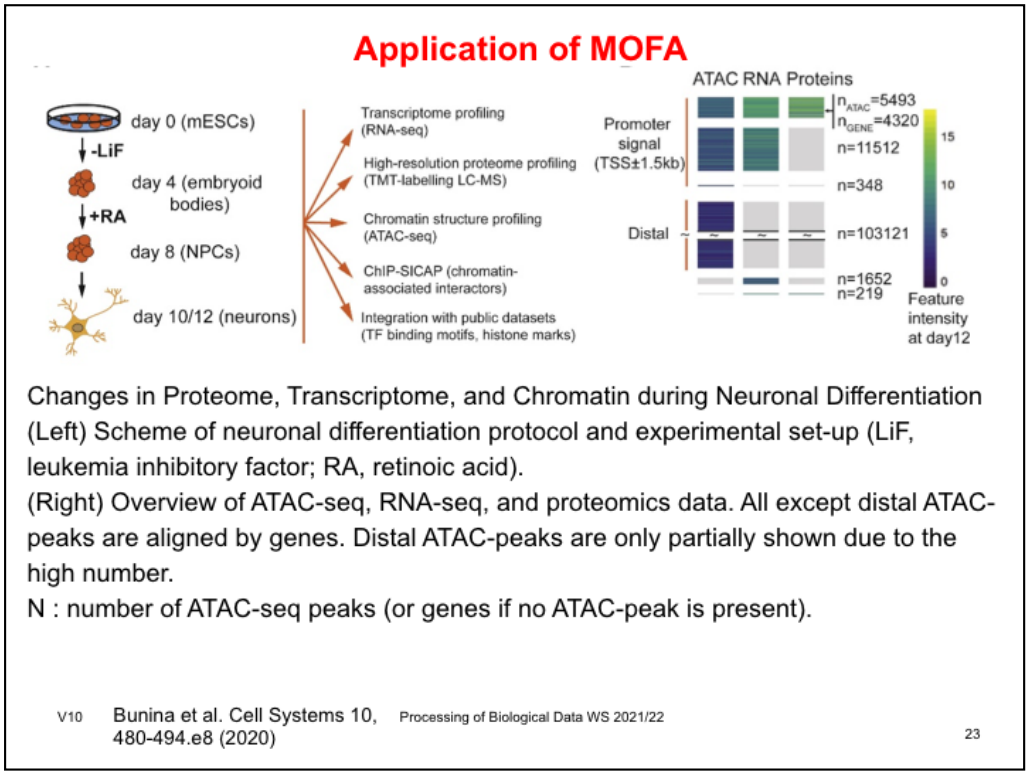
MOFA R package version 1.2.0 was used for the analysis (Argelaguet et al., 2018).

ATAC-seq peak counts (4 replicates, 4 time points), RNA gene counts (4 replicates, 4 time points) and protein counts (2 replicates, 4 time points), all variance-normalized, were used as input to the model with default parameters and 3% factor drop threshold.

The downstream analysis of the model output was performed with ranked lists of top factor loadings (genes or proteins or ATAC-seq peaks) in each data modality (converted to ensembl gene IDs) as input for gene set enrichment analysis (GSEA (Subramanian et al., 2005)), using mouse gene ontology annotations as a reference list.

Each ATAC-seq peak was linked to the nearest gene and these nearest gene lists were used for GSEA.

ATAC-seq measures chromatin accessibility.



Changes in Proteome, Transcriptome, and Chromatin during Neuronal Differentiation
(Left) Scheme of neuronal differentiation protocol and experimental set-up (LiF, leukemia inhibitory factor; RA, retinoic acid).
(Right) Overview of ATAC-seq, RNA-seq, and proteomics data. All except distal ATAC-peaks are aligned by genes. Distal ATAC-peaks are only partially shown due to the high number.
N : number of ATAC-seq peaks (or genes if no ATAC-peak is present).

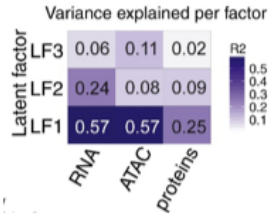
(Left) The differentiation protocol transforms mouse ESCs into glutamatergic neurons.

Briefly, ESCs were cultured on feeder-free gelatin-coated plates for 2 passages in ESC medium containing 20 ng/ml LIF protein (leukemia inhibitory factor).

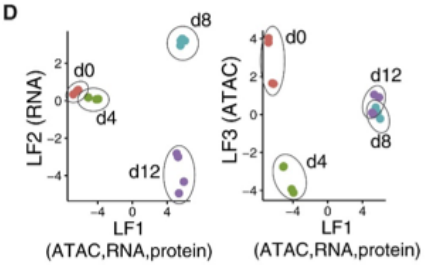
Differentiation starts upon transfer of the cells and removal of LIF from the medium, leading to the formation of embryoid bodies.

On days 4 and 6, retinoic acid (RA) at a final concentration of 5 μM was added to the medium.

Application of MOFA



(Top) Variance explained by latent factors (LFs) identified with MOFA is shown for each dataset.
LF1 explained the majority of variance in all three layers whereas LF2 and LF3 specifically explain RNA and ATAC-seq data, respectively.



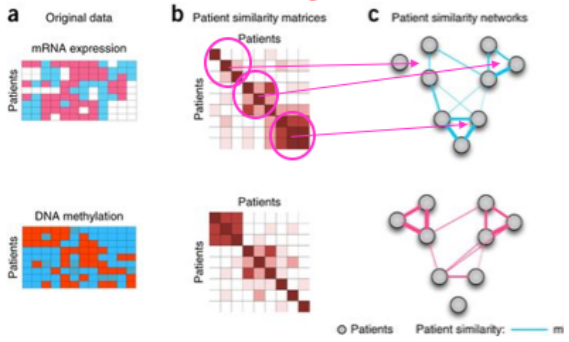
(Bottom) scatterplots of samples projected to LF1 versus LF2 (left), and LF1 versus LF3 (right).
d0 stands for „day 0“ at the start of the differentiation protocol.

V10 Bunina et al. Cell Systems 10, 480-494.e8 (2020) Processing of Biological Data WS 2021/22 24

MOFA identified three latent factors (LFs) that explained a major part of the variance in at least one dataset (Top figure).

(Bottom) The common factor (LF1) separated early (days 0 and 4) from late (days 8 and 10/12) differentiation, suggesting that drastic changes in cellular processes after neural induction strongly involve all three regulatory layers.

Method 2: Similarity Network Fusion (transformation-based)



Aim of SNF: discover patient subgroup clusters.

Huang et al. Front Genet. 8: 84 (2017)

Anna Goldenberg

(a) Example representation of mRNA expression and DNA methylation data sets for the same cohort of patients.

(**b**) Patient-by-patient similarity matrices for each data type.

(**c**) Patient-by-patient similarity networks, equivalent to the patient-by-patient data. Patients are represented by nodes and patients' pairwise similarities are represented by edges.

V10 Wang et al. Nature Methods 11, 333 (2014) Processing of Biological Data WS 2021/22 25

The method „**similarity network fusion**“ was developed by the group of Anna Goldenberg at Toronto.

This is the paper that presented SNF:
https://www.nature.com/articles/nmeth.2810

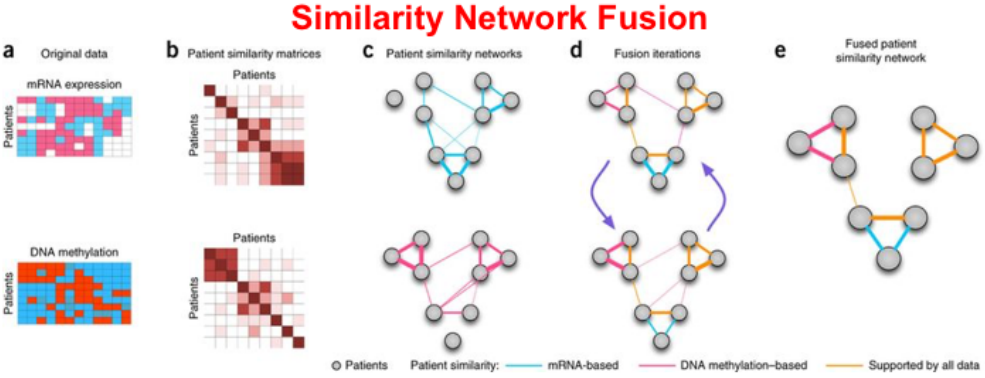
SNF follows a very intuitive principle. Shown here are only the first steps of the algorithm.

(a) Illustrates the raw data.

(b) Based on the data of (a) one computes the similarity between all pairs of samples (here: patients), e.g. by the measure of cosine similarity or any other suitable definition.

(c) The pairwise similarities are converted into edge weigts of a patient-vertex graph. In the upper row, the strongest similarities are found for the 3 bottom right node pairs. In the right figure, this is represented by „thick“ edges. In the lower row, the highest similarities are observed in the top left corder of the similarity matrix (middle figure). This then leads to thicker lines between the top nodes

You will implement SNF in assignment #5.



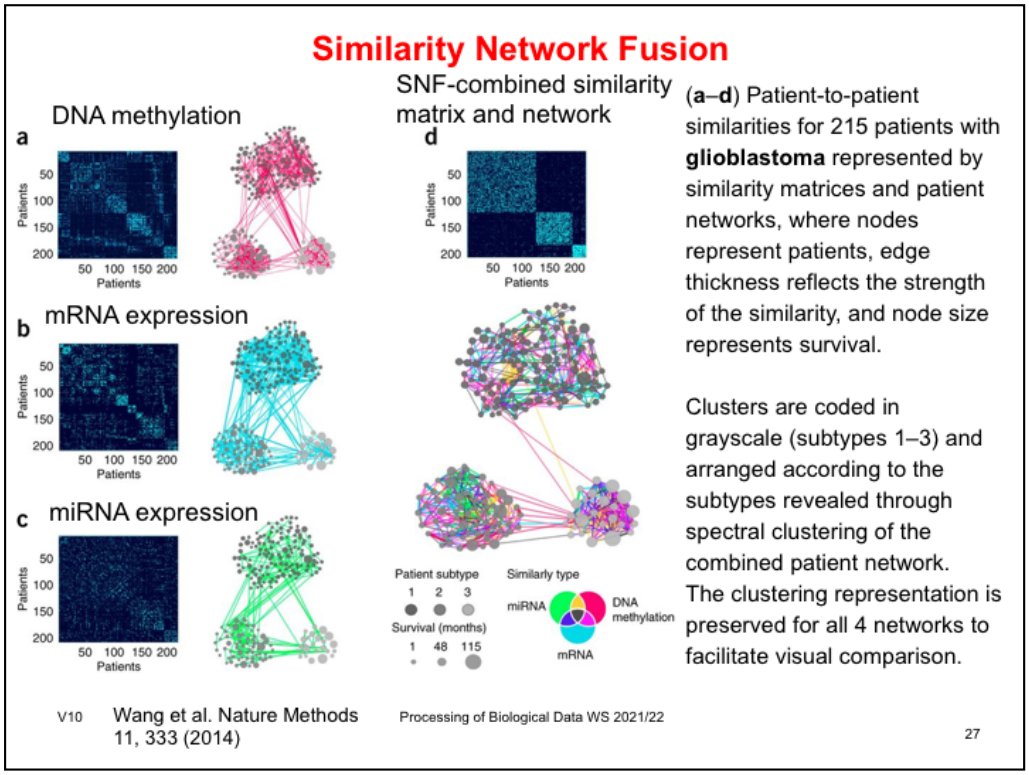
(**d**) Network fusion by SNF iteratively updates each of the networks with similarity information from the other networks, making them more similar with each step.

(**e**) The iterative network fusion results in convergence to the final fused network. Edge color indicates which data type has contributed to the given similarity.

V10 Wang et al. Nature Methods 11, 333 (2014) Processing of Biological Data WS 2021/22 26

In step (d), an iterative exchange takes place between the networks representing different data types.

In (e), only one converged network remains that represents the consensus or average of the different networks.



This example was presented by the Goldenberg group in their SNF paper. In the literature, there are differing opinions whether there exist 2, 3, or 4 subgroups of GBM patients.

The figures on the left represent 3 different data types for a group of 215 glioblastoma patients: DNA methylation (1,491 genes), mRNA expression (12,042 genes) and miRNA expression (534 miRNAs)

As expected, networks built using a single data type yielded very different patterns supports of patient similarity. For example, DNA methylation strongly supports connectivity in the smallest patient cluster (a), whereas mRNA expression supports similarity in the medium-sized cluster (b). It is difficult to discern patterns in the patient-similarity network based on miRNA data alone (c). The **fused network** gives a **much clearer picture** of **clustering** in this set of patients with GBM, illustrated by the tightness of connectivity within clusters and relatively few edges between clusters (d).

The **smallest cluster** (subtype 3) corresponds to the previously identified **IDH subtype** consisting of younger patients with a substantially more favorable prognosis. All patients with an *IDH1* mutation for whom the information was available (n = 14 patients) belong to this cluster. Subtype 1 patients had a favorable response to temozolomide (TMZ), a drug commonly used to treat GBM.

The network analysis goes beyond subtyping. Each edge in the fused network is colored by the data type(s) that contributed to the given similarity. A multicolor cluster means that no single data type or combination support patient similarity across GBM. Most edges were supported by at least two data types: 49.5% of all patient similarities (edges) were due to two data types, 17.2% were supported by all three data types and the remaining 33.3% of the edges were supported by only one data type, with strong enough similarity that those edges remained prominent in the fused network.

The GBM analysis highlights 3 important features of the network-based integrative approach:

(i) the ability to detect common as well as complementary signals;

(ii) the ability to reduce noise by aggregating across multiple types of data; and

(iii) insight into the relative importance of each data source for determining patient similarity, thus refining our understanding of the heterogeneity within each subtype.

Application of SNF: pancreatic ductal adenocarcinoma

Pancreatic ductal adenocarcinoma (PDAC) is an **aggressive disease** that typically presents at an advanced stage and is refractory to most treatment modalities.

A whole-exome sequencing study of pancreatic cancer identified a large number of **mutations** and somatic **copy number alterations** (SCNAs) that alter the function of many key **oncogenes** and **tumor suppressor genes**, including KRAS, TP53, SMAD4, and CDKN2A.

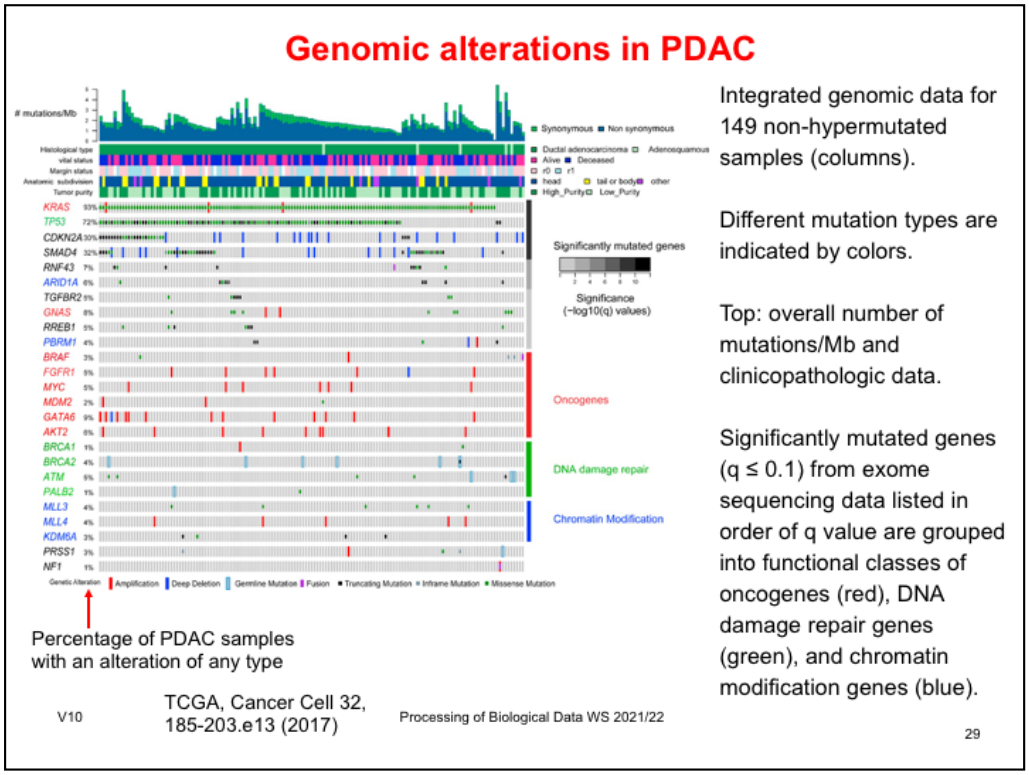
Germline alterations in DNA damage repair genes such as BRCA1, BRCA2, PALB2, or ATM give rise to **genomic instability** in a subset of PDACs.

The majority of PDACs harbor **complex chromosomal rearrangement** patterns.

TCGA, Cancer Cell 32, 185-203.e13 (2017)

Link to this paper:
https://www.sciencedirect.com/science/article/pii/S1535610817302994

This is an example where a different group (here, the TCGA consortium) applied the SNF tool.



Whole-exome sequencing identified somatic DNA alterations, including single nucleotide variants (SNVs), small insertions and deletions (indels), and SCNAs.

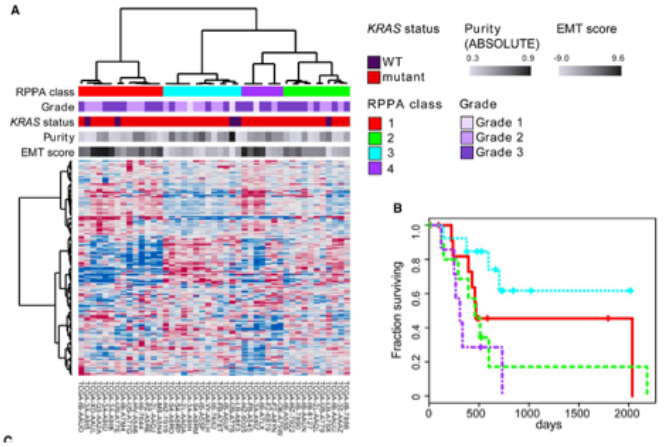
Significant recurrent mutations were identified in the genes KRAS, TP53, CDKN2A, SMAD4, RNF43, ARID1A, TGFβR2, GNAS, RREB1, and PBRM1.

The authors also observed recurrent mutations in several genes at false discovery rates (FDRs) above a threshold of q = 0.1, including mutations in other known oncogenes, DNA damage repair genes, and chromatin modification genes.

About definition of „margin“, see https://www.cancer.gov/publications/dictionaries/cancer-terms/def/margin

The edge or border of the tissue removed in cancer surgery. The margin is described as negative or clean when the pathologist finds no cancer cells at the edge of the tissue, suggesting that all of the cancer has been removed. The margin is described as positive or involved when the pathologist finds cancer cells at the edge of the tissue, suggesting that all of the cancer has not been removed.

RRPA profiles: protein concentrations



(A) Unsupervised consensus clustering of RPPA protein expression data for 45 of the 76 high-purity samples.

(B) Cox survival analysis between clusters ($p = 0.045$).

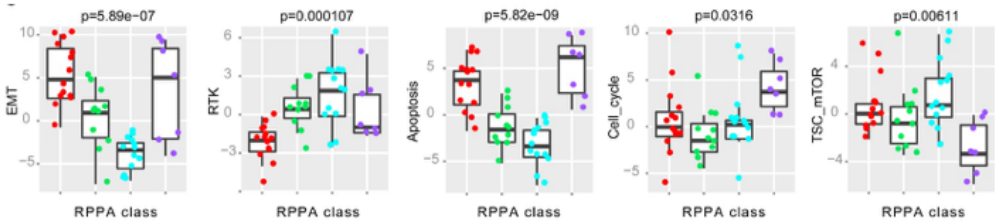
TCGA, Cancer Cell 32, 185-203.e13 (2017)

Unsupervised consensus clustering of protein expression measured on a 192-antibody array for 45 of the 76 high-purity samples identified four clusters (panel A) that exhibited significant differences in survival (panel B).

„RPPA“ stands for „reverse phase protein array“ and measures protein concentrations.

KRAS is a member of the RAS-signaling cascade that transmits external growth signals to the cell. 93% of the samples carry KRAS mutations.

Genomic alterations in PDAC



X-axis: 4 RPPA classes (red, green, blue, violet) – see previous slide

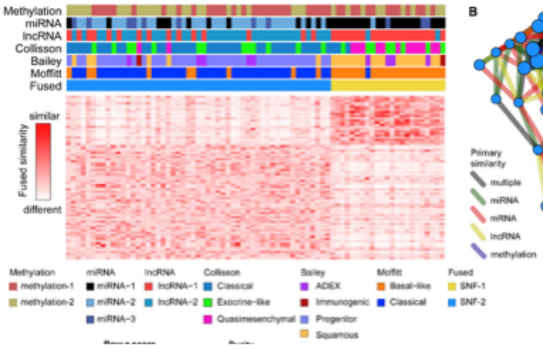
Differences in proteomic pathway activity scores across RPPA cluster/class

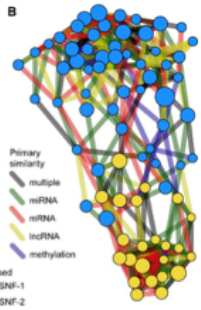
TCGA, Cancer Cell 32, 185-203.e13 (2017)

Analysis of pathway activity between clusters identified significantly different scores for epithelial-to-mesenchymal transition (EMT), apoptosis, TSC/mTOR, cell_cycle, and receptor tyrosine kinase (RTK) pathways.

Tumors from cluster 3 (light blue), which had better survival (see panel B), were characterized by low EMT and apoptosis pathway activity, but high TSC/mTOR and RTK activity.

PDAC: integrated analysis with SNF





(Left) Integrated clustering of methylation, miRNA, lncRNA, and mRNA data using Similarity Network Fusion (SNF) on high-purity samples.

(Right) Network fusion diagram of the 2 integrated clusters: each node is a sample, with node color indicating the SNF cluster and node size proportional to ABSOLUTE purity.
Edges are colored according to the datatype giving the strongest similarity between patients.
Nodes positioned in between the top and bottom clusters generally have lower purity, reflecting the weaker signal for molecular classification.

TCGA, Cancer Cell 32, 185-203.e13 (2017)

Link to this paper:
https://www.sciencedirect.com/science/article/pii/S1535610817302994

To integrate information from multiple platforms, the authors performed Similarity Network Fusion (SNF), which has been shown to produce homogeneous, clinically relevant subtypes in multiple TCGA studies.

They applied SNF to the high-purity cohort using sample-to-sample similarities derived from mRNA, miRNA, and DNA methylation. They found a two-cluster solution that was independent ($p = 0.79$) of tumor purity and a three-cluster (plus one outlier) solution that was associated ($p = 0.025$) with tumor purity.

The clusters defined by SNF were highly concordant with results obtained from miRNA, lncRNA, or mRNA alone.

Rethink: why do we do analysis of omics-data?

(1) Analysis of general phenomena

- Which genes/proteins/miRNAs control certain cellular behavior?
- Which ones are responsible for diseases?
- Which ones are the best targets for a therapy?

(2) We want to help an individual patient

- Why did he/she get sick?
- What is the best therapy for this patient?

This is a short reflection on the whole course „processing of biological data“.

In principle, this review should come at the end of the last lecture next week, but that lecture will probably end with some recommendations how to prepare for the final oral exam. Therefore I have moved these slides to the end of this lecture.

In the course, we have covered a number of techniques for preprocessing of data. Often, we need to decide before the analysis which samples to include and which samples to eliminate because e.g. too many data points are lacking.

The criteria for our decisions will – on the one hand – depend on how much data is available. We will address this on the next slide.

Also, the criteria will depend on the research question we ask.

Do we want to analyze a general phenomenon? Or do we want to help an individual patient?

Rethink: how should we treat omics-data?

(1) Analysis of general phenomena

- We typically have „enough“ data + we are interested in very robust results
- -> we can be **generous** in **removing problematic data** (low coverage, close to significance threshold, large deviations between replicates …)
- We can/should also **remove outliers** and **special cases** from the data because we are interested in the general case.

The recommendation formulated here only apply in the case when we have „enough“ data.

E.g. in machine learning, a role of thumb is that we should have at least 5 data points for every feature that we use in a regression model/ML model. If we use n features, we should have > 5^n data points.

Rethink: how should we treat omics-data?

(2) We want to help an individual patient

- Usually we only have 1-3 data sets for this patient (technical replicates)

  we cannot remove any of this data

  if there exist technical problems with the data, we need to find a practical solution for this because the patient needs to be treated

- If there are problems in the data, we have to report this together with our results -> low confidence in the result or in parts of the result

For a single patient, we will typically only get 1-3 technical replicates. This is it. We have to live with this data and do „the best we can“.

Outlook

Insights gained from omics approaches to disease are mostly comparative.

We compare omics data from healthy and diseased individuals and assume that this difference is directly related to disease.

However, in complex phenotypes both "healthy" and "disease" groups are heterogeneous with respect to many **confounding factors** such as population structure, cell type composition bias in sample ascertainment, batch effects, and other unknown factors.

E.g. sex is one of the major determinants of biological function, and most diseases show some extent of sex dimorphism.
Thus, any personalized treatment approaches will have to take sex into account.

Differentiating causality from correlation based on omics analysis remains an open question.

V10 Hasin et al. Genome Biology 18:83 (2017) Processing of Biological Data WS 2021/22 36

After completing the analysis of a data set, we will (hopefully) arrive at some statistically significant conclusions. Does this mean that these conclusions are „true"?

Yes, they are true given the data we analyzed. But this may not necessarily mean that they are true in a biological / medical sense.

The reason is that – sometimes – the derived conclusions are affected by additional **confounding factors**. A well-known example is the question whether drinking coffee increases your risk of cancer. See

https://cebp.aacrjournals.org/content/25/6/951.long

for the latest update on this issue.

Previous epidemiologic studies had evaluated the potential association between coffee consumption and risk of lung cancer, but the results were not consistent. An important aspect to consider is the potential confounding effect from **tobacco smoking**, a known cause of lung cancer, which in many populations is associated with coffee drinking. An positive association between coffee drinking and lung cancer risk is justified by the fact that coffee contains agents which may cause cancer under experimental conditions, such as acrylamide, which is formed at very low levels during the roasting of coffee beans. In contrast, other agents present in coffee have been reported to exert an anticarcinogenic effect, including the diterpenes cafestol and kahweol.

Anyhow, this study concluded with good news: „**when the potential confounding effect from smoking is controlled for, coffee drinking does not appear to be a lung cancer risk factor.** “

We will turn to this issue – the analysis of confounding factors – in our last lecture next week.