V12 Multi-omics data integration

Program for today:

- *Staphylococcus aureus* Africa project – analysis for confounding variables

- Overview multivariate analysis for omics projects

- Case study: gene-regulatory network for breast cancer

- Case study: single cell methylation and expression data

Relevant slides for written exam on July 18

Lecture	Slides
1	32-43
2	2-12
3	9-16, 22-31
4	38-46
5	5-11, 22
6	-
7	2
8	1-7, 10-29, 41-44
9	22, 24
10	-
11	7-25
12	1-11, 21-24
Material (algorithms, protocols) from all 5 assignments	

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.

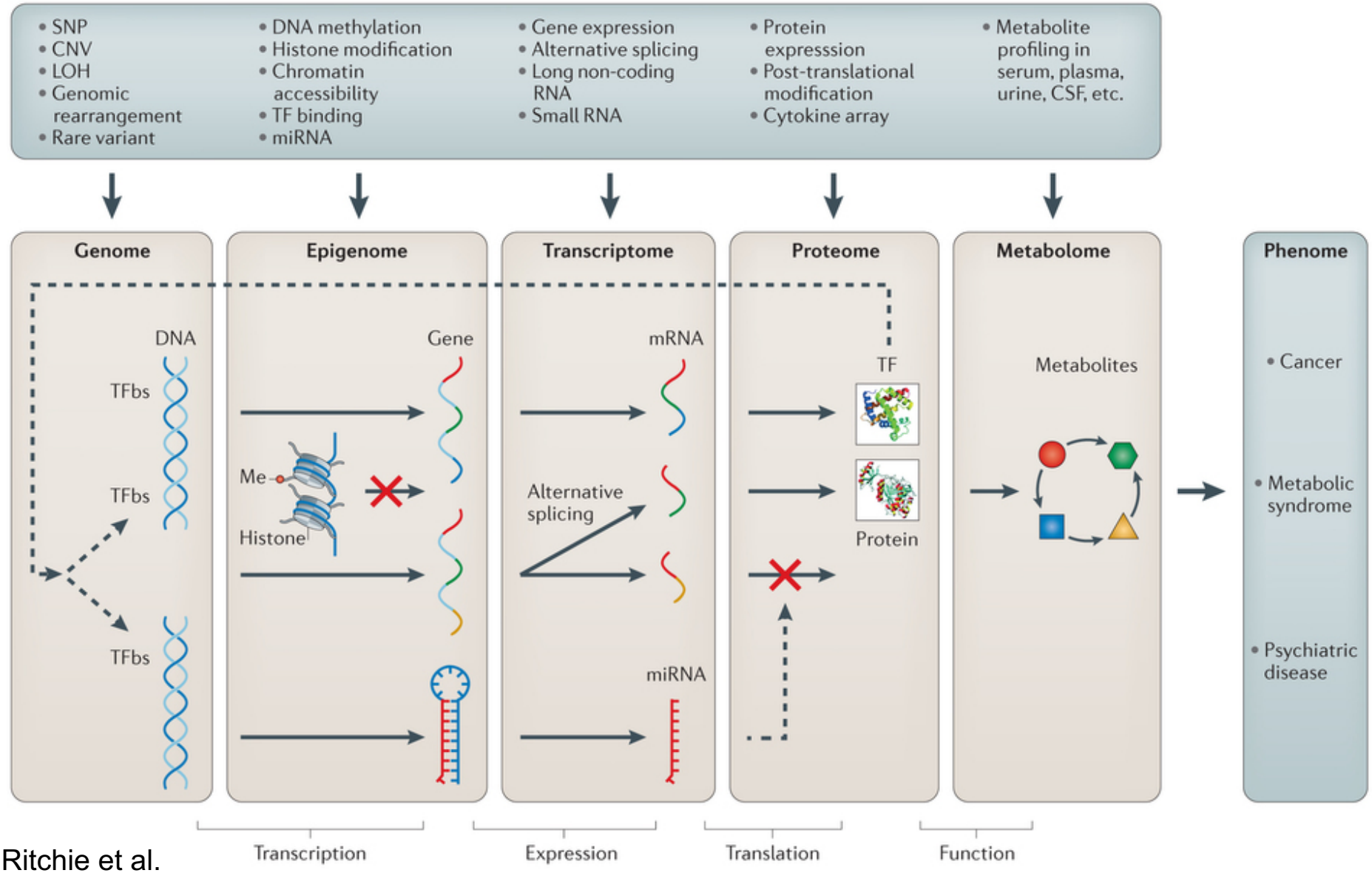
Main motivation behind combining different data sources:

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 V11, 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

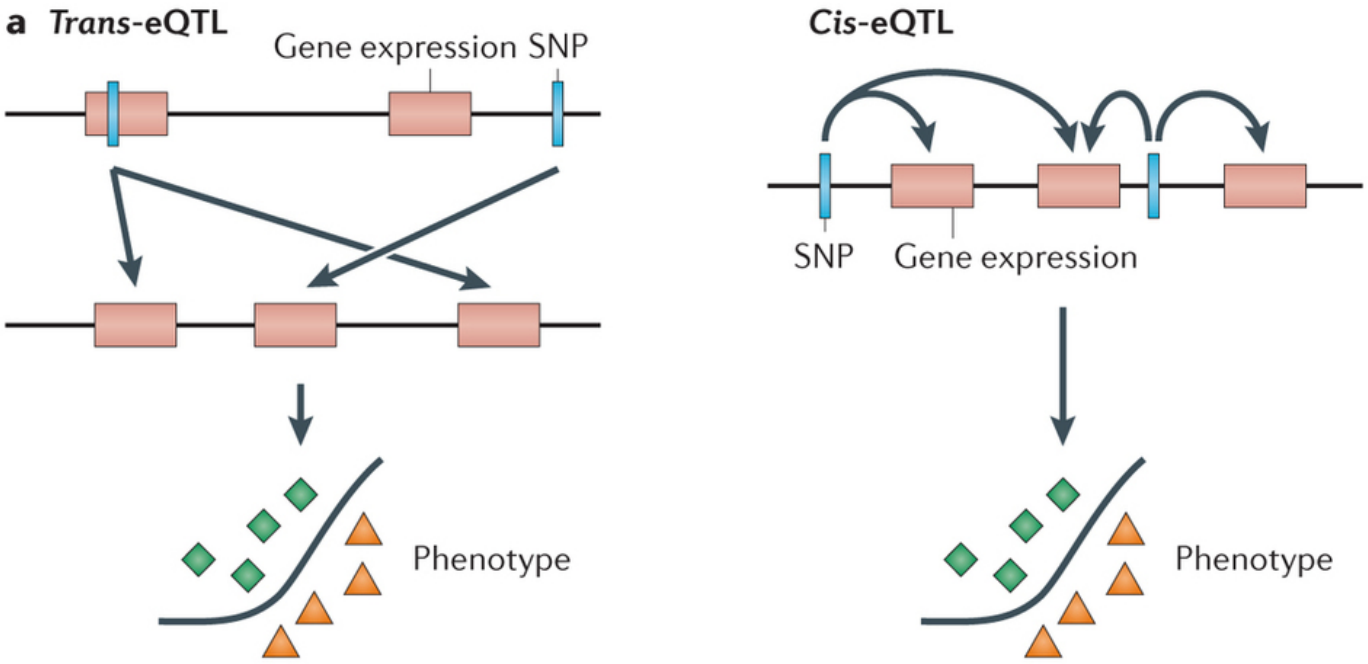
Steps: (1) associate SNPs with phenotype; filter by significance threshold

(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

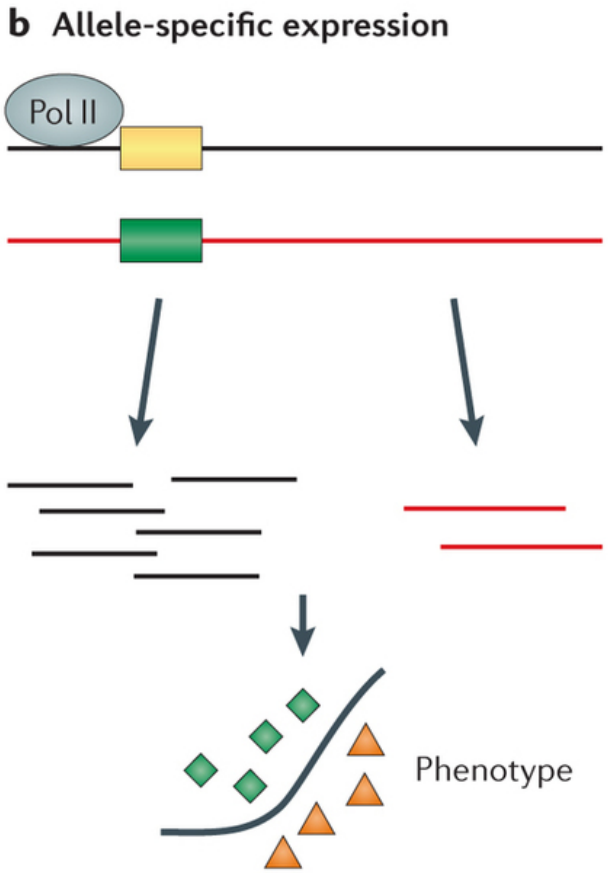
Cis-eQTL: effect on nearby gene



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.

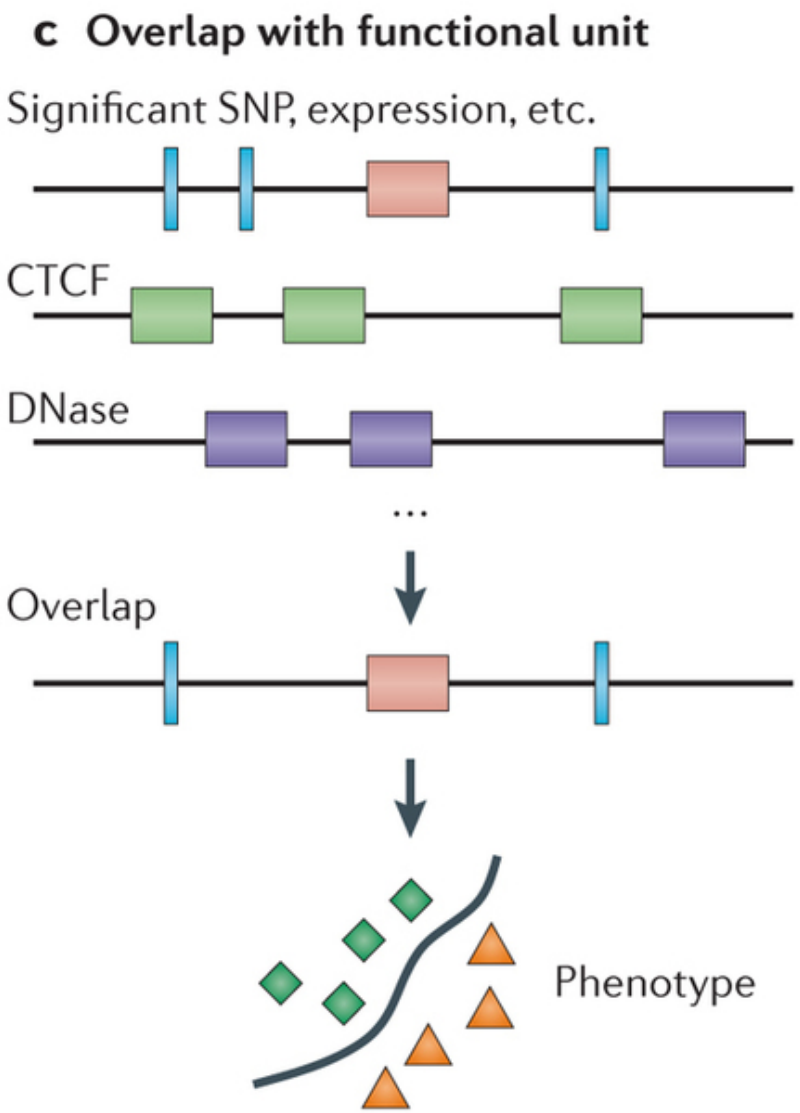
ASE analysis tests whether the maternal or paternal allele is preferentially expressed.

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.

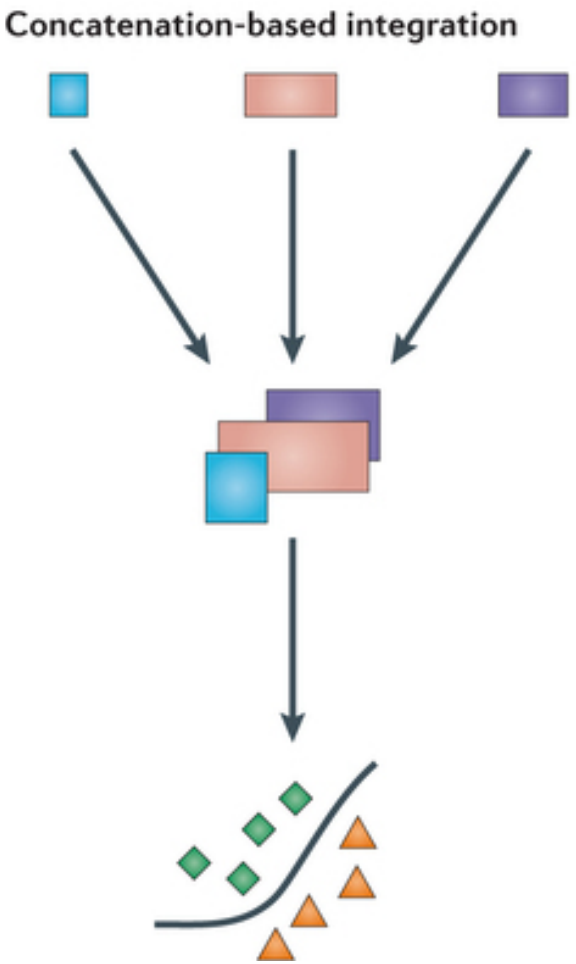
Nature Rev Genet **16**, 85 (2015)

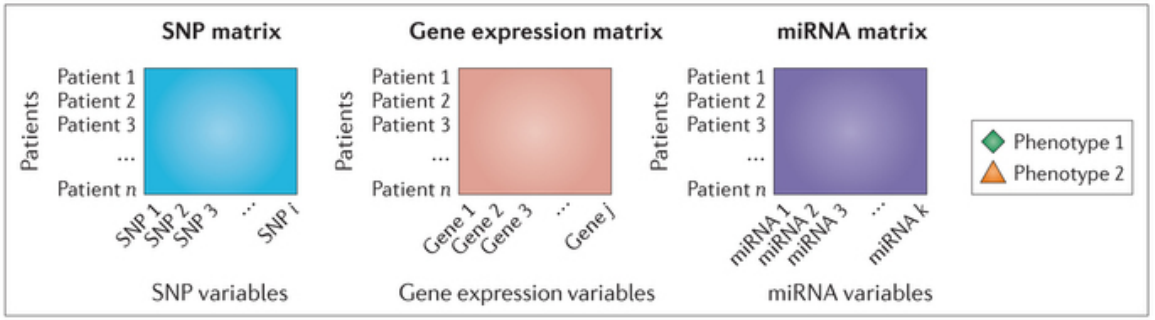
Domain knowledge overlap involves a two-step analysis:
(1) an initial association analysis is performed at the SNP or gene expression variable.

(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.

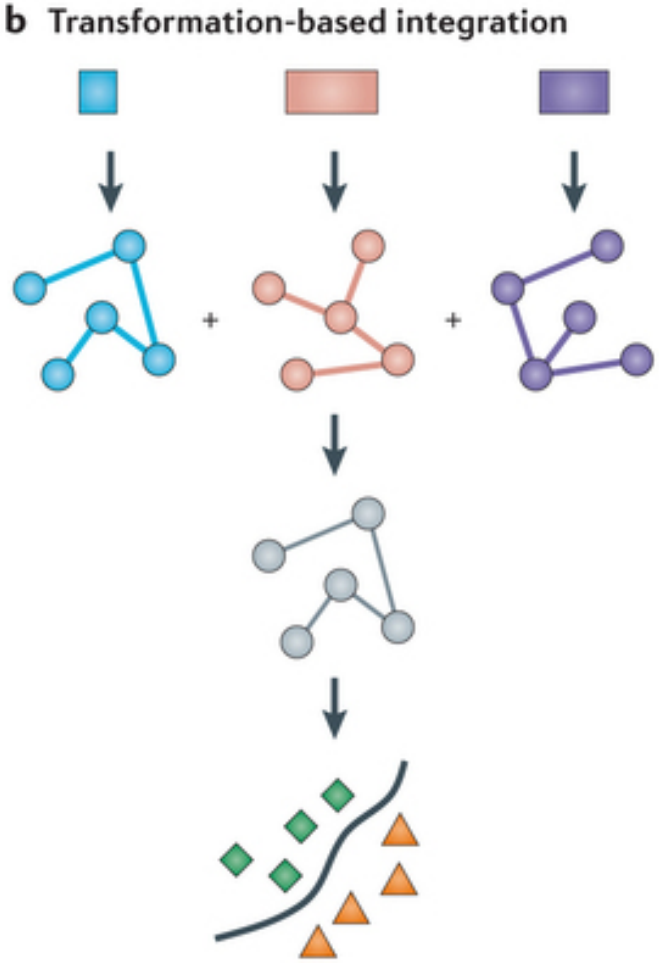
Challenges:

- 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



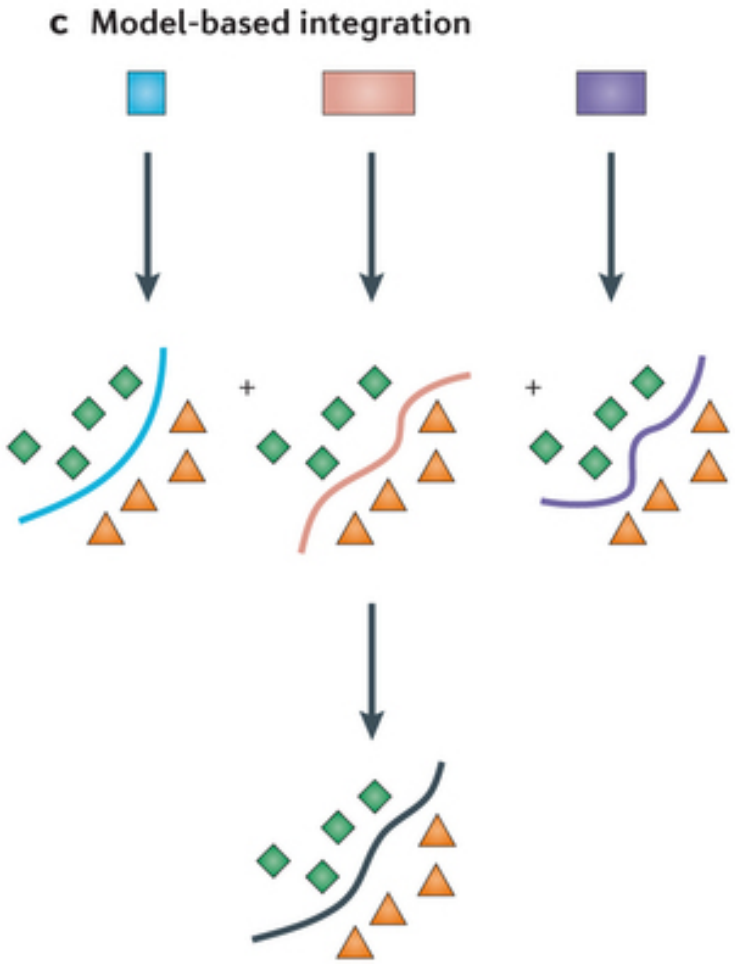
b | Transformation-based integration involves performing mapping or data transformation of the underlying data sets before analysis.

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)

Case study: multi-omics analysis of lung cancer

Lung cancer is the leading cause of death from cancer in the United States resulting in over 150,000 deaths per year.

Non-small cell lung cancer is the predominant form of the disease.

The 5 year survival rate for Non-Small Cell Lung cancer (NSCLC) patients is only about 21%.

Lung cancer treatment is therefore moving rapidly towards an era of personalized medicine, where the molecular characteristics of a patient's tumor will dictate the optimal treatment modalities.

Yan et al. *Scientific Reports* **7**, 333 (2017)

Multi-omics analysis of lung cancer

NSCLC patients with EGFR mutations show significantly improved responses to treatment with Tyrosine kinase inhibitors, e.g., **gefitinib** or **erlotinib** that target this receptor kinase.

However, almost all of these patients eventually **relapse** due to development of **resistance** through various mechanisms.

This paradigm of tumor rewiring in the face of targeted treatment is evident from many clinical trials.

This suggests that the identification of complementary targets will be necessary to improve survival and the probability of long term cures.

Yan et al. *Scientific Reports* **7**, 333 (2017)

New targets?

The ubiqutin protease system (UPS) regulates a variety of basic cellular pathways associated with cancer development. Aberrations in the UPS are implicated in the pathogenesis of various human malignancies.

The cullin-RING ligases (CRLs) are the largest E3 ligase family involved in the UPS. In mammals, there are 7 different cullins (e.g. CUL4) and 2 rings (RBX1 and 2).

Recently, many DDB1-CUL4 associated factors (DCAFs) have been identified and serve as substrate receptors to execute the degradation of proteins.

However, the specific CUL4A-DCAF nexus that contributes to human cancer development remains largely unknown.

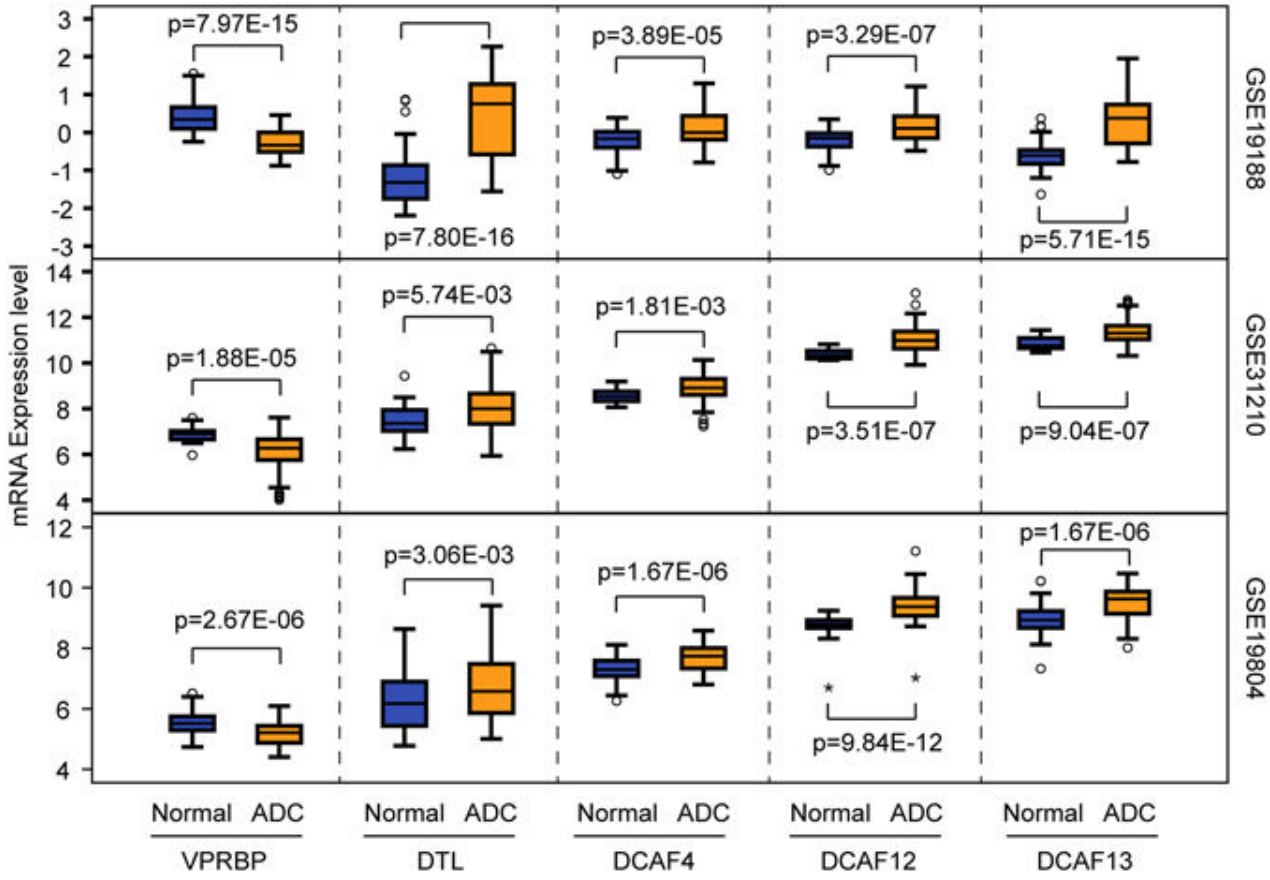
This study analyzed whether **combining data** on gene expression and DNA copy number variations for 19 well-defined DCAFs in human lung adenocarcinomas (LuADCs) could be helpful for predicting patient **prognosis**.

Yan et al. *Scientific Reports* **7**, 333 (2017)

Expression of DCAFs

Out of 19 DCAFs, *VPRBP* gene expression was significantly **downregulated** in lung adenocarcinoma compared to normal lung tissue.

On the other hand, *DTL*, *DCAF4*, *DCAF12* and *DCAF13* were significantly **upregulated** in LuADCs.



Increased or decreased gene expression cut-off: 2-fold and adjusted p < 0.05.

lung adenocarcinoma (orange).

normal lung tissue (blue)

3 different GEO data sets.

Yan et al. *Scientific Reports* **7**, 333 (2017)

DNA copy number variations

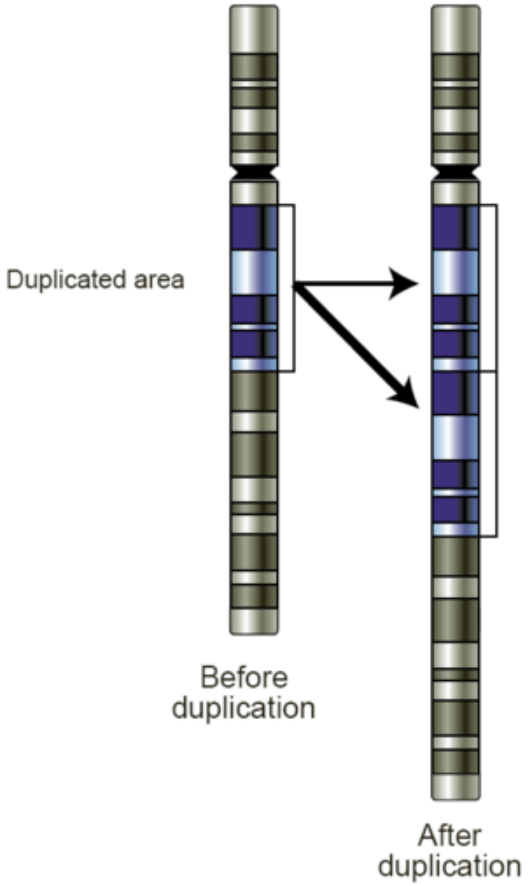
Copy number variation (**CNV**) is a phenomenon in which sections of the genome are repeated and the number of repeats in the genome varies between individuals in the human population.

CNVs are a type of structural variation: a considerable number of base pairs are duplicated or deleted

CNVs occur in humans and in a variety of other organisms including *E. coli*.

Approximately two thirds of the entire human genome is composed of repeats.

4.8-9.5% of the human genome can be classified as CNVs.



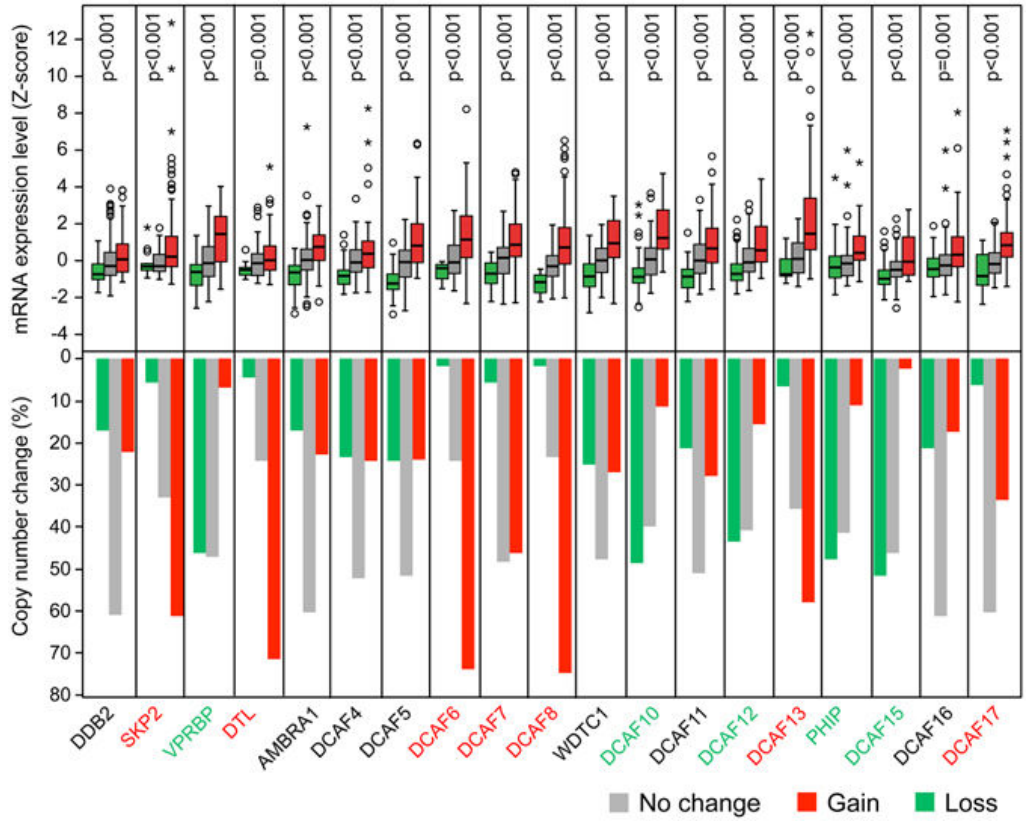
www.wikipedia.org

Correlated expression with copy number variations

DCAF gene expression is strongly correlated with its DNA copy number in lung adenocarcinoma.

Top panel: relationship between tumor DNA copy number and gene expression for 19 DCAFs in LuADCs.

Bottom panel: percent of tumors with DNA copy number change.



Tumors with increased DNA copy number (Gain) are indicated in red and those with decreased DNA copy number (Loss) in green. Tumors with no change in DNA copy number are indicated in gray.

Yan et al. *Scientific Reports* **7**, 333 (2017)

Pronognistic value of DCAF expression?

Evaluated their prognostic value of DCAFs for LuADC patients in a large public clinical microarray database using Kaplan-Meier plots.

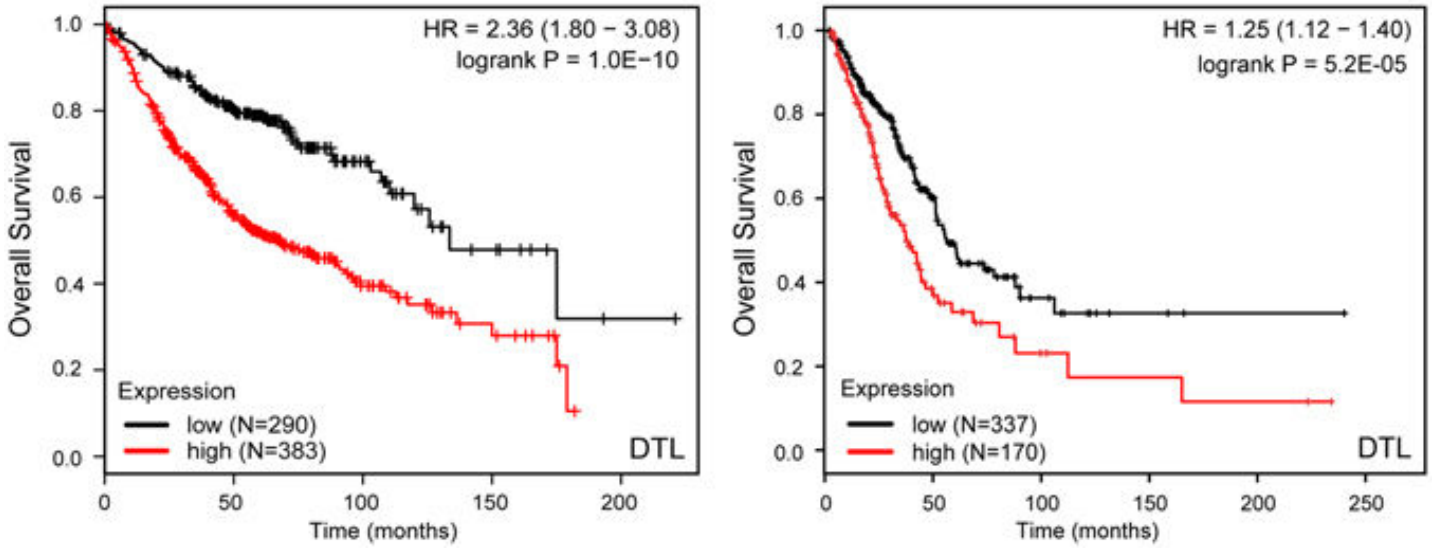
Patients were divided into 2 groups based on the expression levels of each individual DCAF.
Subsequently, the effect of high or low expression level of these DCAFs on the overall survival (OS) was evaluated using Cox regression analysis, the Kaplan-Meier survival curve and log-rank test.

→ high transcriptional levels of *DTL* and *DCAF15* are significantly associated with shortened overall survival (OS) whereas
high transcriptional levels of *DDB2*, *DCAF4* and *DCAF12* favor good prognosis.

Re-analysis with TCGA data only confirmed *DTL*.

www.wikipedia.org

DTL gene expression is associated with survival in LuADCs

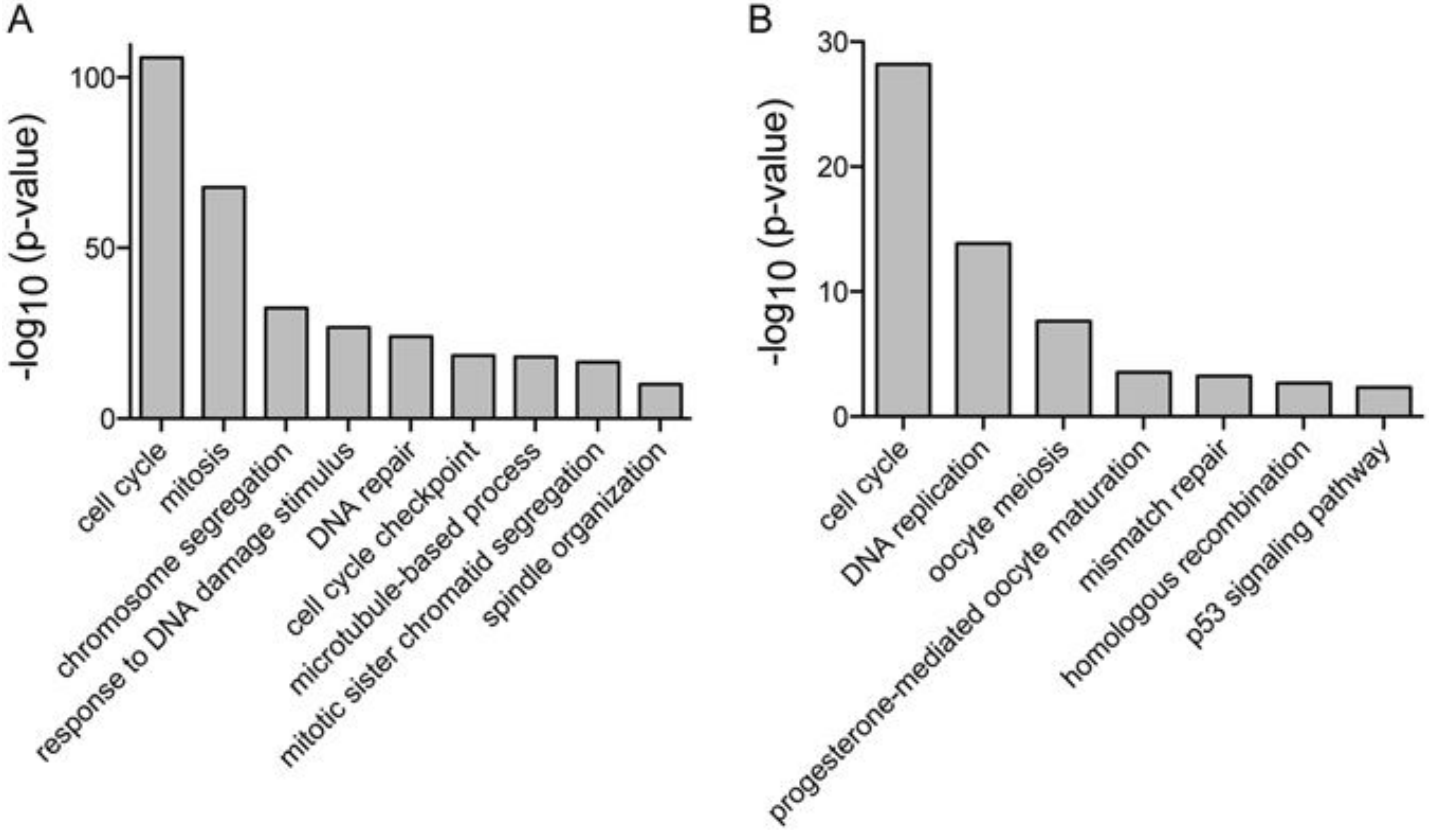


Survival risk curves are shown for *DTL* expression in LuADCs using KM-plotter (probe ID 222680_s_at) (**A**) and TCGA (**B**).

Low and high expression level of *DTL* are drawn in black and red respectively.
In TCGA, patients were divided into tertiles based on DTL expression levels. The top tertile was defined as the high DTL expression cohort and the remaining patients were defined as the low DTL expression cohort.
The p-value represents the equality of survival curves based on a log-rank test.

Yan et al. *Scientific Reports* **7**, 333 (2017)

GO annotations: potential role of DTL



Gene Ontology (**A**) and KEGG pathway (**B**) analysis of genes that are transcriptionally co-expressed with DTL in LuADCs.

DTL-correlated genes are enriched in cell cycle and DNA repair pathways.

The levels of 25 proteins were significantly associated with DTL overexpression in LuADCs, which include significant decreases in protein level of the tumor supressor genes such as PDCD4, NKX2-1 and PRKAA1.

Yan et al. *Scientific Reports* **7**, 333 (2017)

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

Rethink: which data to be measured / analyzed?

- Case study: CNVs strongly correlated with expression data

-> don‘t provide more information on the consequences,
only on the possible reason why expression changes

- Some epigenetic marks are also correlated with gene expression
-> don‘t provide more information

- Most useful: complementary data e.g. on protein activity (phosphorylation)

Thoughts on improving this lecture

- Topics of the lecture

- Topics of the assignment