V13 Multi-omics data integration

Program for today:

- Data integration methods – overview II (see also V12)

- Similarity network fusion
- Multiomics factor analysis

- Rethink data analysis

- Results of course evaluation

Benefits of multi-omics data

(1) 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.

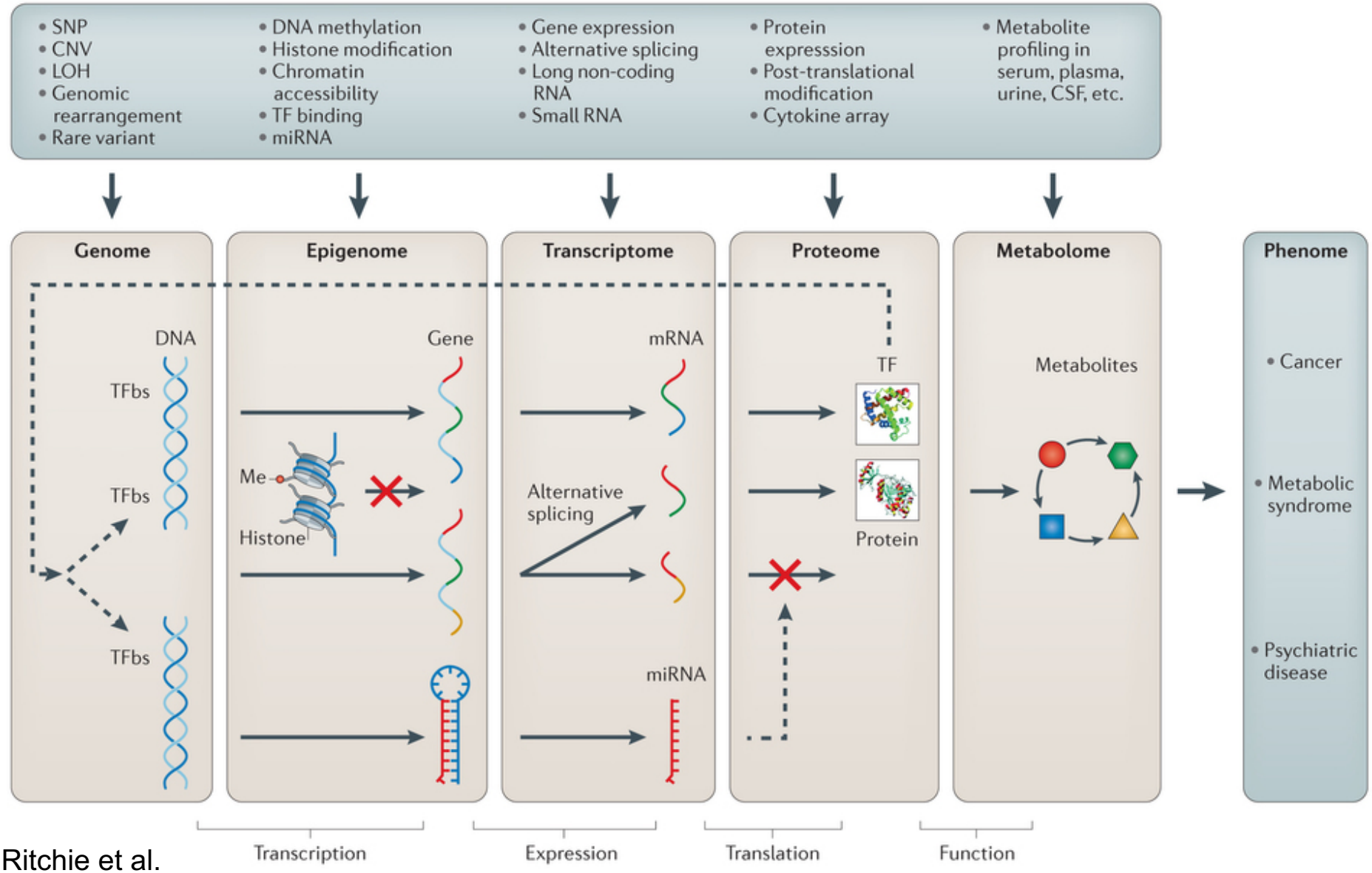
Main motivation behind combining different data sources:

Identify genomic factors and their interactions
that **explain** or **predict disease risk.**

Ritchie et al.

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

Multi-omics: genotype -> phenotype mapping



Ritchie et al.

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

Nature Reviews | Genetics

Methods for data integration

In V12, we saw that there are network-based and Bayesian approaches.

However, there exists another basic classification of data integration methods:

(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: eQTL analysis

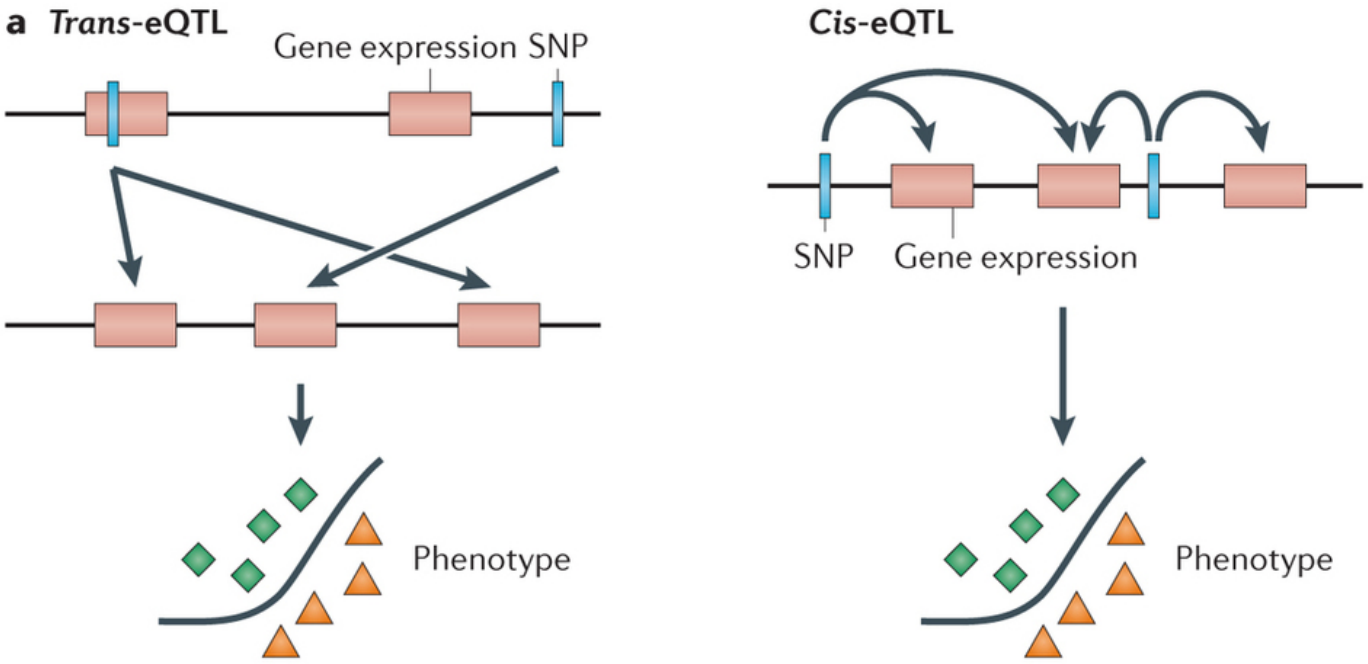
Steps: (1) associate SNPs with phenotype; filter by significance threshold

(2) Test 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

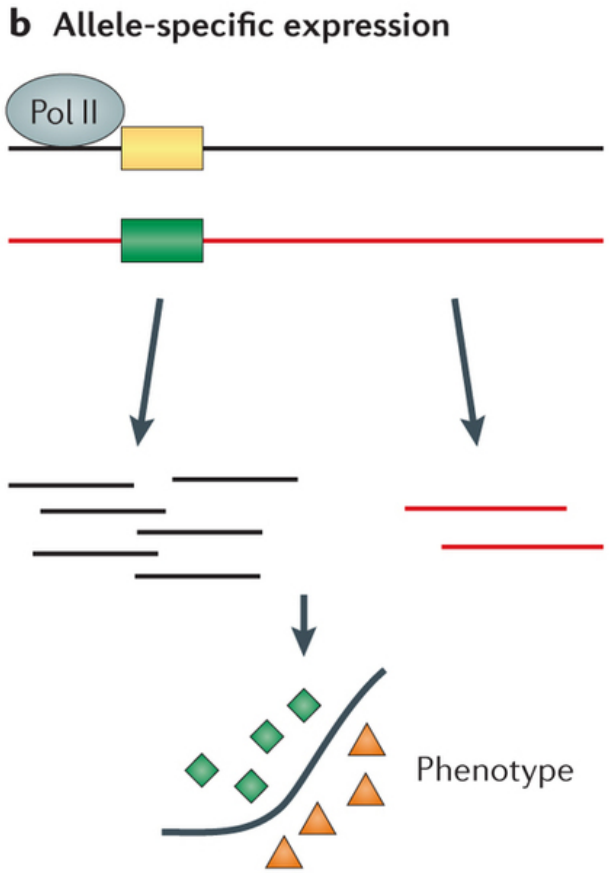
Cis-eQTL: effect on nearby gene



Ritchie et al.

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

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.

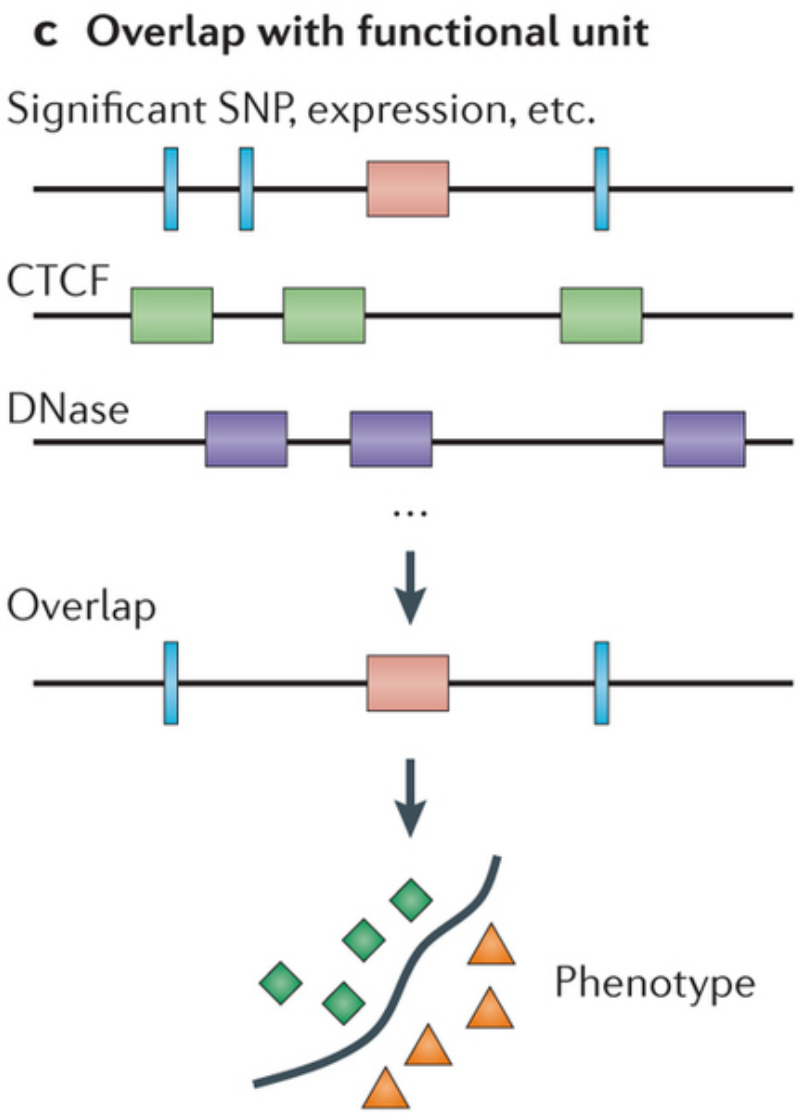
ASE analysis tests whether the maternal or paternal allele is preferentially expressed.

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

Ritchie et al.

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

Multi-staged analysis: domain knowledge overlap



Ritchie et al.

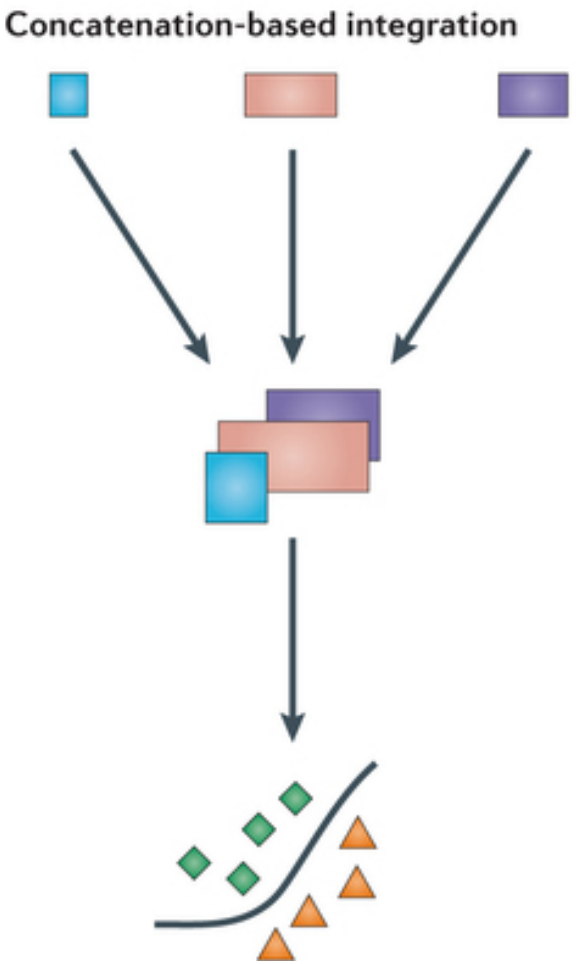
Nature Rev Genet **16**, 85 (2015)

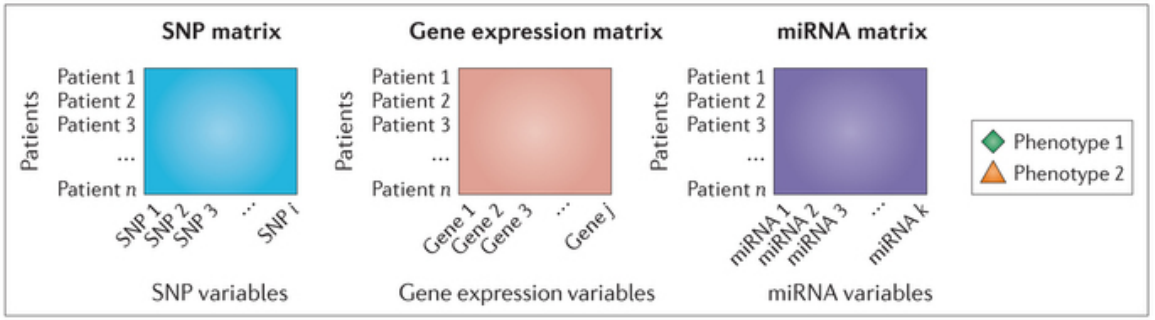
Domain knowledge overlap involves a two-step analysis:
(1) an initial association analysis is performed at the SNP or gene expression variable.

(2) This is followed by the annotation of the significant associations with knowledge generated by other biological experiments.

This approach enables the selection of association results with functional data to corroborate the association.

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.

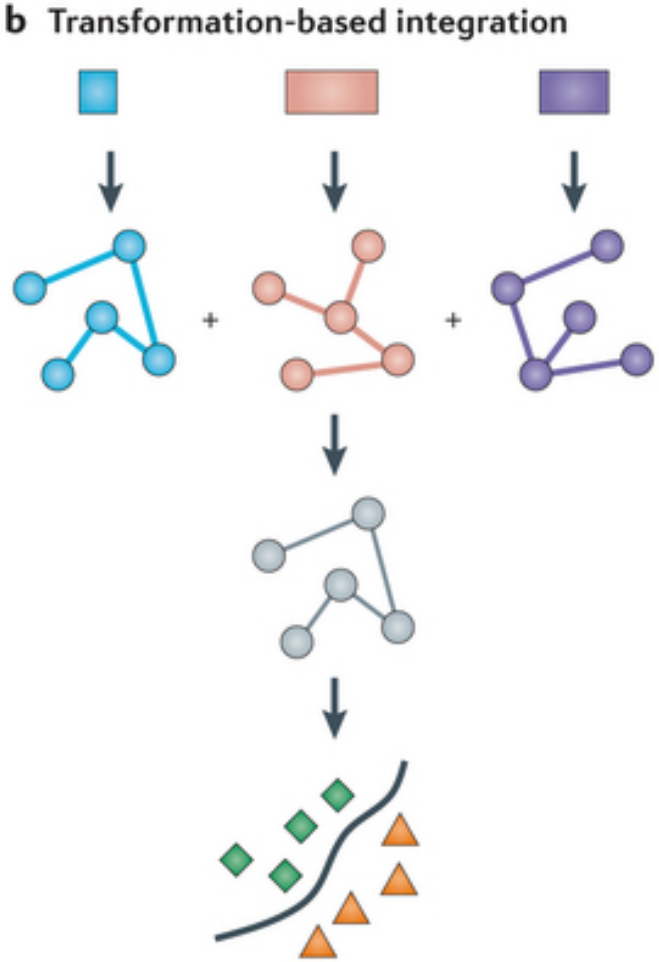
Challenges:

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

Ritchie et al.

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

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.

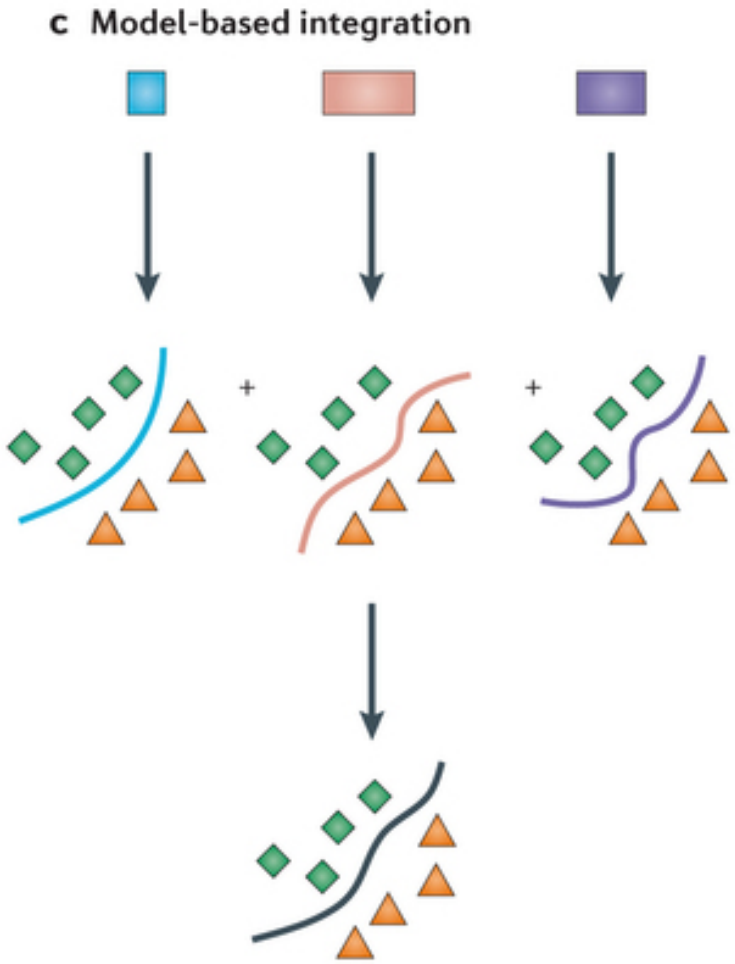
It has most “support” from the 3 initial trees.

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

Ritchie et al.

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

Meta-dimensional analysis: model-based integration



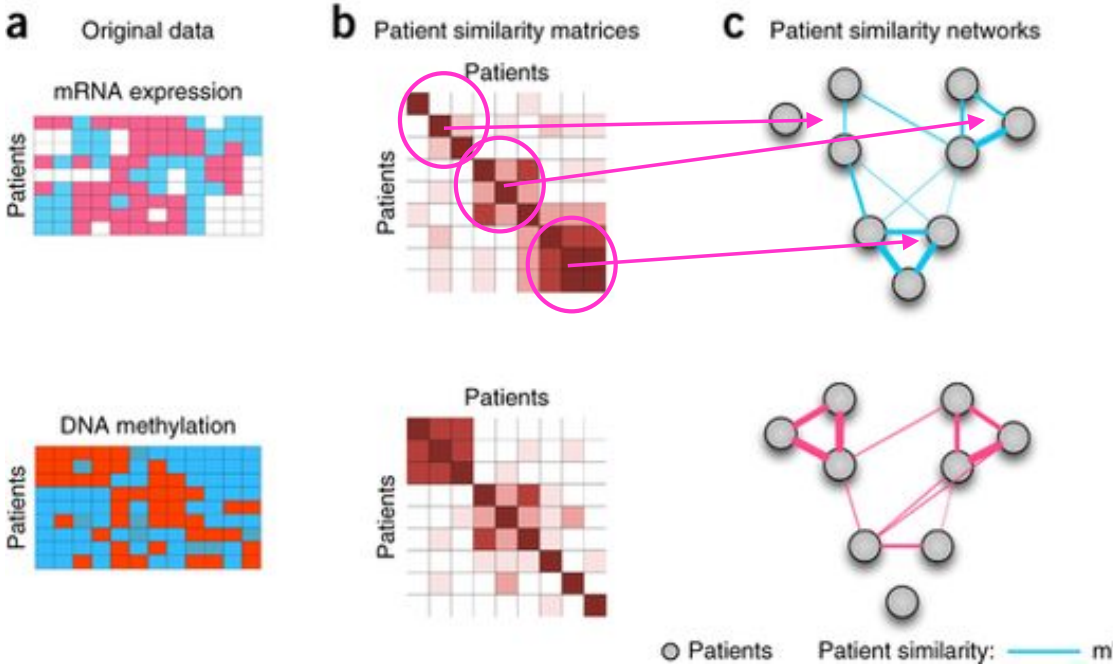
c | Model-based integration is the process of performing analysis on each data type independently.

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)

Method 1: Similarity Network Fusion



Aim of SNF: discover patient subgroup clusters.

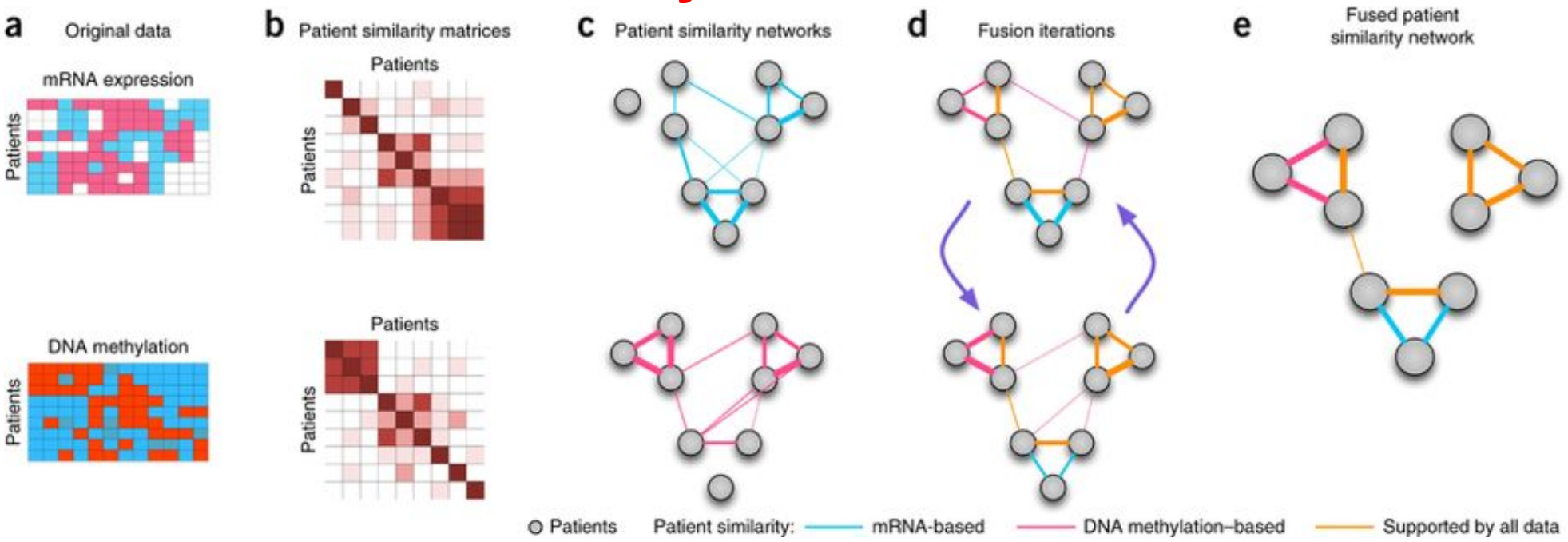
Huang et al. Front Genet. 8: 84 (2017)

(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.

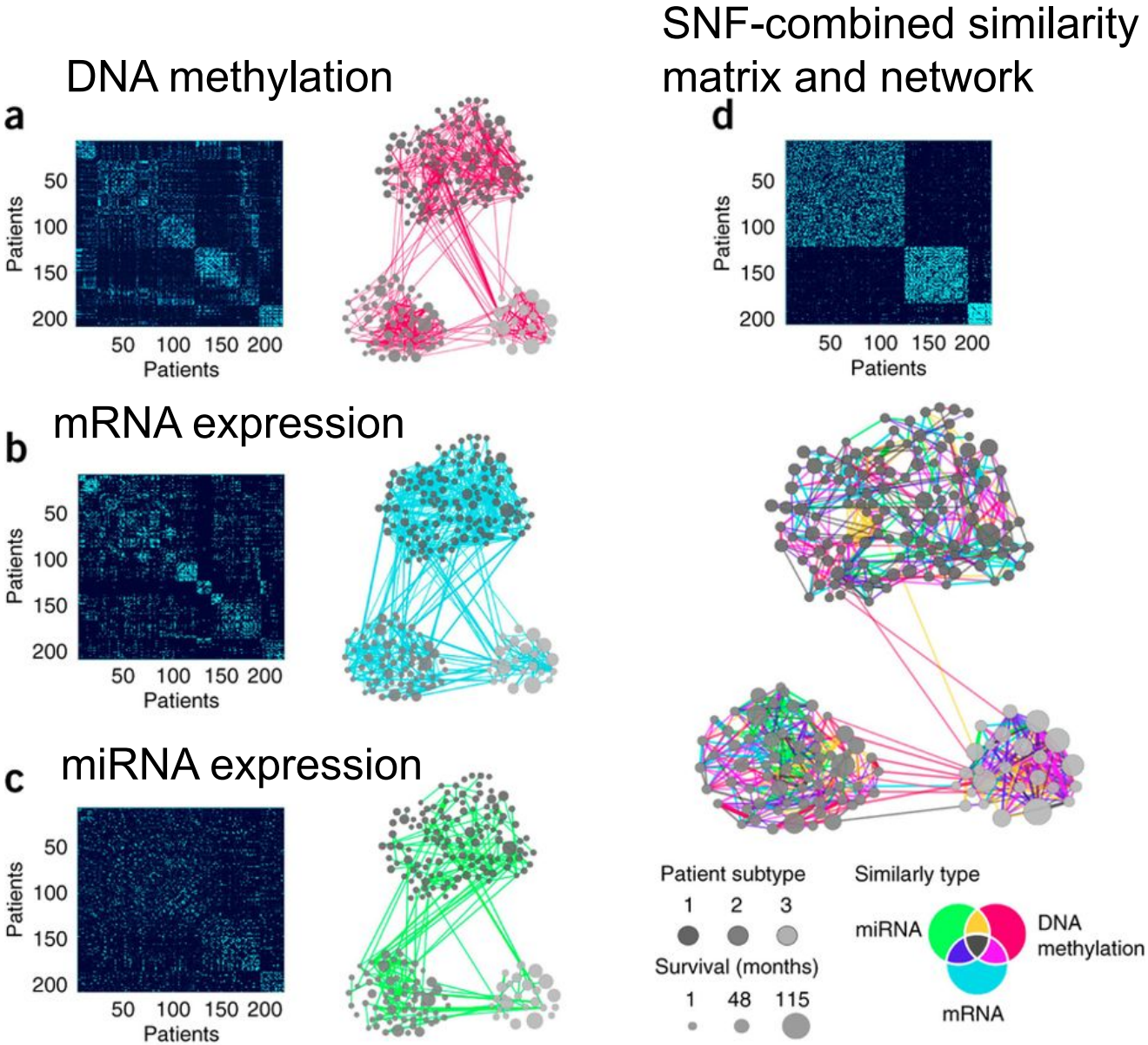
Similarity Network Fusion



(**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.

Similarity Network Fusion

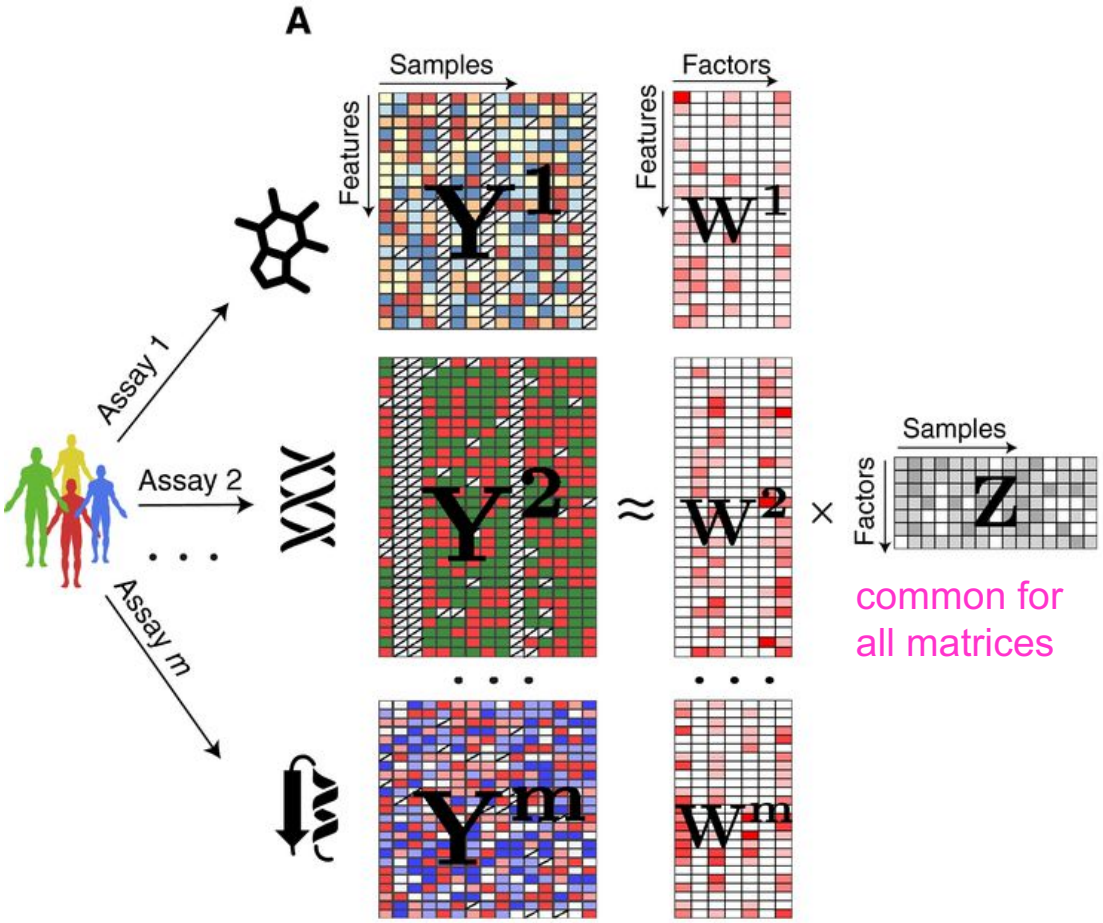


(**a**–**d**) Patient-to-patient similarities for 215 patients with **glioblastoma** represented by similarity matrices and patient networks, where nodes represent patients, edge thickness reflects the strength of the similarity, and node size represents survival.

Clusters are coded in grayscale (subtypes 1–3) and arranged according to the subtypes revealed through spectral clustering of the combined patient network. The clustering representation is preserved for all 4 networks to facilitate visual comparison.

V13 Wang et al. Nature Methods 11, 333 (2014)

Method 2: Multiomics Factor Analysis



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

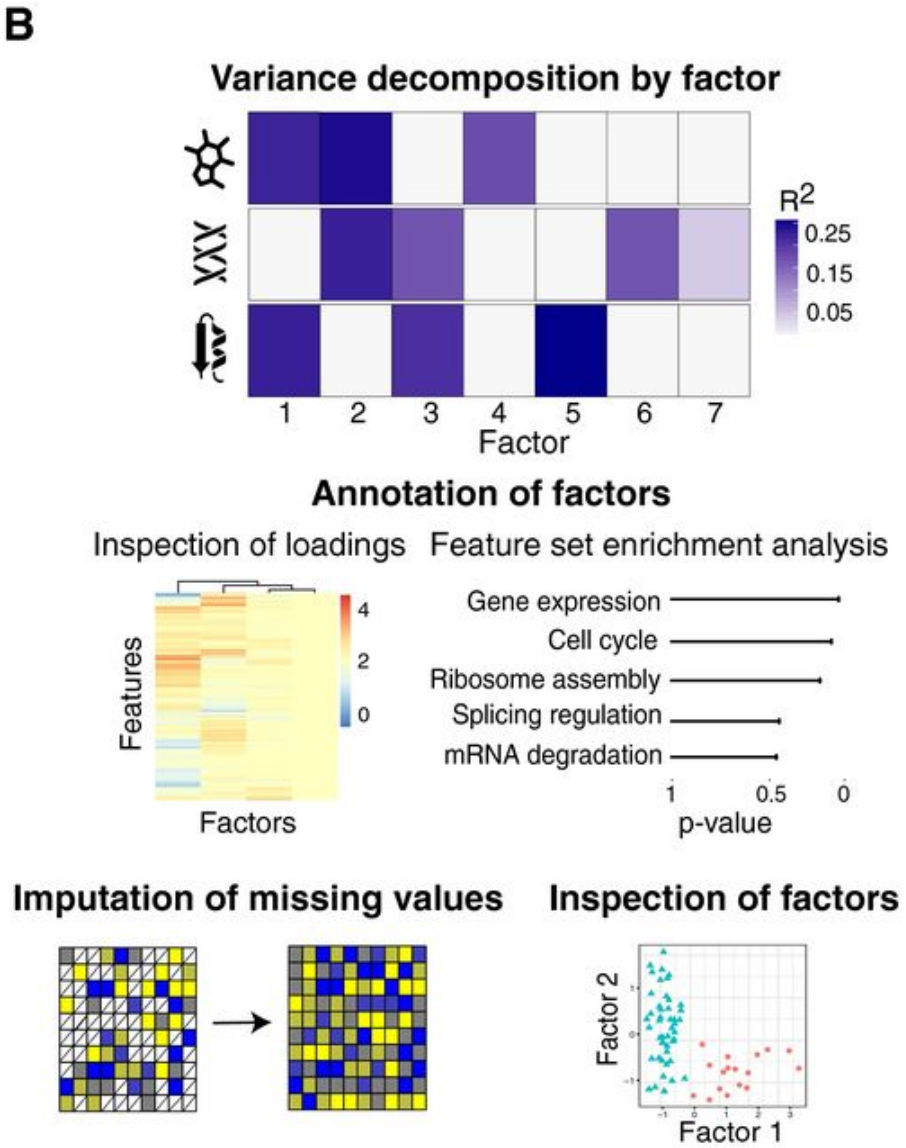
Model overview: MOFA takes *M* data matrices as input ($\mathbf{Y}^1$,…, $\mathbf{Y}^M$), one or more from each data modality, with co-occurrent samples but features that are not necessarily related and that can differ in numbers.

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.

Argelaguet et al Mol Syst Biol. 14, e8124 (2018)

Multiomics Factor Analysis

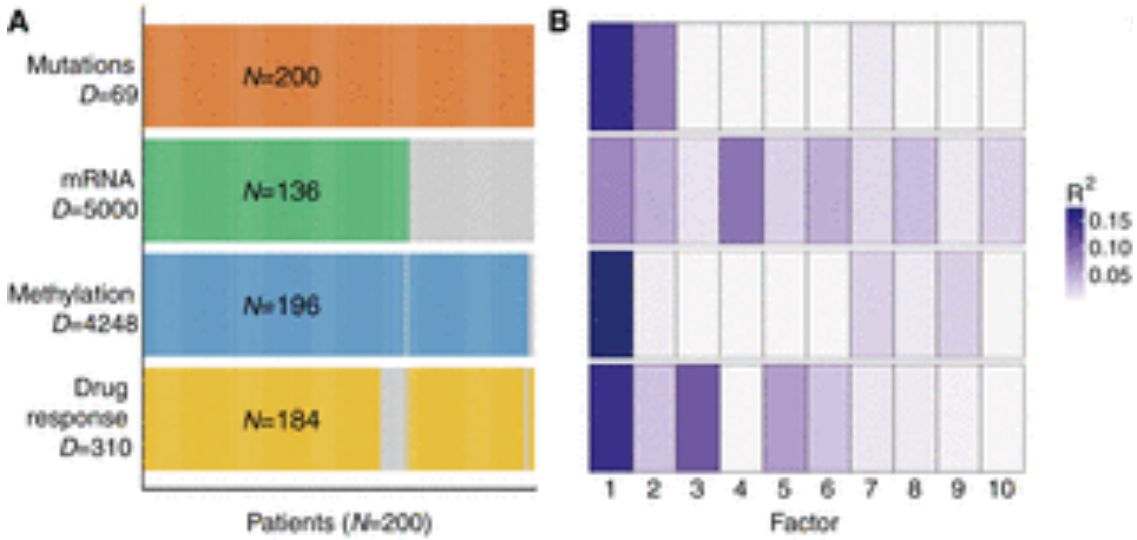


MOFA can be viewed as a generalization of principal component analysis (PCA) to multi-omics data.

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.

V13 Argelaguet et al Mol Syst Biol. 14, e8124 (2018)

Multiomics Factor Analysis



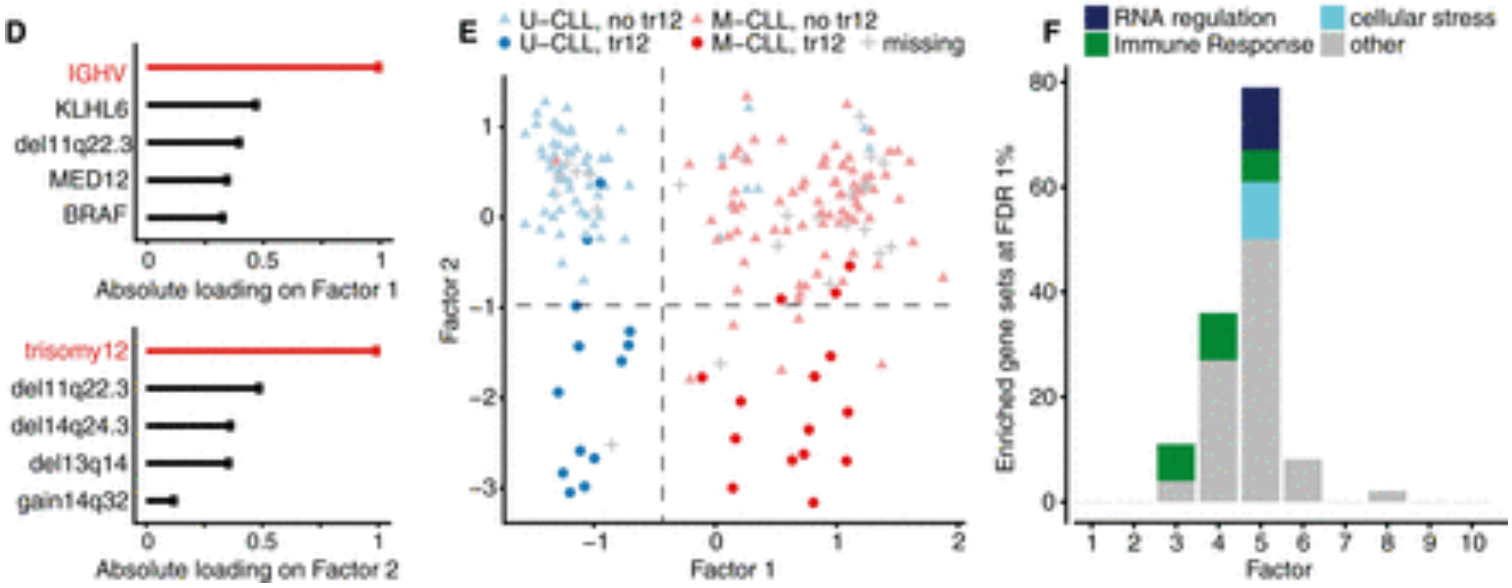
Application of MOFA to a study of chronic lymphocytic leukaemia

A. Study overview and data types. 4 data modalities are shown in different rows and *N* samples in columns. Missing samples are shown using grey bars.

MOFA identified 10 factors.

(**B**) Proportion of total variance explained (R^2) by individual factors for each assay.

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.

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?

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 remove outliers and special cases from the data because we are interested in the general case.

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

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.

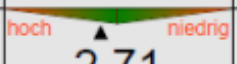
E.g. Sex is one of the major determinants of biological function, and most diseases show some extent of sex dimorphism. Thus, any personalized treatment approaches will have to take sex into account.

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

Relevant slides for written exam on Feb 25, 2019

Lecture	Slides
1	15, 16, 18, 27-39
2	4,6, 9, 14, 22-24
3	7-10, 14-29, 37, 45
4	1-4, 6, 8-18
5	3-5, 18, 20-24, 35-36
6	3 (only Hi-C), 8-10, 12-19, 23-26, 28-31
7	-
8	4-5, 8
9	all
10	2-4, 8-9, 14-15, 21-25
11	33-36
12	7-9, 17-18
13	2, 4, 11-12, 14
Material (algorithms, protocols) from all 5 assignments	

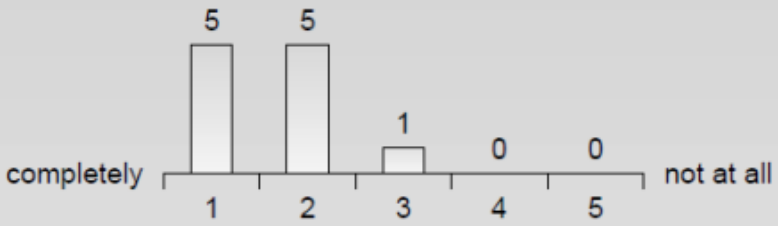
Course evaluation

	Course		Comparison		
Scale	M	SD	M	SD	N
Lecturer	1.73	0.56	1.83	0.67	1562
Structure	1.8	0.63	2.11	0.87	1562
Topic	1.73	0.61	1.99	0.87	1562
Requirements	 2.71	0.43			
Organization	1.87	0.77	1.74	0.79	1562
Overall Assessment	2	0.91	2.16	0.91	1562

Course evaluation

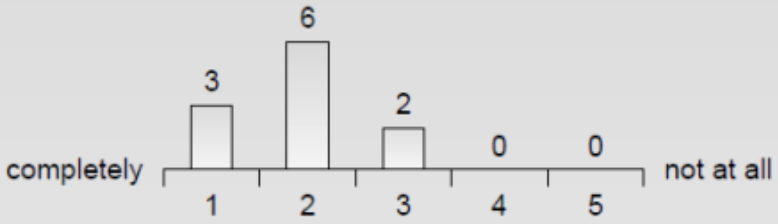
Lecturer

The lecturer was enthusiastic and motivated.



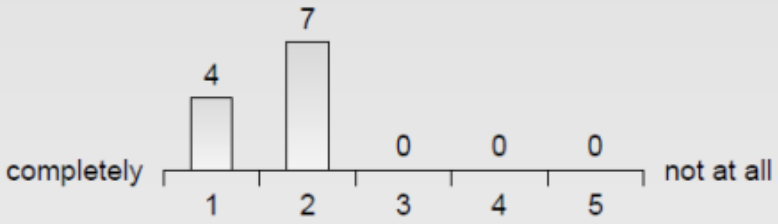
N= 11
M= 1.64
SD= 0.67
k.A.= 0

I was able to follow the pace of the lecturer.



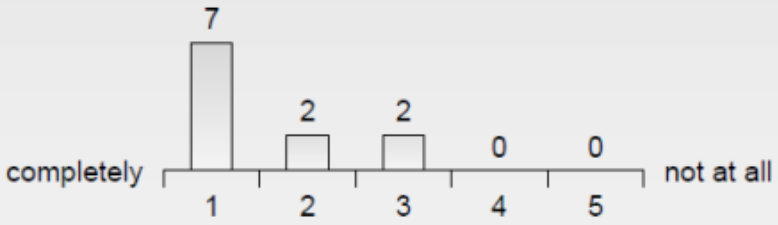
N= 11
M= 1.91
SD= 0.7
k.A.= 0

The lecturer provided a good learning and working atmosphere.



N= 11
M= 1.64
SD= 0.5
k.A.= 0

The lecturer has always been well prepared.

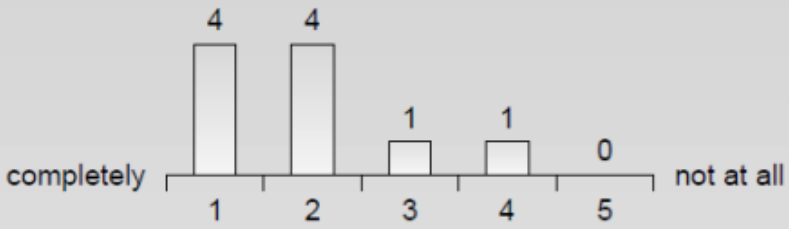


N= 11
M= 1.55
SD= 0.82
k.A.= 0

Course evaluation

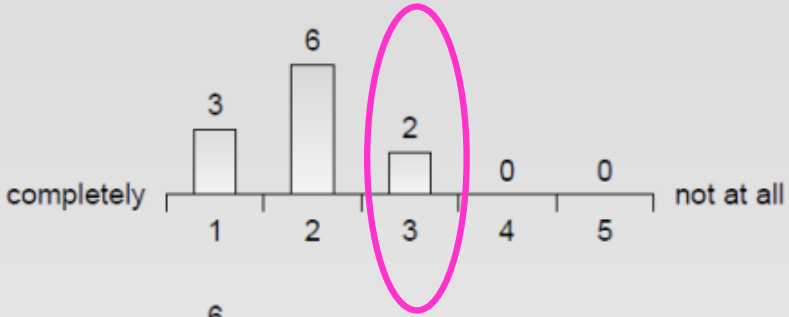
Lecturer

The lecturer was very competent.



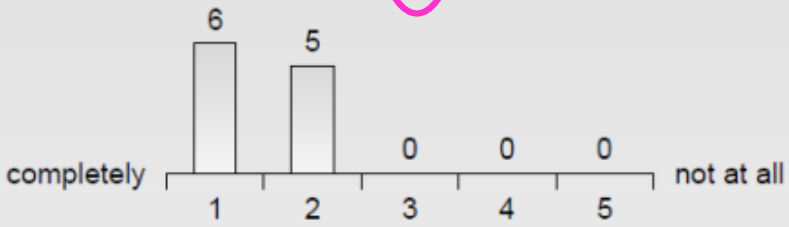
N= 11
M= 1.9
SD= 0.99
k.A.= 1

The lecturer was able to put complicated ideas across.



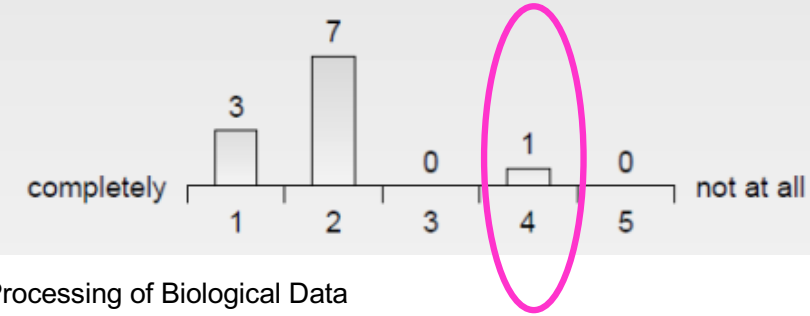
N= 11
M= 1.91
SD= 0.7
k.A.= 0

It was important to the lecturer that the participants benefitted from the course.



N= 11
M= 1.45
SD= 0.52
k.A.= 0

The lecturer motivated the participants.

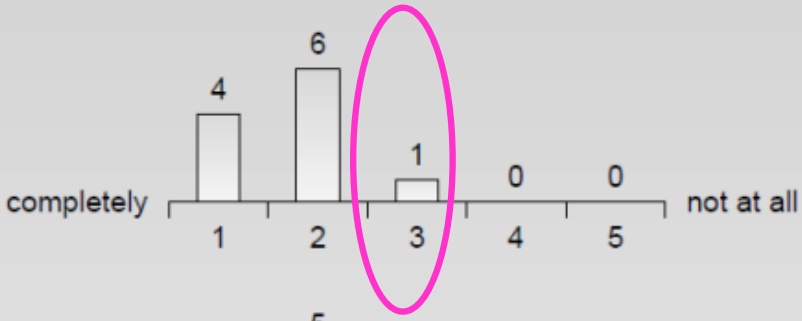


N= 11
M= 1.91
SD= 0.83
k.A.= 0

Course evaluation

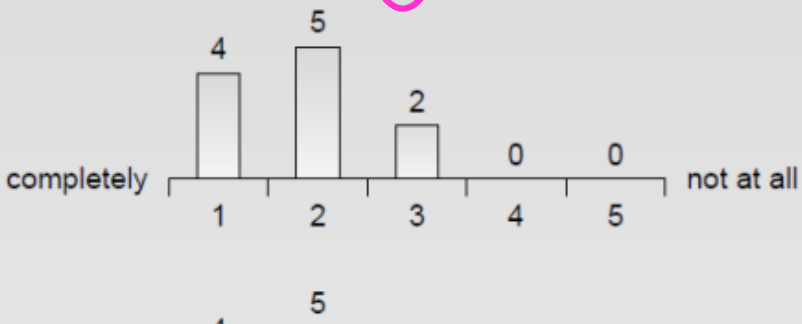
Structure

The learning objective was clear to me.



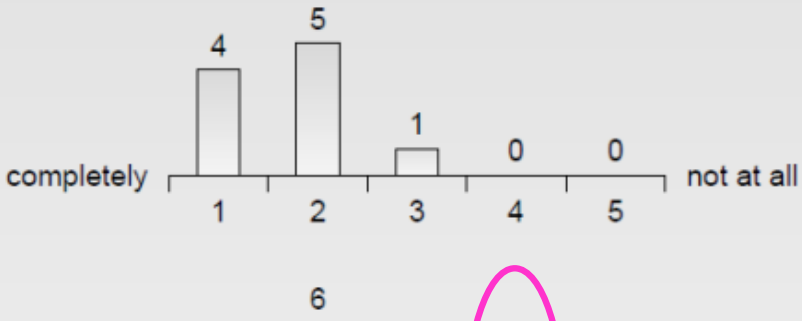
N= 11
M= 1.73
SD= 0.65
k.A.= 0

The educational objectives were well defined from the beginning.



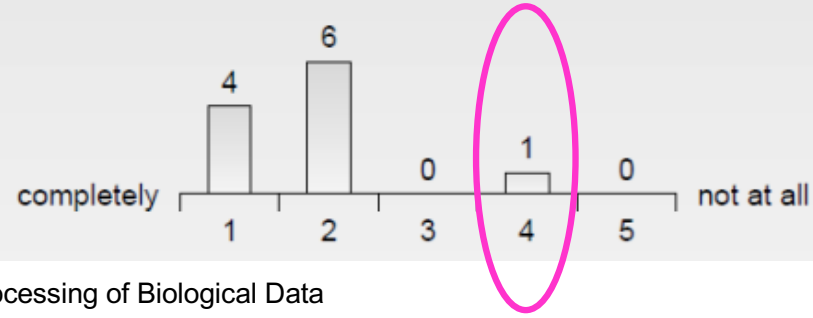
N= 11
M= 1.82
SD= 0.75
k.A.= 0

The course was well structured and comprehensible.



N= 11
M= 1.7
SD= 0.67
k.A.= 1

The structure of the content was logical/easy to follow.



N= 11
M= 1.82
SD= 0.87
k.A.= 0

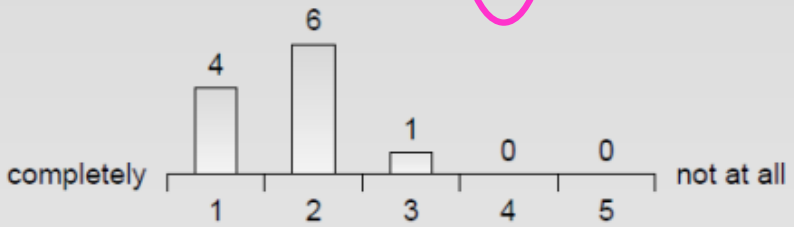
Course evaluation

Topic

I was already interested in the subject of the course before I signed up for it.

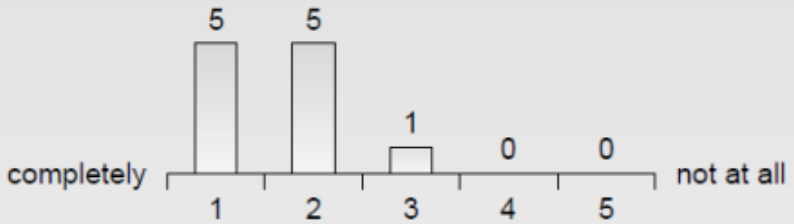
N= 11
M= 1.82
SD= 0.98
k.A.= 0

I believe that I have learned important facts in this course.



N= 11
M= 1.73
SD= 0.65
k.A.= 0

The topic of the course is relevant.

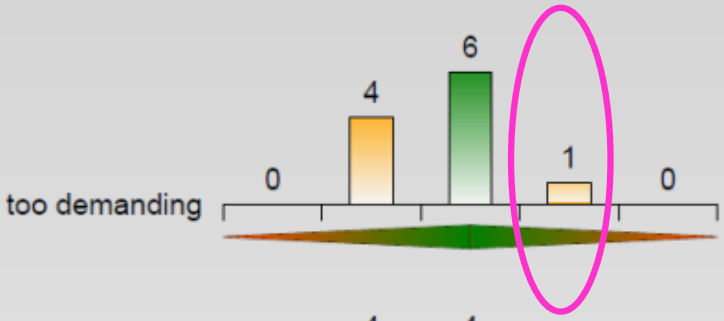


N= 11
M= 1.64
SD= 0.67
k.A.= 0

Course evaluation

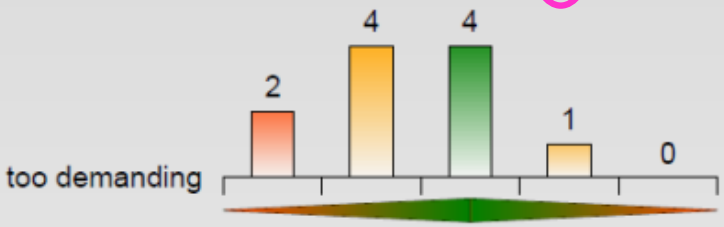
Requirements

The difficulty of the content was…



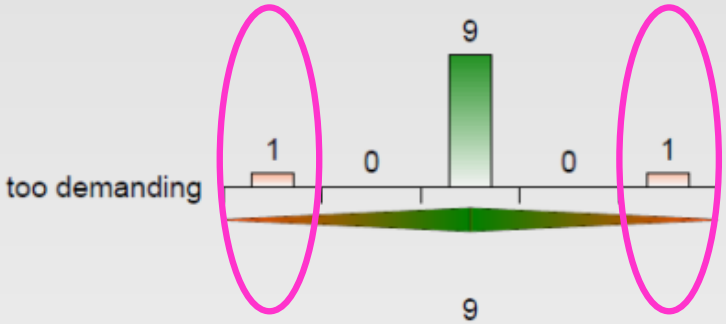
N= 11
M= 2.73
SD= 0.65
k.A.= 0

The amount of the content was…



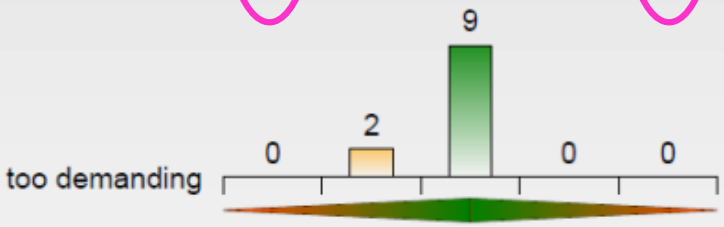
N= 11
M= 2.36
SD= 0.92
k.A.= 0

The requirements of the course were…



N= 11
M= 3
SD= 0.89
k.A.= 0

The amount of time required for the course (including preparation and follow-up) was…

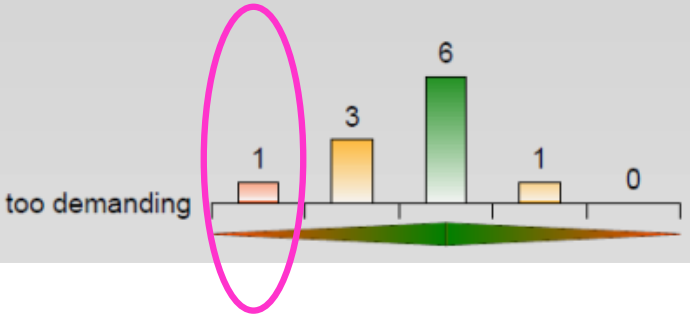


N= 11
M= 2.82
SD= 0.4
k.A.= 0

Course evaluation

Requirements

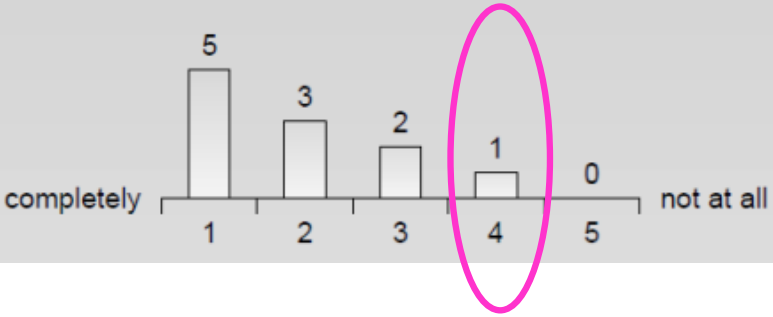
Overall, I felt the course to be…



N= 11
M= 2.64
SD= 0.81
k.A.= 0

Organization

Altogether, the course was well organized.



N= 11
M= 1.91
SD= 1.04
k.A.= 0

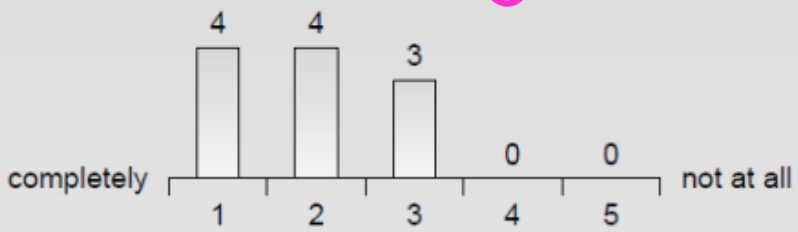
Course evaluation

Organization

Concerning the organizational aspects of the course (i.e. place, time, performance requirements) I was informed well.

N= 11
M= 1.64
SD= 0.92
k.A.= 0

I was satisfied with the accessibility of necessary learning material.

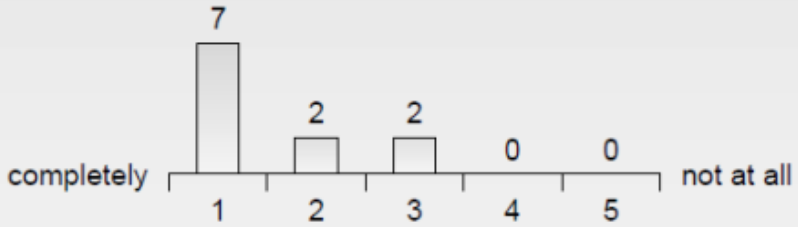


N= 11
M= 1.91
SD= 0.83
k.A.= 0

Organizational issues were dealt with in time and in detail.

N= 11
M= 2.36
SD= 1.36
k.A.= 0

The course was running smoothly during the semester.

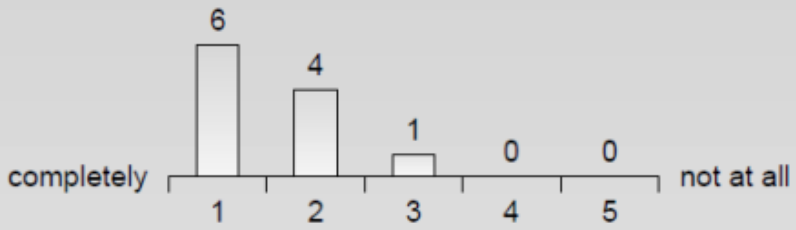


N= 11
M= 1.55
SD= 0.82
k.A.= 0

Course evaluation

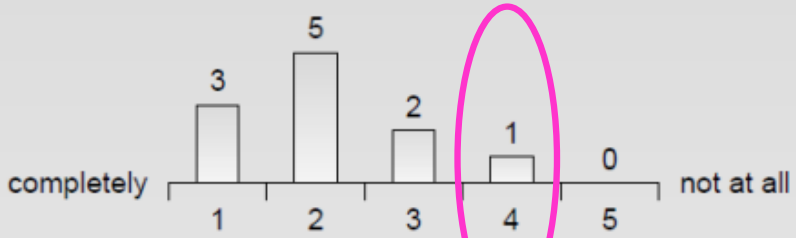
Overall Assessment

Overall, this was a good course.



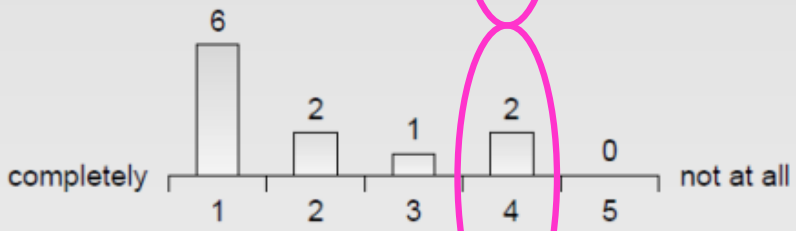
N= 11
M= 1.55
SD= 0.69
k.A.= 0

I learned a lot in this course.



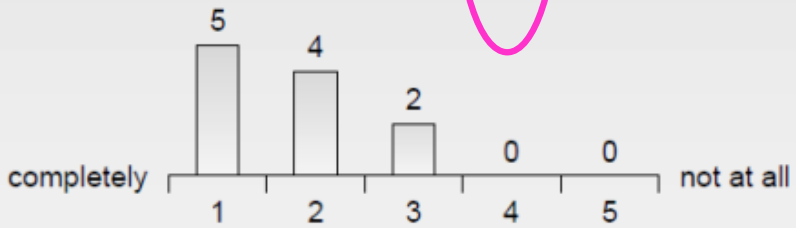
N= 11
M= 2.09
SD= 0.94
k.A.= 0

The course fulfilled my expectations.



N= 11
M= 1.91
SD= 1.22
k.A.= 0

I would recommend the course.

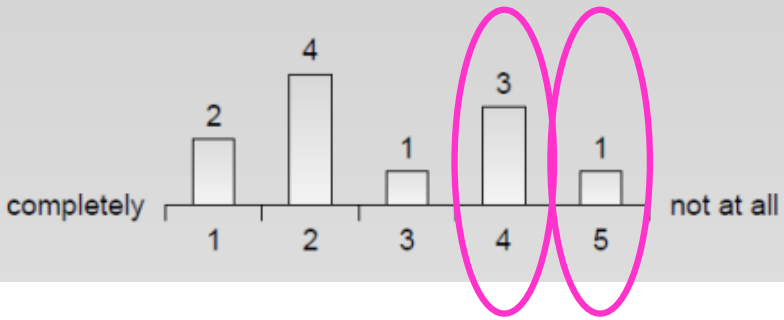


N= 11
M= 1.73
SD= 0.79
k.A.= 0

Course evaluation

Overall Assessment

In terms of its quality, this course was as good as the best course I have ever attended.



N= 11
M= 2.73
SD= 1.35
k.A.= 0

Course evaluation

Further remarks: I especially appreciated

"The varied topics & applicability to real life research or Data Science."

"Diversity of topics."

"The way the prof is communicating."

"Structure of the course."

"The way of explaining the concepts with real-time examples and experiments."

Course evaluation

Further remarks: I did not like

"The tutorial every 2 weeks was [?] confusing. Few assignments --> screwing even one up leads to big consequence. The degree of difficulty was [?] and they clearly cultivated useful skills."

"Theory & Tutorial isn't correlated."

"The way explained."

"The theory part and assignment are not always matching."

"Sometimes the assignments are much more than the content of slides. We need to spend lots of time to search for that and figure out."

Course evaluation

Further remarks: Suggestions for improvements

"Variables of formulary should always be clearly listed on the slides"

"Please try to cover the topic that we cover in tutorial. (As tutorial handles move practice stuff that has been covered at all)"

"Maybe teaching in more simple language with more illustrations."