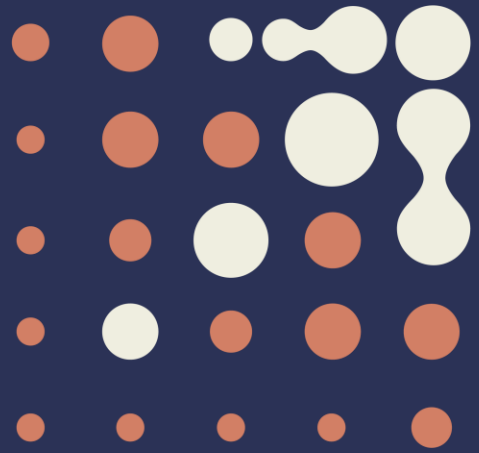
Medical Genomics: Transcriptomics, multi-omics, and beyond

N. Alcala

Rare Cancers Genomics Team

November 10th 2021



International Agency for Research on Cancer

World Health Organization

RARE CANCERS GENOMICS

Plan

Part I. Transcriptomics

- **Concepts**: tissue heterogeneity and microenvironment
- **Techniques**: sequencing strategies (bulk, single-cell, spatial)
- **Resources**: medical transcriptomics databases
- **Analysis**: calling somatic variants, supervised and unsupervised analyses

Part II. Multi-omics

- **Concepts**: complementarity of 'omic layers
- **Analysis**: tools for integration

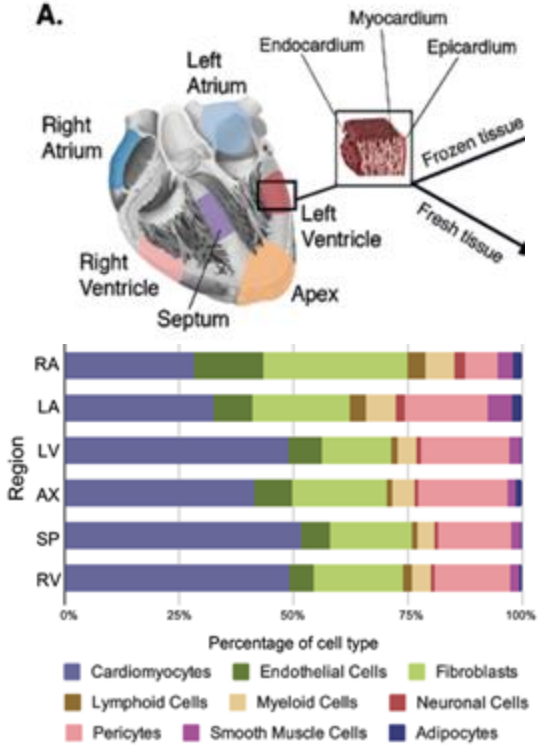
Part III. Integration with other types of medical data

- **Concepts**: medical imaging and digital pathology
- **Analysis**: deep learning and integration with whole-slide pathological images

Part I. Transcriptomics | *Concepts*

Tissue heterogeneity

- Tissues are made of **mixtures of cells**
- Although not novel, **the investigation of tissue heterogeneity** gained novel traction with new sequencing technologies
- For instance, even the most studied organs like the heart are still under fundamental research investigations



Adult human heart cell composition. Percentage of cell types estimated from single-cell RNA-seq. *Source: Litviňuková et al. Biorxiv 2020.*

Part I. Transcriptomics | *Concepts*

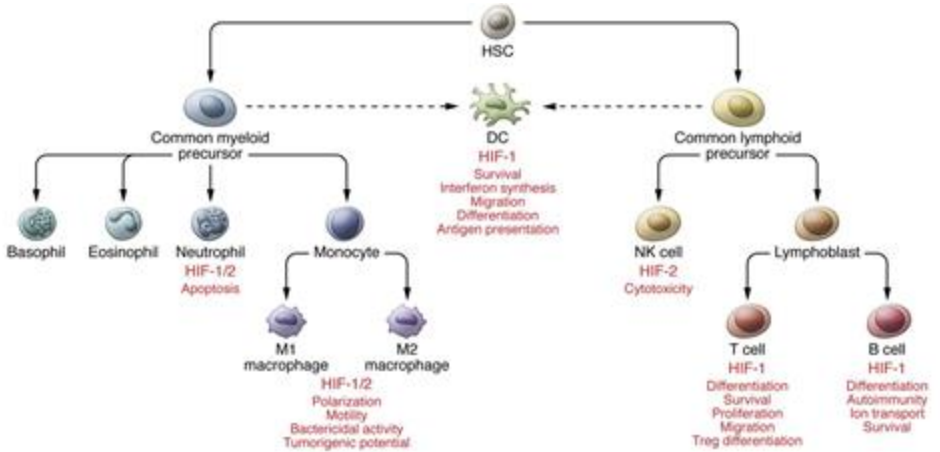
Tissue heterogeneity: Stroma and Microenvironment

Stromal cells (connective tissue cells)

- **Fibroblasts**: synthesize the extracellular matrix and collagen, initiate inflammation and immune response

Immune cells

- **Dendritic cells**: present antigens
- **Macrophages**: perform phagocytosis
- **T cells**: cytotoxic (CD8+), helper (CD4+)
- **Neutrophils**: promote inflammation, phagocytosis

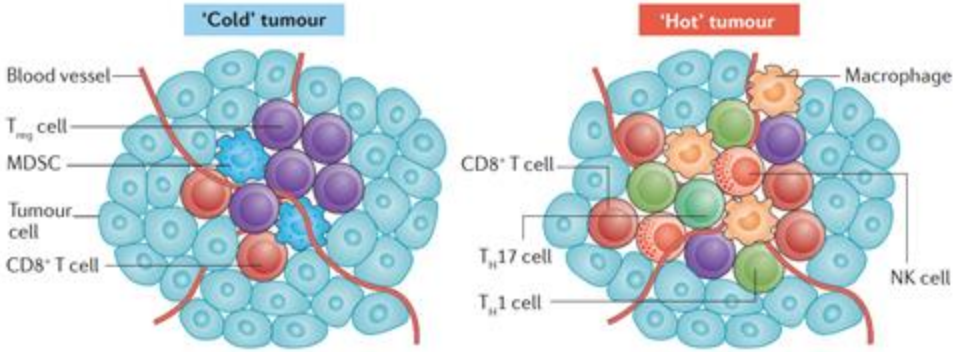


Immune cell differentiation. *Source: Taylor et al. J. Clin Invest 2016.*

Part I. Transcriptomics | *Concepts*

Tissue heterogeneity: Tumor microenvironment (TME)

- Tumors have various amounts and compositions of **Tumor Infiltrating Lymphocytes (TILs)**
- **TILs influence disease progression**
- **Cold tumors** have few TILs, and mostly from lowly differentiated cells
- **Hot tumors** have many TILs and effector immune cells with antitumor functions (NK, CD8+ T cells)

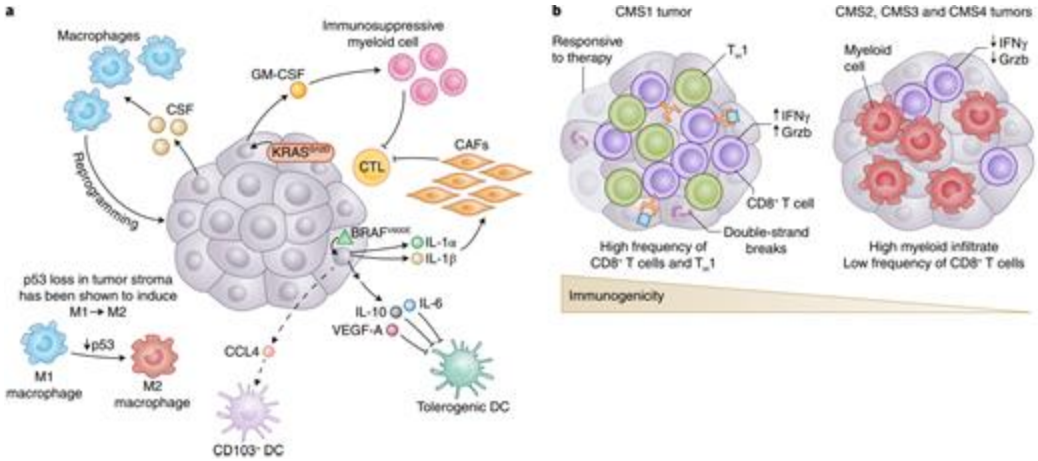


Tumors differ in their level of infiltration. *Source: Nagarsheth et al. Nat Rev Immun 2017.*

Part I. Transcriptomics | *Concepts*

Tissue heterogeneity: Tumor microenvironment (TME)

- **Tumors can establish protumoral and immunosuppressive environments**
- They **recruit stromal and immune cells** to suppress the immune response (e.g., cancer associated fibroblasts), promote metastasis (e.g. using tumor-associated macrophages) by increasing angiogenesis (blood vessel formation providing nutrients to the tumor)

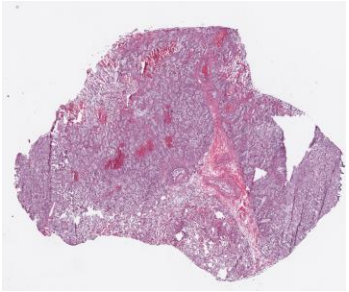


Tumors genotypes and phenotypes shape the TME. **a**. For example, in melanoma KRAS somatic alterations promote the recruitment of immunosuppressive cells. **b**. DIfferent subtypes of colorectal carcinoma (CMS1-4) present different levels genomic instability and TILs (hot vs cold tumors) that influence response to immunotherapy. *Source: Binnewies et al. Nature Medicine 2018*.

Part I. Transcriptomics | *Techniques*

Bulk sequencing: preparation

1. **Tissue collection**: Surgical resection of the tissue
2. **Medical diagnosis** (pathological review): Formalin-Fixed Paraffin-Embedded (FFPE) block preserving the tissue structure but damaging DNA and RNA, and stained with Hematoxylin and Eosin (H&E) to allow microscopic examination
3. **RNA extraction**: biopsy dipped in liquid nitrogen and stored at -80°C to create a Fresh Frozen sample, preserving DNA and RNA but difficult to read for diagnosis
4. **Library preparation**: mRNA purified and fragmented, reverse transcription, complementary DNA (cDNA) synthesis, end repair and A-tailing, adapter ligation, purification and amplification (PCR) to create the final cDNA libraries
5. **Sequencing**



Lung Adenocarcinoma. Source: cancer digital slide archive

Part I. Transcriptomics | *Techniques*

Bulk sequencing: preparation

Notes:

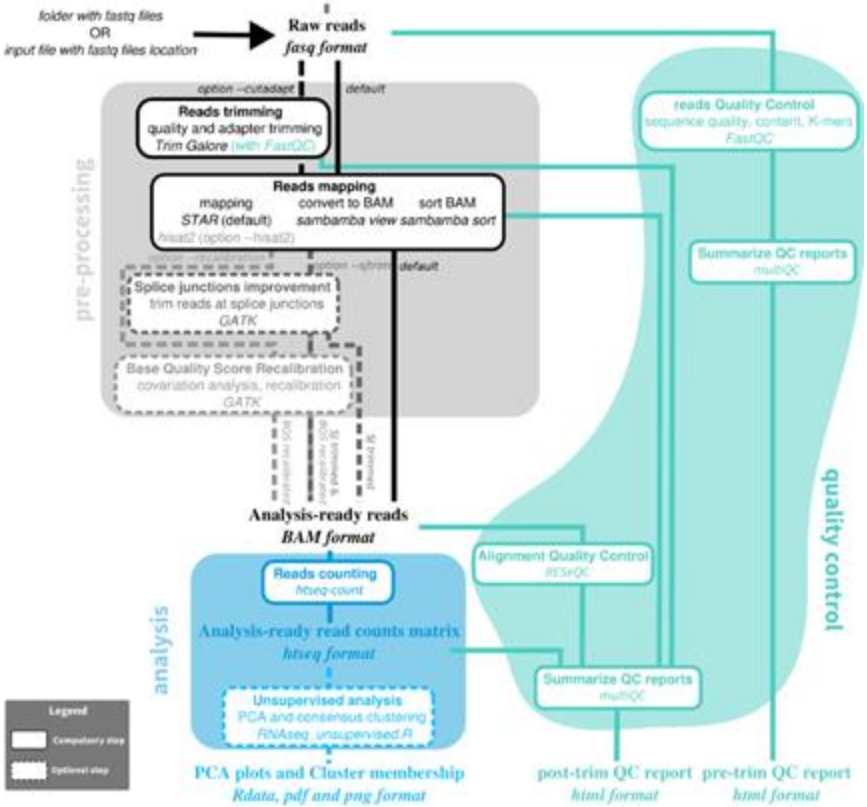
- *For cancer transcriptomics, adjacent **normal tissue is often not sequenced**, because of difficulty to ensure that a tissue has actually entirely normal transcriptome, thus **transcriptomics often study variation within diseased tissue and not the difference between normal and diseased tissue***

- *Steps (3)-(5) require **patient consent for molecular analyses and collection of de-identified data**, reviewed by an ethics committee.*

Part I. Transcriptomics | *Techniques*

Bulk sequencing: processing

- Quality control
- Mapping (e.g., STAR) or pseudoalignment (kallisto, salmon)
- *Optional: preparation for calling (splice junction trimming, base quality score recalibration)*
- *Optional: local realignment to improve splice junction and indel identification*
- Quantification at gene and transcript level



Part I. Transcriptomics | *Techniques*

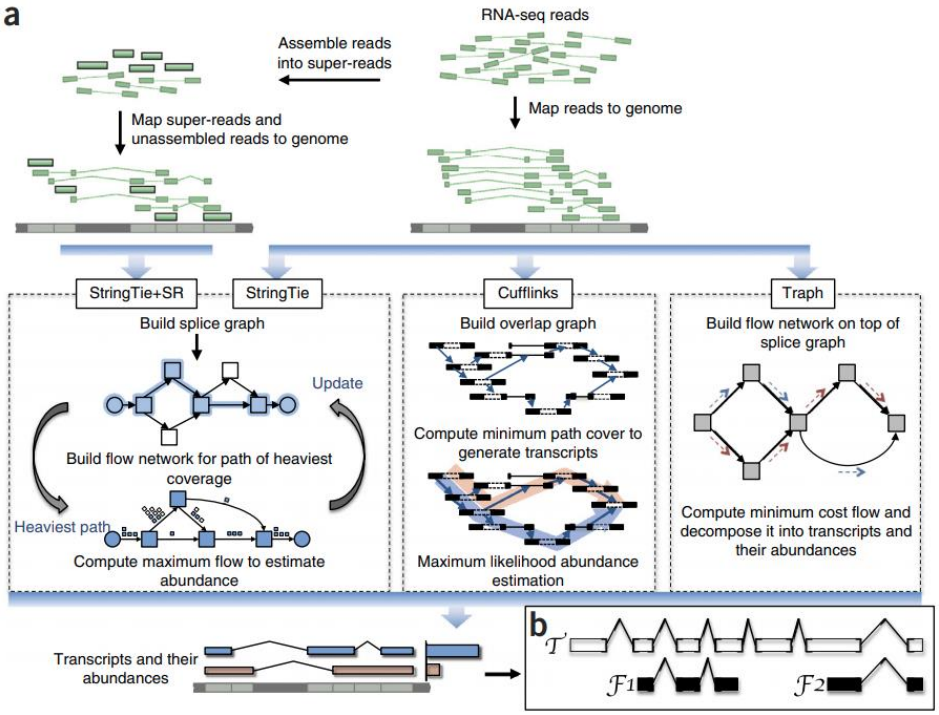
Bulk sequencing: processing

- Quantification at gene and transcript level - for **genes**, count the number of reads aligned to exon

Part I. Transcriptomics | *Techniques*

Bulk sequencing: processing

- Quantification at gene and transcript level
  - for genes, count the number of reads aligned to exon
  - for **transcripts**, more complicated, need approximate algorithm to guess to which transcripts the read belongs
- **Final result: table with for each gene (resp. transcript) and each sample, the amount of expression**

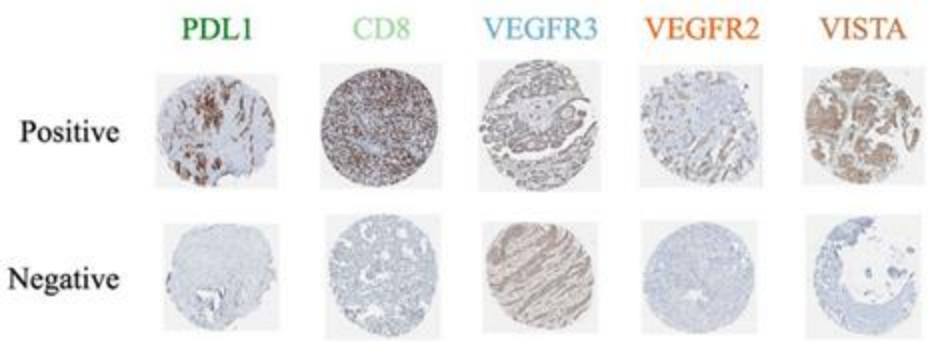


Part I. Transcriptomics | *Techniques*

Bulk sequencing: confirming the results

Because of these uncertainties, confirming the results is necessary

- ***Validation** using the same cohort: duplicates with the same or another technique (e.g., immunohistochemistry to quantify protein expression)*



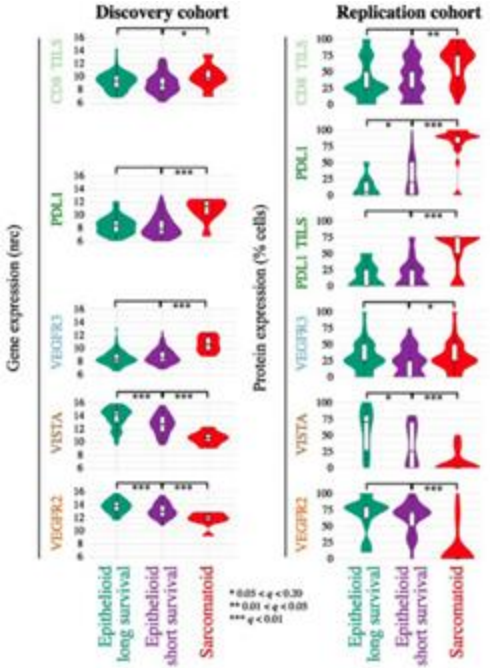
Technical validation of a five-gene panel on 103 malignant pleural mesothelioma. Tissue MicroArray (TMA) IHC staining representing the positive and negative references of the tested protein expression. *Source: Alcala et al. Ebiomedicine 2019.*

Part I. Transcriptomics | *Techniques*

Bulk sequencing: confirming the results

Because of these uncertainties, confirming the results is necessary

- ***Validation*** *using the same cohort: duplicates with the same or another technique (e.g., immunohistochemistry to quantify protein expression)*
- ***Replication*** *of main results using another cohort*



Replication of the prognostic value of a five-gene panel for malignant pleural mesothelioma. **Left.** Gene expression levels in the discovery cohort (*n*=113). **Right**. Protein expression levels in the replication cohort, for the three sets (*n*=74)
Source: Alcala et al. Ebiomedicine 2019.

Part I. Transcriptomics | *Techniques*

Single-cell sequencing: principle

Goal: Quantify the level of expression of genes and transcripts of each individual cell of a tissue

- Track cell differentiation
- Quantify tissue heterogeneity
- Quantify diversity of microbiome

Different methods

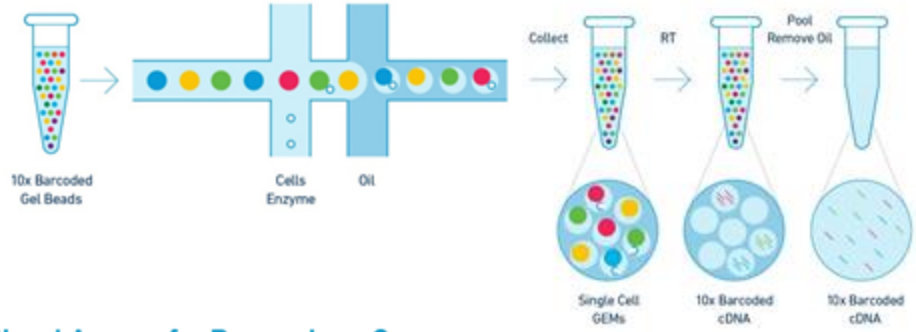
- **Droplet based (10X genomics) -> most used technique**
- Plate-based with unique molecular identifiers (UMIs): CEL-seq, MARS-seq
- Plate-based with reads: Smart-seq2
- Other: sci-RNA-seq, Seq-Well

Part I. Transcriptomics | *Techniques*

Single-cell sequencing: principle

Droplet based (10X genomics)

- Barcoded Gel Beads are attached to each cell to form Gel Bead in EMulsion (GEMs) of nL size
- Reverse transcription (to obtain cDNA) is performed in each GEM, attaching identifiers to each cell
- Amplification, library generation, sequencing are performed as in classical RNA-seq experiment
- Each read is paired with a barcode read with cell identifier + Unique Molecular Identifier (UMI)



Source: 10X genomics.

Part I. Transcriptomics | *Techniques*

Single-cell sequencing: principle

Goal: Quantify the level of expression of genes and transcripts of each individual cell of a tissue

Depth-cell number trade-off:

- Current technologies can sequence 100 to 100,000 cells, with 1,000 to 100,000 reads/cell
- **More cells** help **identify rare cell subtypes**
- **More depth/cell** allows to **identify subtle differences in expression** between cell types

Part I. Transcriptomics | *Techniques*

Single-cell sequencing: processing

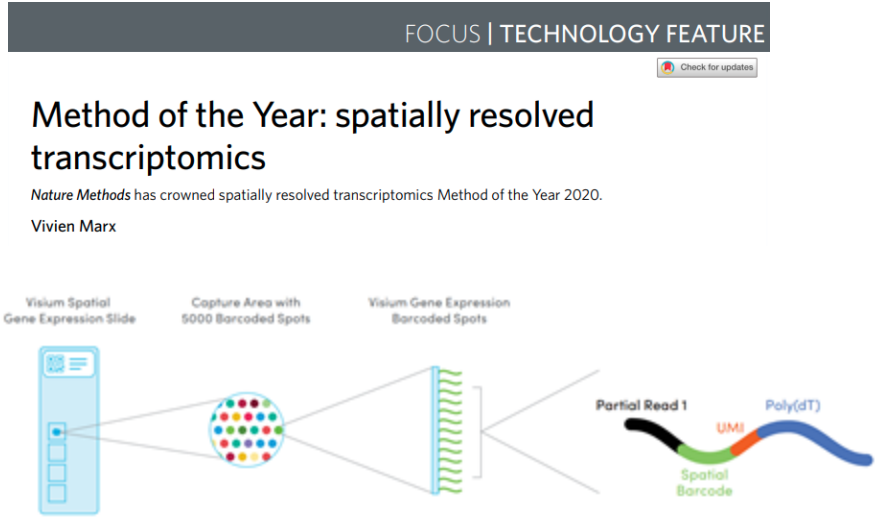
Processing is similar to bulk RNA-seq but taking into account barcodes

- barcode-aware alignment (e.g., STARsolo)
  - error-correction and demultiplexing of cell barcodes
  - standard alignment (e.g., STAR)
  - deduplication of UMIs
- Quantification of UMIs

Part I. Transcriptomics | *Techniques*

Spatial transcriptomics: principle

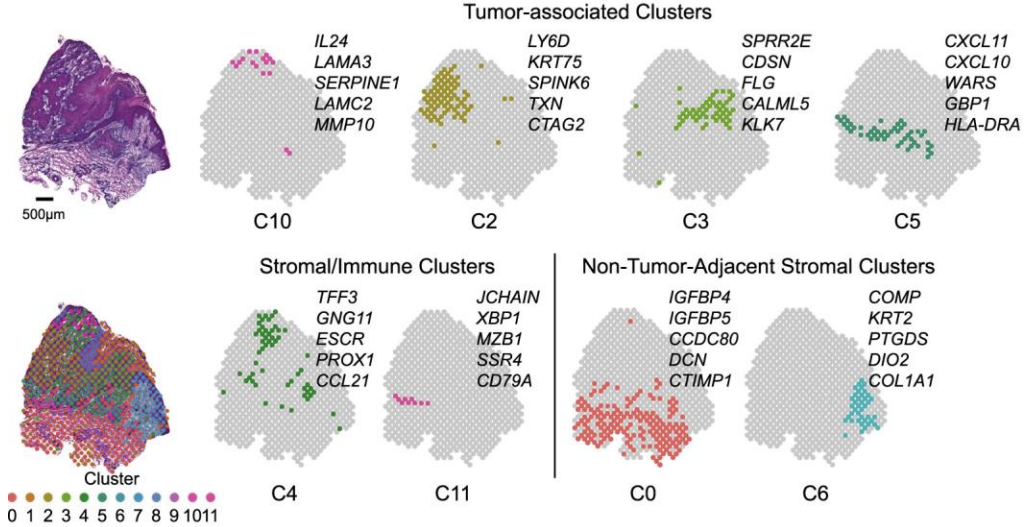
1. Fresh-frozen tissue section placed on array with capture probes that bind to RNA
2. cDNA is synthesized from captured RNA and sequencing libraries prepared
3. libraries are sequenced



Spatial composition of the Visium Spatial Gene Expression slide. Each slide contains four Capture Areas with approximately 5000 barcoded spots, which in turn contain millions of spatially-barcoded capture oligonucleotides. Tissue mRNA is released and binds to the barcoded oligos, enabling capture of gene expression information. *Source: 10X genomics*.

Part I. Transcriptomics | *Techniques*

Spatial transcriptomics: application



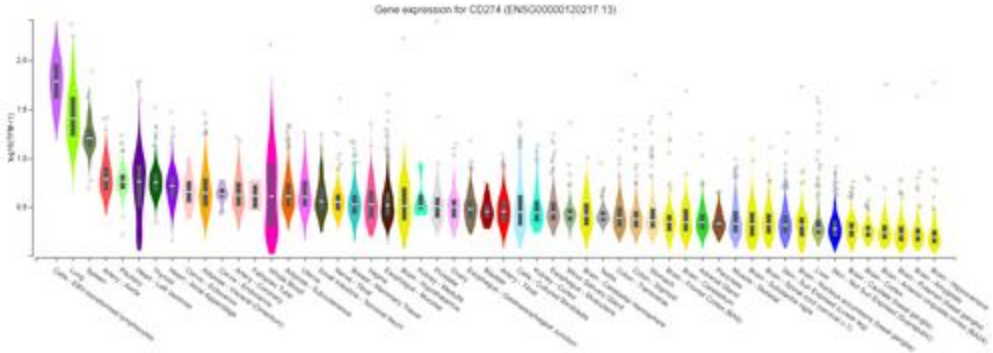
Spatial transcriptomics of a cutaneous squamous cell carcinoma tumor. Top left: H&E slide; bottom left: expression clusters. *Source: Ji et al. Cell 2020.*

Part I. Transcriptomics | *Resources*

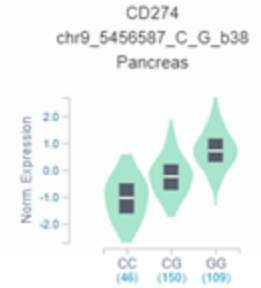
The Genotype-Tissue Expression (GTEx) project

Database of tissue-specific gene expression and regulation

- 54 non-diseased tissue sites for 1000 individuals with WGS/WES, and RNA-Seq
- gene expression, expression quantitative tra loci (eQTL), and histology images



Expression of immune checkpoint gene PD-L1 in 52 tissues.



Example Pancretic eQTL. Gene expression varies as a function of genotype at locus chr9 5456587

Part I. Transcriptomics | *Resources*

The Cancer Genome Atlas (TCGA) project

Database of cancer multi-omic data for

- Tumors from 33 primary sites
- RNA-seq data under controlled access (requires research institute affiliation)
- Processed gene expression data (read counts and FPKM) open-access

Web interface of the genomic data portal hosting the TCGA data. *Source:* https://portal.gdc.cancer.gov/.

Part I. Transcriptomics | *Resources*

The Cancer Genome Atlas (TCGA) project

Database of cancer multi-omic data for

- Tumors from 33 primary sites
- RNA-seq data under controlled access (requires research institute affiliation)
- Processed gene expression data (read counts and FPKM) open-access
- Data can be visualized and basic analyses can be performed

Example analysis that can be performed on the GDC portal. Survival analysis in pancreatic cancer based on *KRAS* mutational status. *Source:* https://portal.gdc.cancer.gov/.

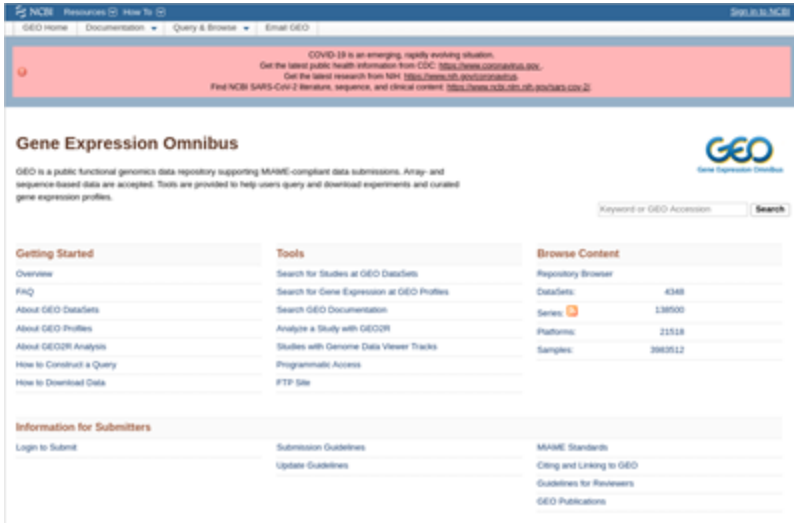
Part I. Transcriptomics | *Resources*

The Gene Expression Omnibus (GEO) repository

Database of expression data (arrays and RNA-seq)

- Includes human data
- All data is open-access

Will be used for the practicals.



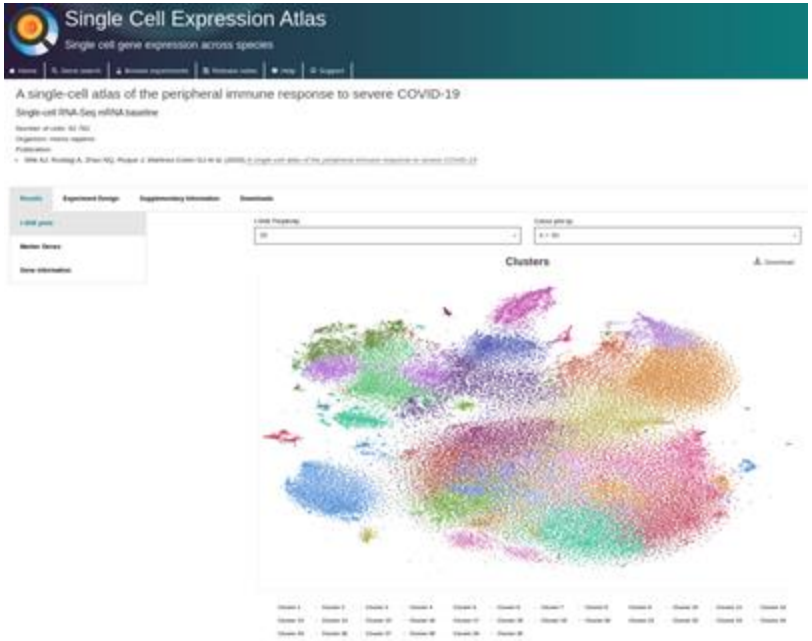
Web interface of the gene expression omnibus repository. *Source:* https://www.ncbi.nlm.nih.gov/geo/.

Part I. Transcriptomics | *Resources*

The Single Cell Expression Atlas

Database of scRNA-seq data

- Processed gene expression data (read counts) open-access



Web interface of the single-cell expression atlas. scRNA-seq of immune response to severe COVID-19 (t-SNE). *Source:* https://www.ebi.ac.uk/gxa/sc/home.

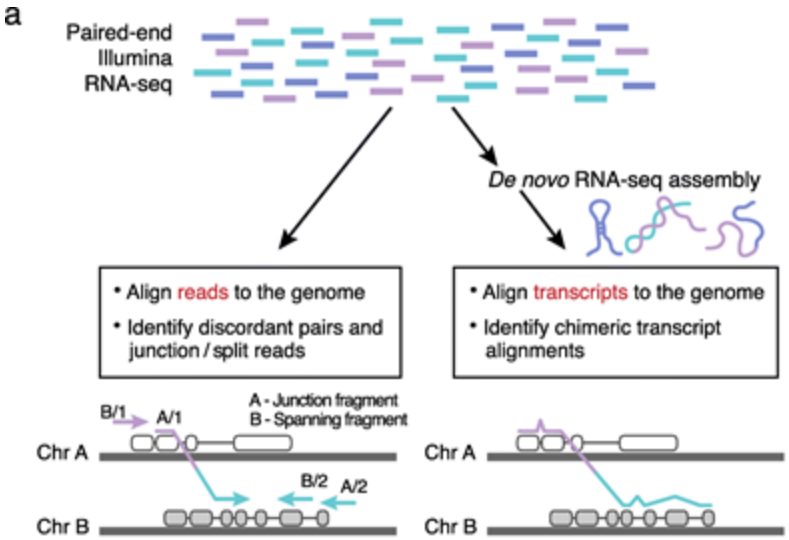
Part I. Transcriptomics | *Analysis*

Variant discovery: gene fusion identification

Goal: discover chimeric genes formed of 2 other genes

Medical relevance: many cancers are driven by oncogenic fusion genes

Methods: Using splice junctions identified during mapping (discordant read-pairs or split reads), identify



Schematic of gene-fusion identification workflow. *Source: Haas et al. Genome Biology 2019.*

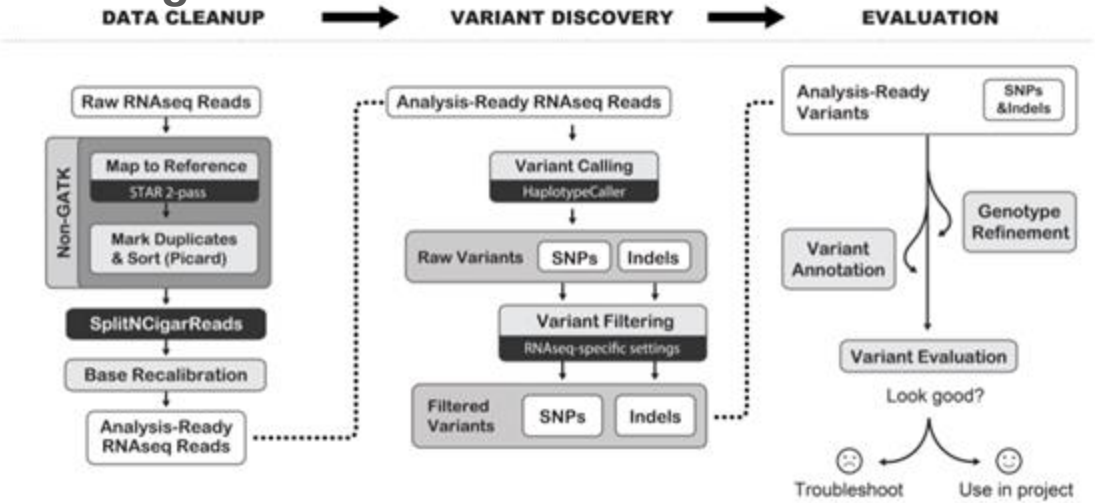
Part I. Transcriptomics | *Analysis*

Variant discovery: small variant calling

Goal: discover (or validate) small somatic variants (single nucleotide polymorphism or indels)

Medical relevance: many diseases are driven by small variants

Methods: Mapping to reference, and heavy filtering using estimated sequencing error rates and databases of known germline variants



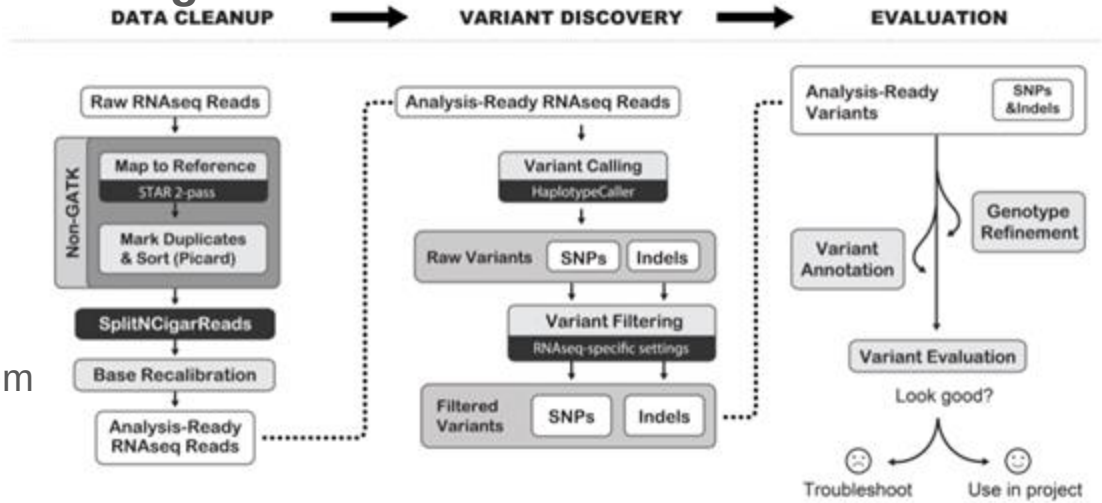
Schematic of the Genome Analysis ToolKit (GATK) best practices for small variant discovery from RNA-seq; *Source: https://gatk.broadinstitute.org*

Part I. Transcriptomics | *Analysis*

Variant discovery: small variant calling

Caveats:

- High false positive rate (due to sequencing error and high depth at some locations)
- High false positive rate (due to variants in low-expression genes)
- Useful for validation of mutations from WGS/WES
- Useful for allele specific expression quantification



Schematic of the Genome Analysis ToolKit (GATK) best practices for small variant discovery from RNA-seq; *Source: https://gatk.broadinstitute.org*

Part I. Transcriptomics | *Analysis*

Supervised analyses (i): differential expression analysis

Goal: explain biological differences between different conditions

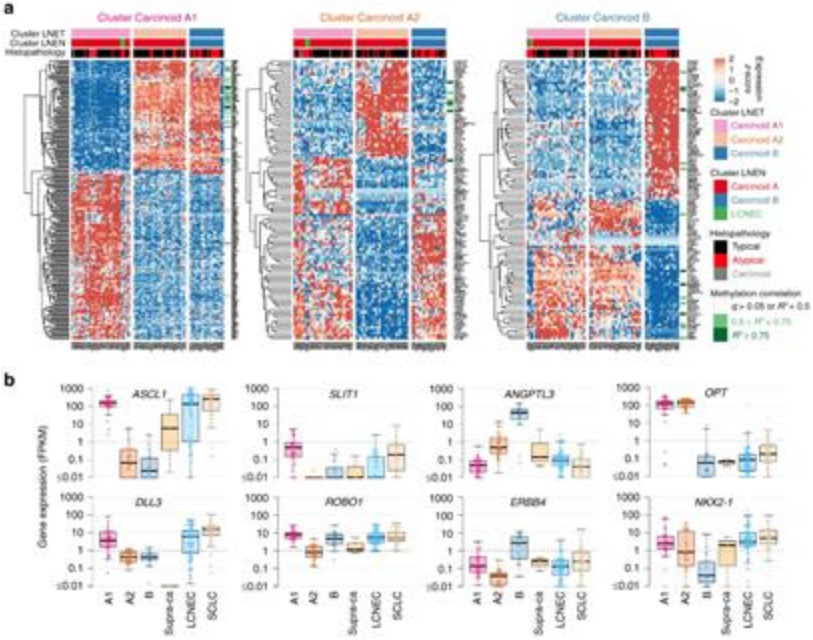
- Fitting model, correcting for confounding variables like batch, or accounting for clinical variables such as sex, age, environmental exposure (e.g., edgeR, DESeq2, limma)
- Analyzing list of genes obtained to understand differences (e.g., gene-set enrichment) or identify therapeutic targets

Part I. Transcriptomics | *Analysis*

Supervised analyses (i): differential expression analysis

Goal: explain biological differences between different conditions

- Fitting model, correcting for confounding variables like batch, or accounting for clinical variables such as sex, age, environmental exposure (e.g., edgeR, DESeq2, limma)
- Analyzing list of genes obtained to understand differences (e.g., gene-set enrichment) or identify therapeutic targets
- *Example: differential expression between molecular subtypes of lung cancer*



Differential expression analysis of lung neuroendocrine tumors. a. Heatmaps of DE genes. **b**. DE genes with clinical relevance. *Source: Alcala, Leblay, Gabriel, et al. Nature Communications 2019.*

Part I. Transcriptomics | *Analysis*

Supervised analyses (ii): machine learning

Goal: predict biological or clinical features using molecular data

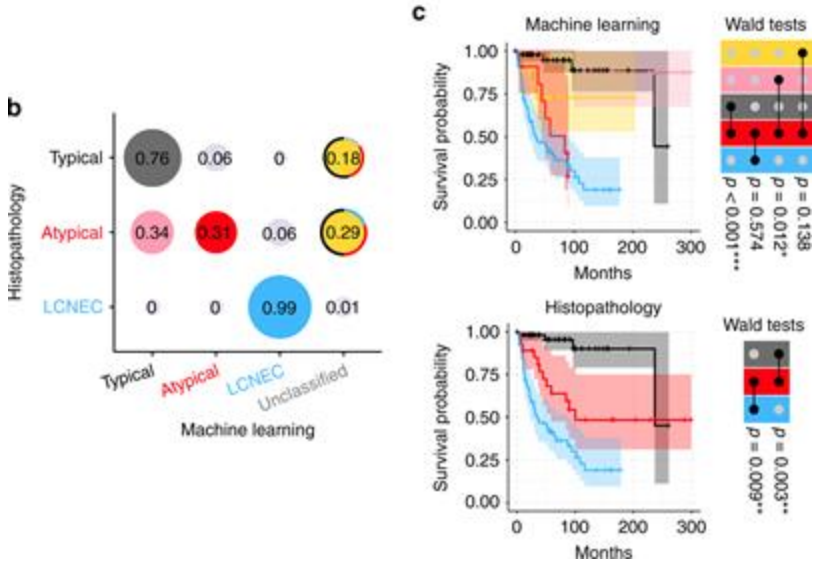
- Normalization of expression (e.g., Variance Stabilization)
- Training model (e.g., random forest, support vector machine, neural network)
- Testing model

Part I. Transcriptomics | *Analysis*

Supervised analyses (ii): machine learning

Goal: predict biological or clinical features using molecular data

- Normalization of expression (e.g., Variance Stabilization)
- Training model (e.g., random forest, support vector machine, neural network)
- Testing model
- *Example: predict tumor histopathological types based on molecular data.*



A random forest classifier stratifies atypical carcinoids into good- and bad-prognosis. b. Confusion matrix of the classifier. **c**. Kaplan-Meier survival curves. Model trained on 186 transcriptomes. *Source: Alcala, Leblay, Gabriel, et al. Nature Communications 2019.*

Part I. Transcriptomics | *Analysis*

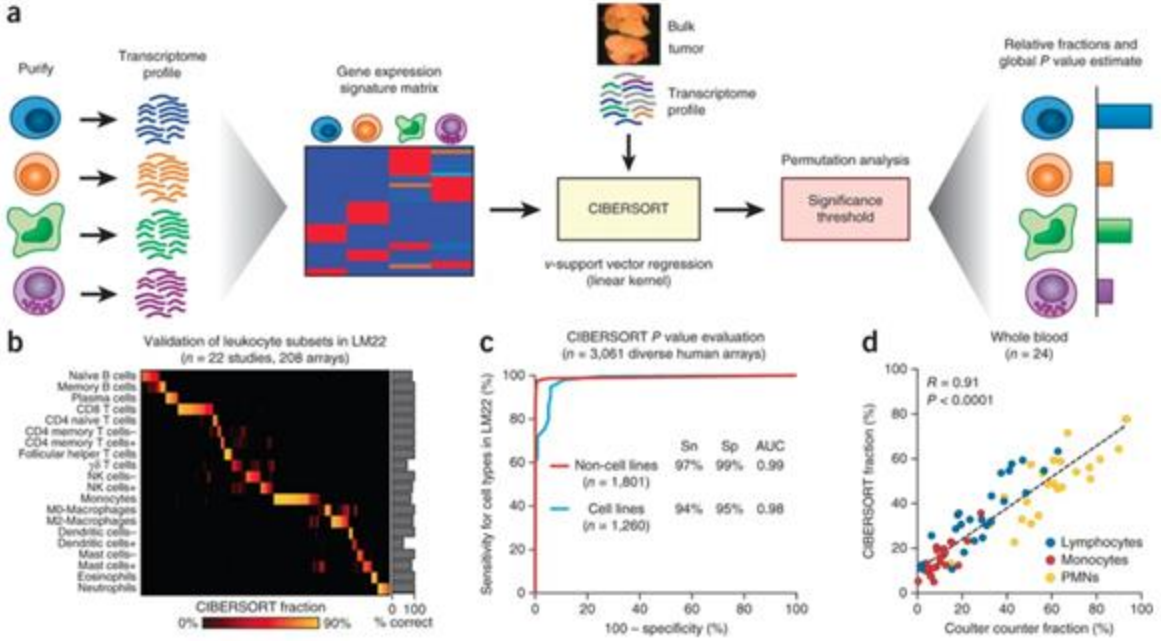
Supervised analyses (iii): Deconvolution

Goal

- infer the composition of the tissue into different cells

Model

- **M** = **F**.**B**, where mRNA mixture **M** and signature matrix **B** are known, and the vector **F** consisting of the fractions of each cell type in the mixture is unknown



CIBERSORT deconvolution workflow. *Source: Newman et al. Nature Methods 2015.*

Part I. Transcriptomics | *Analysis*

Unsupervised analyzes (i)

Goal: identify biological variation without *a priori* (exploratory or hypothesis generating analysis)

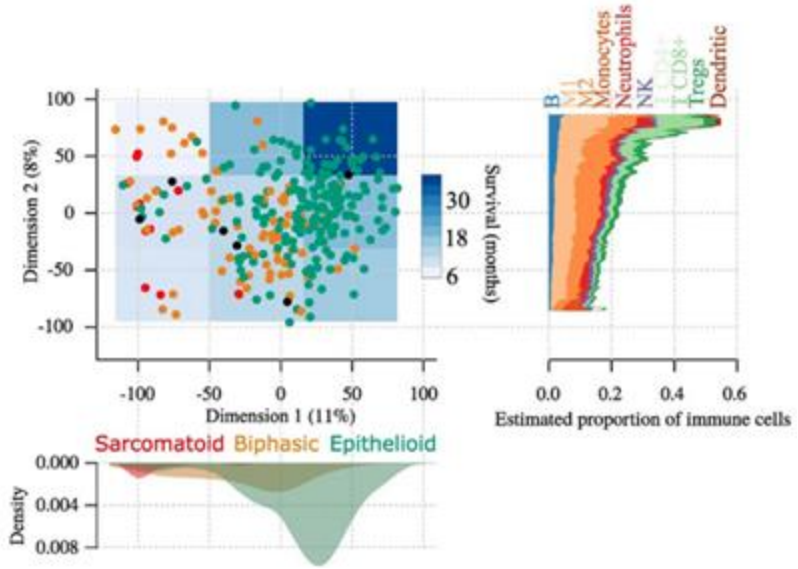
- **Clustering** (e.g., hierarchical, K-means, fuzzy C-means): identifying groups of samples with shared molecular profiles
- **Latent variable identification** (e.g., Principal Component Analysis, Independent Component Analysis, Canonical Correlations Analysis): identifying continuous sources of variation

Part I. Transcriptomics | *Analysis*

Unsupervised analyzes (i)

Goal: identify biological variation without *a priori* (exploratory or hypothesis generating analysis)

- *Example: identifying clinically relevant molecular variation without a priori in rare cancers. Most classifications of tumors use discrete entities, but some molecular data suggest that a continuum is more appropriate.*

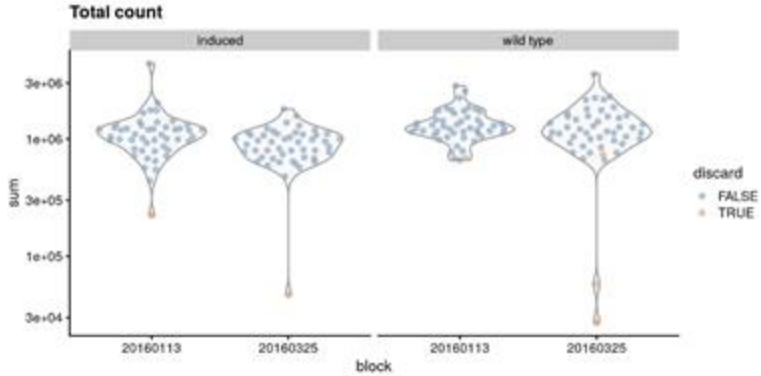


Principal Component Analysis of 210 malignant pleural mesothelioma (MPM) transcriptomes. Blue-colored rectangles represent overall survival. **Bottom**: density of the three histopathological types of MPM on dimension 1. **Right**: proportion of immune cells on dimension 2. *Source: Alcala, Mangiante et al. Ebiomedicine 2019.*

Part I. Transcriptomics | *Analysis*

Single-cell analysis

1. **QC:** Remove low-quality cells (damaged or badly captured), e.g., based on low total counts/cell, proportion of mitochondrial reads and number of non-zero features
2. **Normalization:** counts are normalized for library size differences and transformed to reduce variance
3. **Feature selection**: retaining genes that are highly variable

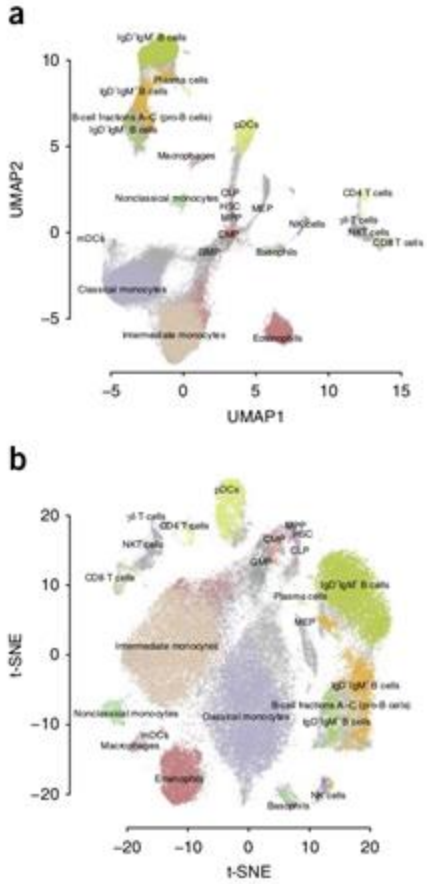


Example QC of scRNA-seq (total count/cell). *Source:* https://osca.bioconductor.org/quality-control.html

Part I. Transcriptomics | *Analysis*

Single-cell analysis

1. **QC:** Remove low-quality cells (damaged or badly captured), e.g., based on low total counts/cell, proportion of mitochondrial reads and number of non-zero features
2. **Normalization:** counts are normalized for library size differences and transformed to reduce variance
3. **Feature selection**: retaining genes that are highly variable
4. **Dimensionality reduction:** compact the data and reduce noise (PCA, or non-linear techniques like t-SNE and UMAP)



2D embedding of scRNA-seq of immune cell populations. a. UMAP. **b**. t-SNE. *Source:* Becht et al. Nature Biotechnology 2019.

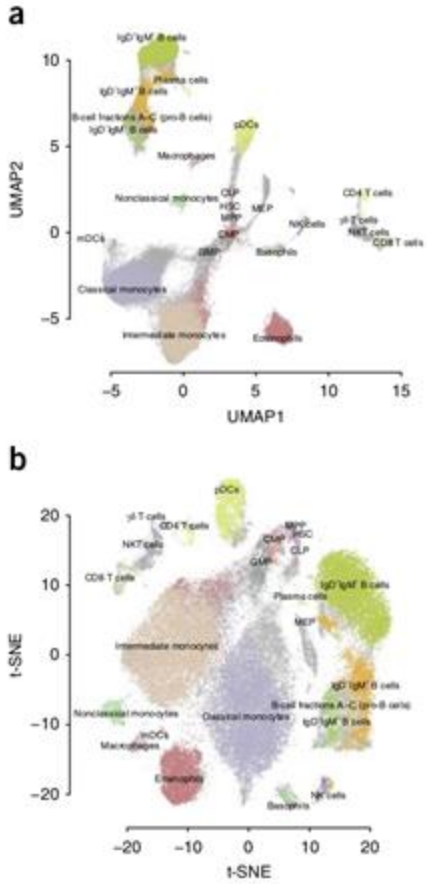
Part I. Transcriptomics | *Analysis*

Single-cell analysis

t-SNE algorithm:

1. **similarities between points in the original (high-dimensional) space** are computed using Gaussian distributions, with variances fitted based on a user-defined parameter (perplexity)
2. **Similarities between points in the output (low-dimensional) space** are computed using Student distributions with 1DF
3. **A cost function** (the Kullback-Leibler divergence between the two distributions) is optimized

Notes: Because similarities decrease exponentially with Euclidean distance between points, there is a high cost to misrepresenting short distances but a low cost for long distances



2D embedding of scRNA-seq of immune cell populations. a. UMAP. **b**. t-SNE. *Source:* Becht et al. Nature Biotechnology 2019.

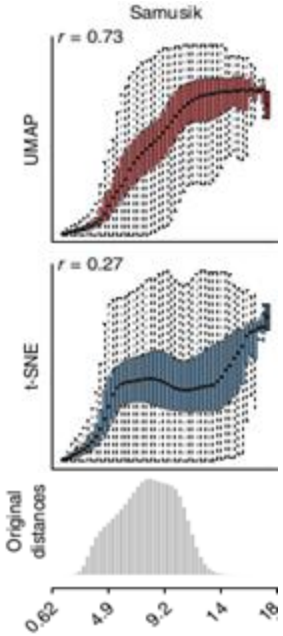
Part I. Transcriptomics | *Analysis*

Single-cell analysis

UMAP algorithm:

1. **similarities between points in the original (high-dimensional) space** are computed using fuzzy simplicial sets memberships
2. **Similarities between points in the output (low-dimensional) space** are computed using Student distributions with 1DF
3. **A cost function** (the cross-entropy) is optimized

Notes: UMAP has a faster running time because cross-entropy is easier to optimize, and is claimed to better preserve long distances

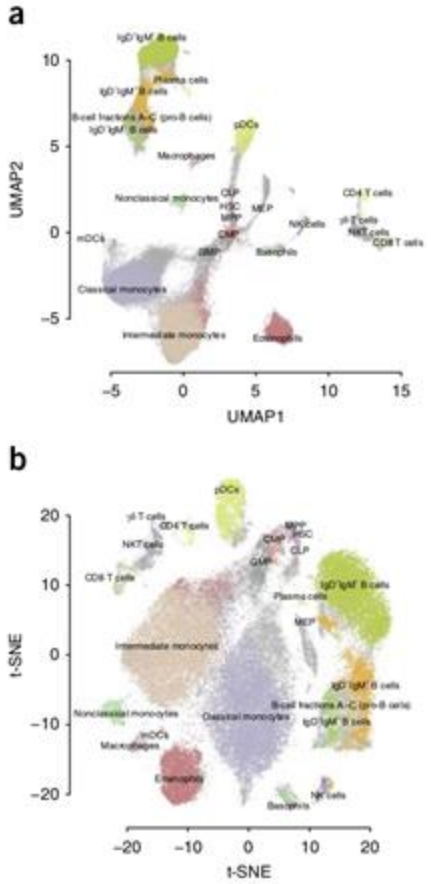


Preservation of original distances by UMAP and t-SNE. *Source:* Becht et al. Nature Biotechnology 2019.

Part I. Transcriptomics | *Analysis*

Single-cell analysis

1. **QC:** Remove low-quality cells (damaged or badly captured), e.g., based on low total counts/cell, proportion of mitochondrial reads and number of non-zero features
2. **Normalization:** counts are normalized for library size differences and transformed to reduce variance
3. **Feature selection**: retaining genes that are highly variable
4. **Dimensionality reduction:** compact the data and reduce noise (PCA, or non-linear techniques like t-SNE and UMAP)
5. **Cell clustering**: group similar expression profiles (biological states)
6. **Differential expression**: identifying marker genes between clusters, aggregating cells to create “pseudo-bulks” with the same sample and label pair, then perform DE



2D embedding of scRNA-seq of immune cell populations. a. UMAP. **b**. t-SNE. *Source:* Becht et al. Nature Biotechnology 2019.

Part II. Multi-omics | *Concepts*

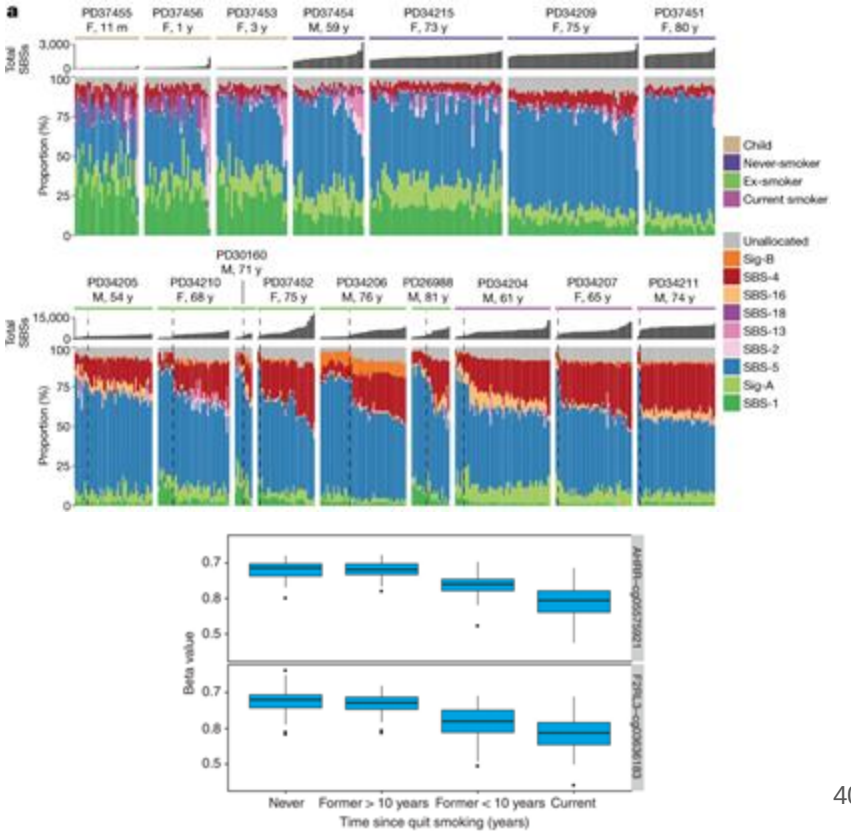
Interactions between ‘omic layers

Alterations in one ‘omic layer impact other layers, for instance

- eQTLs: genomic variants -> transcriptome & proteome
- Epigenome -> transcriptome & proteome

Processes of interest impact multiple layers

- **Environmental exposures** can leave mutational signature (genome) and leave epigenetic marks that impact gene regulation (transcriptome/proteome)

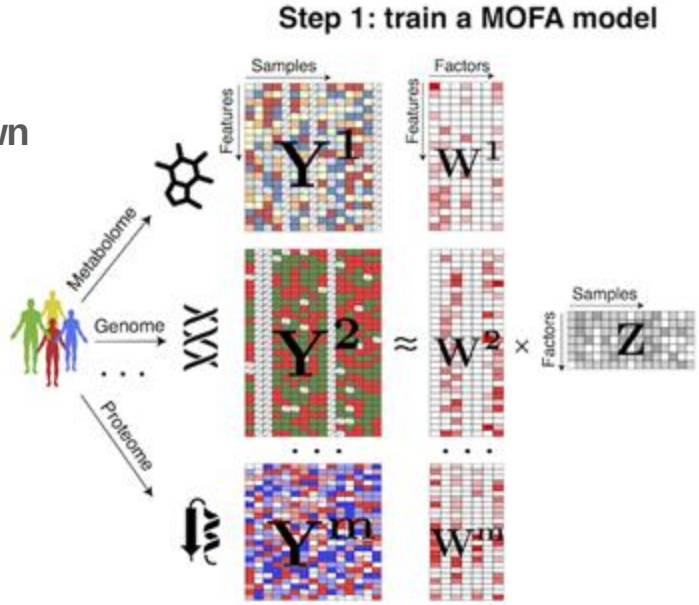


Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Identify **latent factors (unknown continuous variables)** representing biological variation **shared between modalities (e.g., genome, transcriptome)**

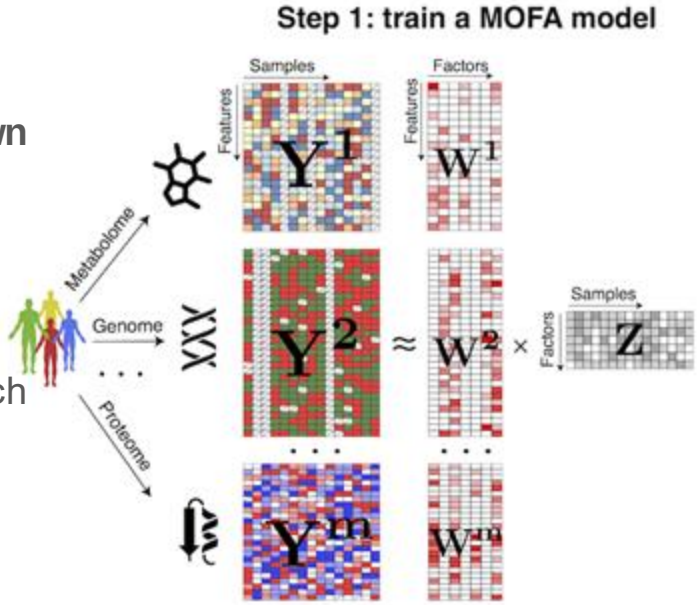


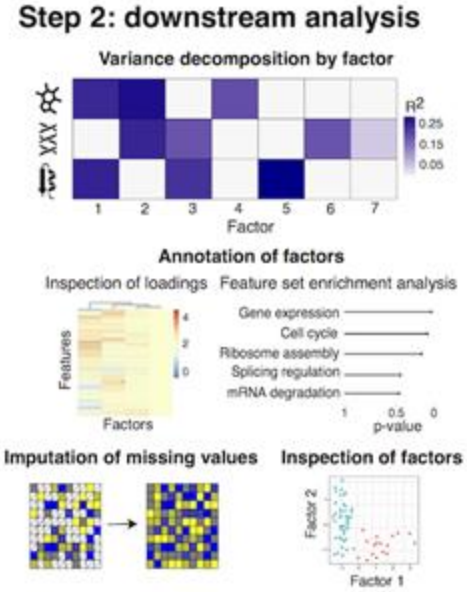
Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Identify **latent factors (unknown continuous variables)** representing biological variation **shared between modalities (e.g., genome, transcriptome)**
- Identify in which 'omic' layer each factor is active
- Downstream analysis to **understand what each factor represents**





International Agency for Research on Cancer

World Health Organization

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

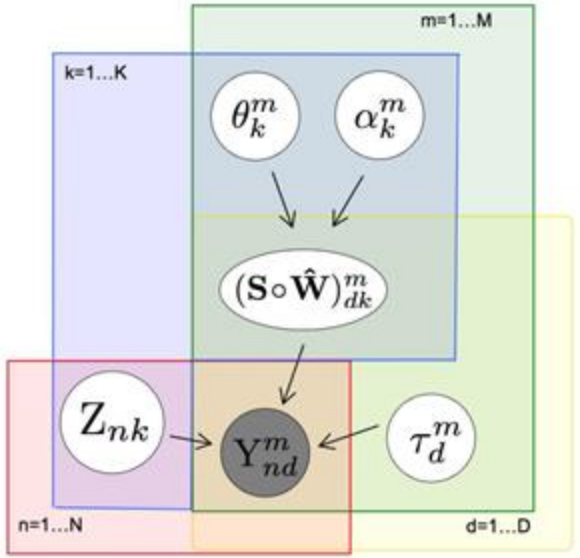
- **Generalization of Principal Component Analysis to multiple modalities** *M*

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- **Generalization of Principal Component Analysis to multiple modalities** *M*
- model $\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^m + \boldsymbol{\varepsilon}^m$,
- where $\mathbf{Y}^m$ is the matrix of observations for each sample *n* (rows) and each feature *d* (columns) for modality *m* (e.g., genomic alterations, expression)
- $\mathbf{Z}$ is the latent factors matrix (*N* by *K*) shared by all modalities *m*
- $\mathbf{W}^m$ is the weights (loadings) matrix (*K* by *M*) of *m*
- $\boldsymbol{\varepsilon}^m$ is the residual noise (column vector of size *N*)



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

Part II. Multi-omics | *Analysis*

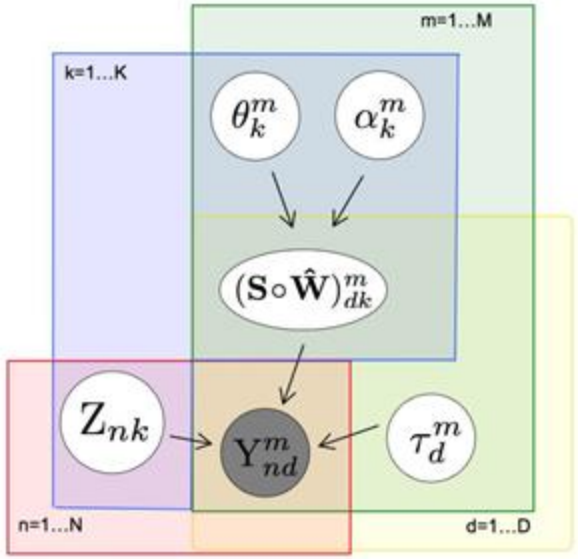
Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Model $\mathbf{Y}^m = \mathbf{ZW}^m + \boldsymbol{\varepsilon}^m$

Important points:

- Because **Z** is estimated from all 'omic' layers *m* and features *d*, the **model handles missing data naturally**
- The sparsity assumptions perform **automatic feature and factor selection**
- **Technical artifacts**, usually restricted to a single modality *k*, are separated from variation with **evidence from multiple modalities**
- **Correlations between modalities** are found (e.g., expression QTLs)



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

Part II. Multi-omics | *Analysis*

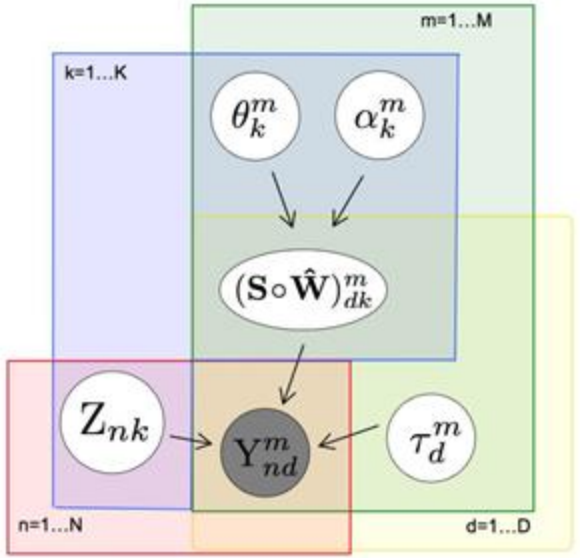
Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

- Model $\mathbf{Y}^m = \mathbf{ZW}^m + \boldsymbol{\varepsilon}^m$

Important points:

- the likelihood formulation implicitly gives more weight to modalities with many features, so **beware of imbalance between input data matrices** (e.g., a mutation matrix of 20 features will not influence much **Z** if an expression matrix with 10,000 features is also provided)



MOFA directed acyclic graph. *Source: Argelaguet et al. Mol Syst Biol 2018.*

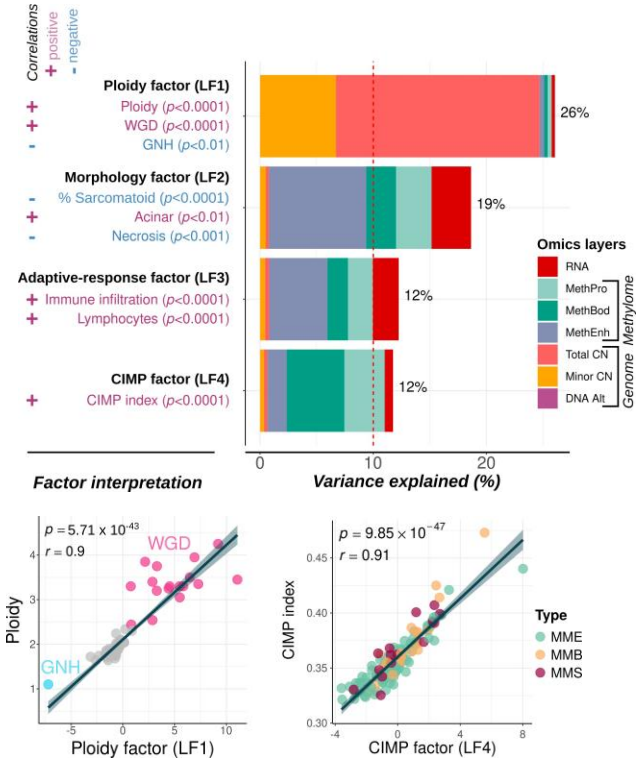
Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

Application to the rare and deadly malignant pleural mesothelioma:

- **Factor 1** mostly explains copy number variation and corresponds to **ploidy**
- **Factor 2** explains gene expression and methylation and separates **cell morphologies**
- **Factor 3** explains gene expression and methylation and separates hot and cold tumors
- **Factor 4** explains mostly methylation and corresponds to the CpG Island methylator phenotype



MOFA of 120 malignant pleural mesothelioma. *Source: Mangiante, Alcala, Di Genova, Sexton-Oates et al. Biorxiv 2021 (under review in Cancer Cell).*

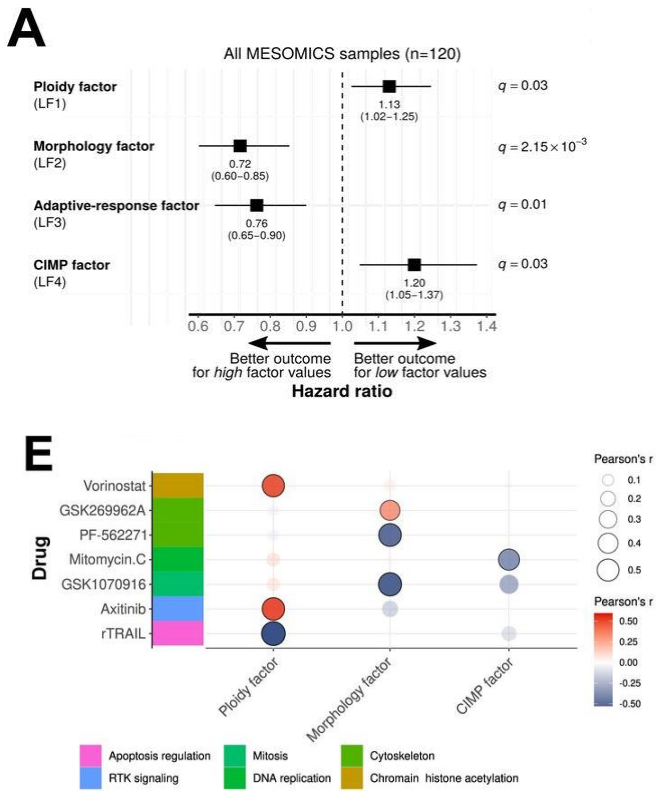
Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Multi-Omics Factor Analysis (MOFA)

Application to the rare and deadly malignant pleural mesothelioma:

- **All 4 factors are associated with patient survival**
- **Factors are associated with different drug responses**



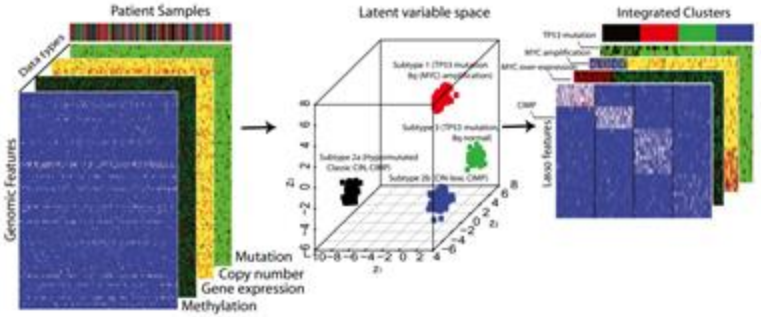
MOFA LFs are associated with clinical behavior. *Source: Mangiante, Alcala, Di Genova, Sexton-Oates et al. Biorxiv 2021 (under review in Cancer Cell).*

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Other tools

- **Integrative clustering** (iCluster+; Mo *et al*. *PNAS* 2013):
  - integrative latent factors identification (similar to MOFA) for dimensionality reduction
  - then clustering in reduced space (*K*-means)
  - *Specificities*: latent factors are not directly interpreted; **emphasis on clustering rather than continuous analyses**

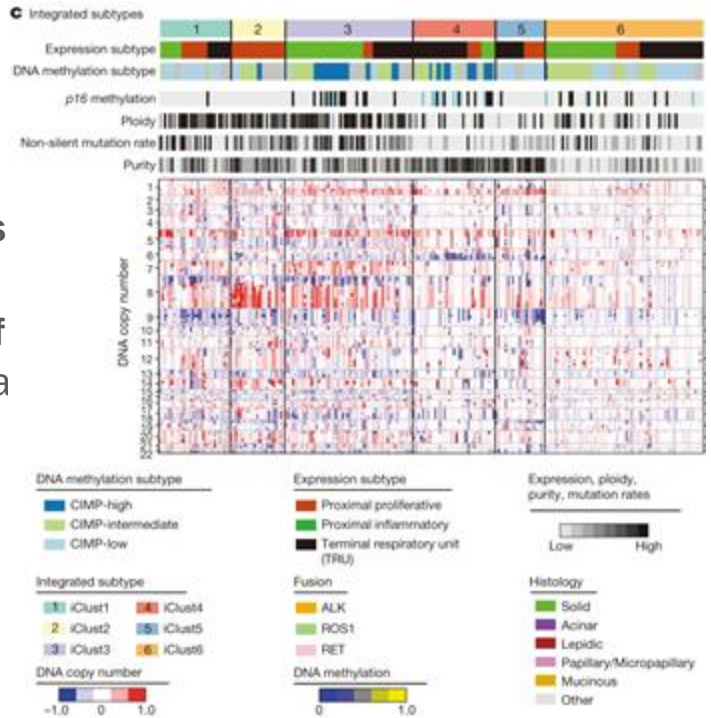


Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Other tools

- **Integrative clustering** (iCluster+; Mo *et al*. *PNAS* 2013):
- **Application (lung cancer): clusters summarize groups** identified in exomes, RNA-seq, and methylation data (top rows), **but fail to accurately represent the continuity of the data** (e.g., CNVs and ploidy do not seem to cluster tha well)



Integrative clustering with iCluster+ of 230 lung adenocarcinomas. *Source: TCGA Nature 2014.*

Part II. Multi-omics | *Analysis*

Tools for integration: unsupervised analyzes

Other tools

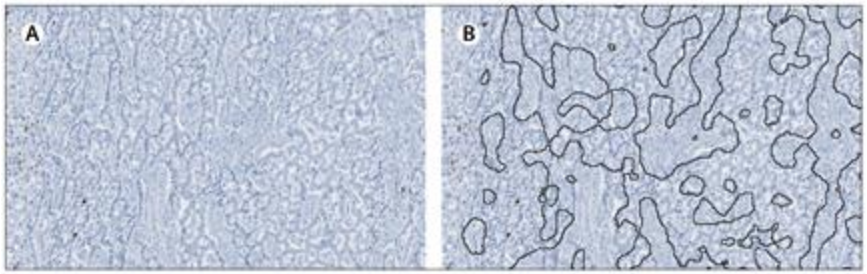
- **Other matrix factorization techniques**: DIABLO (Singh *et al*. Bioinformatics 2019), supervised method which extends Partial Least Squares regression (PLS) to multi-modal data to discriminate between multiple groups

Part III. Integration with other medical data | *Concepts*

Digital pathology

Histopathology: disease diagnosis through microscopic examination of stained tissue sections using histological techniques

Digital pathology: use of digitalized, high-resolution whole-slide images for sharing and analysis



(A) Whole-slide image of patient with a pancreatic neuroendocrine tumour. (B) The non-tumour regions are automatically outlined by a deep learning algorithm. *Source: Niazi et al. The Lancet Oncology 2019.*

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Goal: Predicting molecular features (e.g., molecular alterations, gene expression) solely based on pathology images

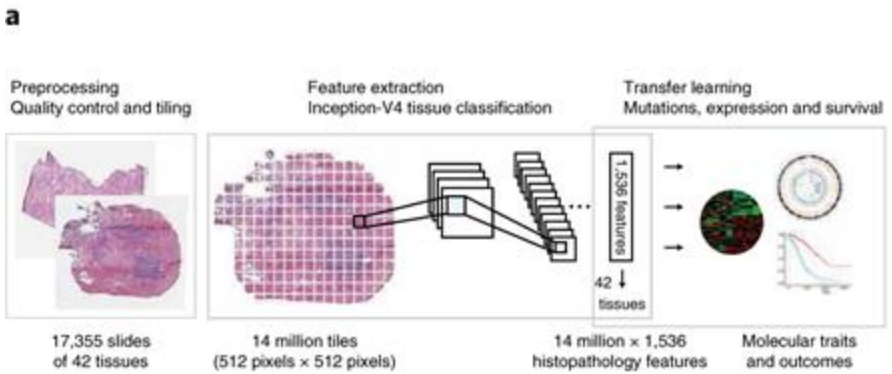
Advantages: once the model is trained, **relatively easy to use** (only requires digital slides) compared to genomic analyses (requiring sequencing and heavy data-processing)

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example: Predicting genomic features from images

- **Tiling**
- **Training on TCGA slides:** classification into 42 tissues (cancer types and normal tissue)



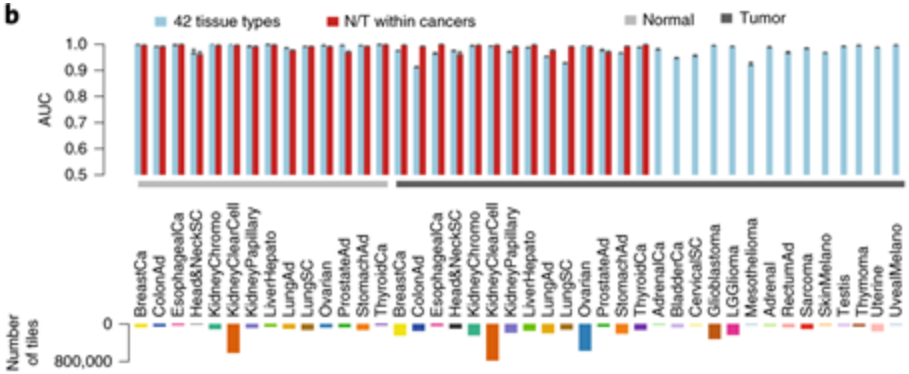
Deep learning workflow to identify clinically relevant genomic features from pathological images. *Source: Fu et al. Nature Cancer 2020.*

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example: Predicting genomic features from images

- **Tiling**
- **Training on TCGA slides:** classification into 42 tissues (cancer types and normal tissue)



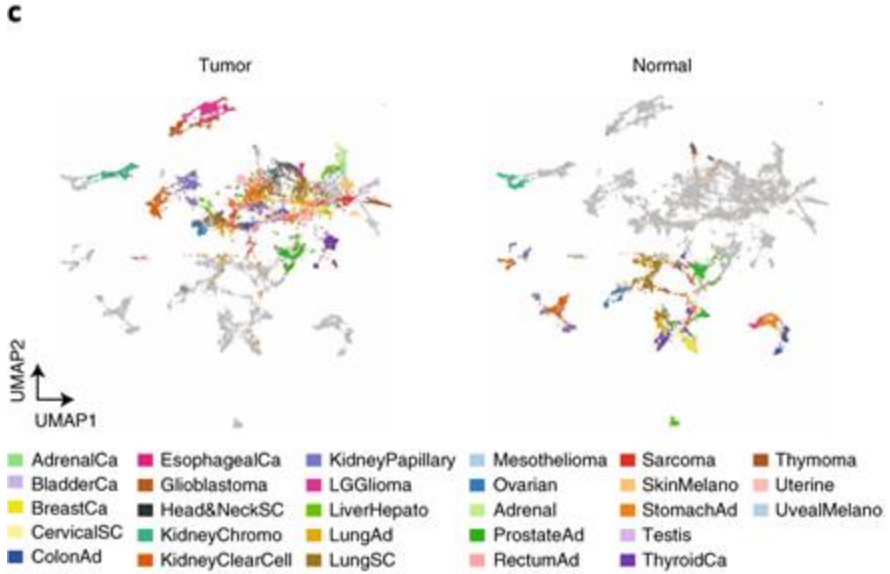
Classification accuracy into 42 tissues. *Source: Fu et al. Nature Cancer 2020.*

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example: Predicting genomic features from images

- **Tiling**
- **Training on TCGA slides:** classification into 42 tissues (cancer types and normal tissue)
- **Extraction of 1,536 features** from last hidden layer of the network



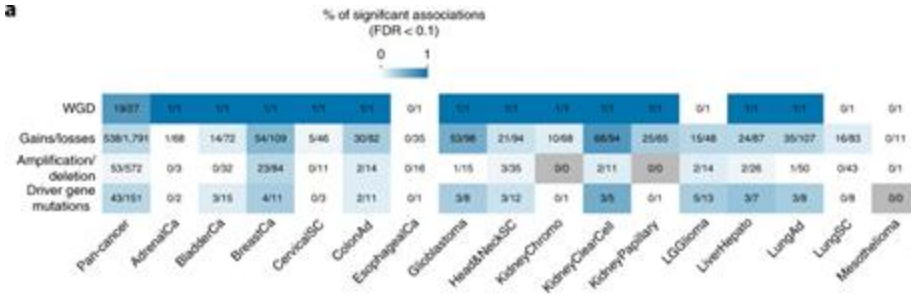
Two-dimensional representation (UMAP) of the 1,536 image features.
Source: Fu et al. Nature Cancer 2020.

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example: Predicting genomic features from images

- **Tiling**
- **Training on TCGA slides:** classification into 42 tissues (cancer types and normal tissue)
- **Extraction of 1,536 features** from last hidden layer of the network
- Use **penalized generalized linear model** regression to predict genomic features (glmnet R package) from the 1,536 features
- Assess **predictive power** using AUC



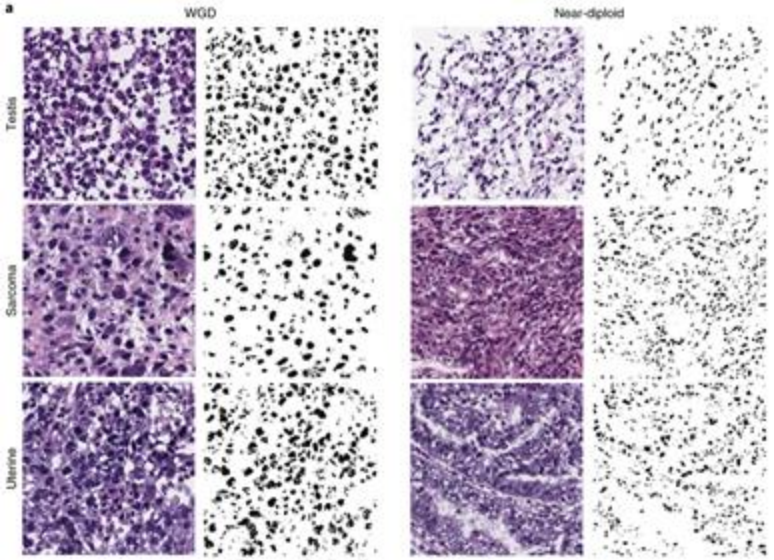
Association between genomic alterations and genomic features.
Source: Fu et al. Nature Cancer 2020.

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example: Predicting genomic features from images

- **Tiling**
- **Training on TCGA slides:** classification into 42 tissues (cancer types and normal tissue)
- **Extraction of 1,536 features** from last hidden layer of the network
- Use **penalized generalized linear model** regression to predict genomic features (glmnet R package) from the 1,536 features
- Assess **predictive power** using AUC



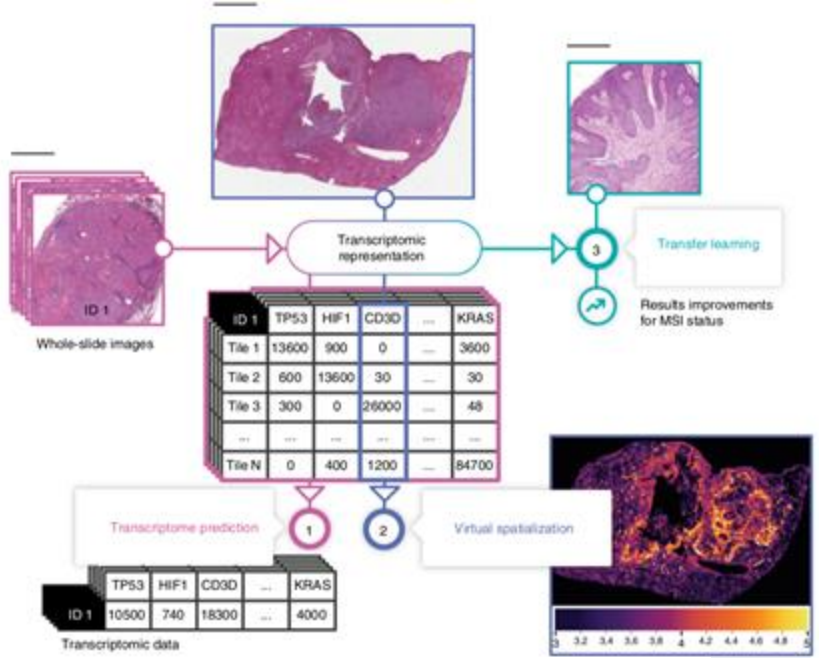
Slides with predicted whole-genome duplication (WGD) present larger nuclei (left) than slides predicted as near-diploid. *Source: Fu et al. Nature Cancer 2020.*

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example 2: Predicting gene expression from images

- **Tiling**
- **Training** on 8,725 patients from TCGA with slides and RNA-seq data: prediction of gene expression (quantitative variable)
- Extraction of scores per tile for interpretation



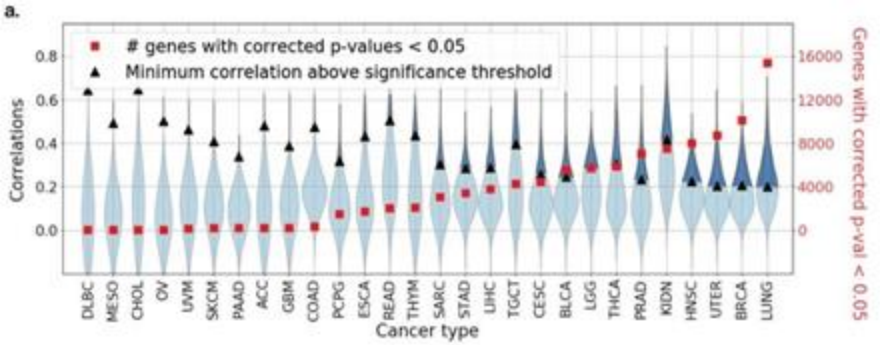
Deep learning workflow to predict gene expression from pathological images. *Source: Owkins Nature Communications 2019.*

Part III. Integration with other medical data | *Analysis*

Integrating genomics and whole-slide images using deep learning

Example 2: Predicting gene expression from images

- **Tiling**
- **Training** on 8,725 patients from TCGA with slides and RNA-seq data: prediction of gene expression (quantitative variable)
- Extraction of scores per tile for interpretation
- Some cohorts are much more amenable to prediction (lung, breast cancer)



Deep learning predictions of gene expression across TCGA cohorts.
Source: Schmauch et al. Biorxiv 2020.