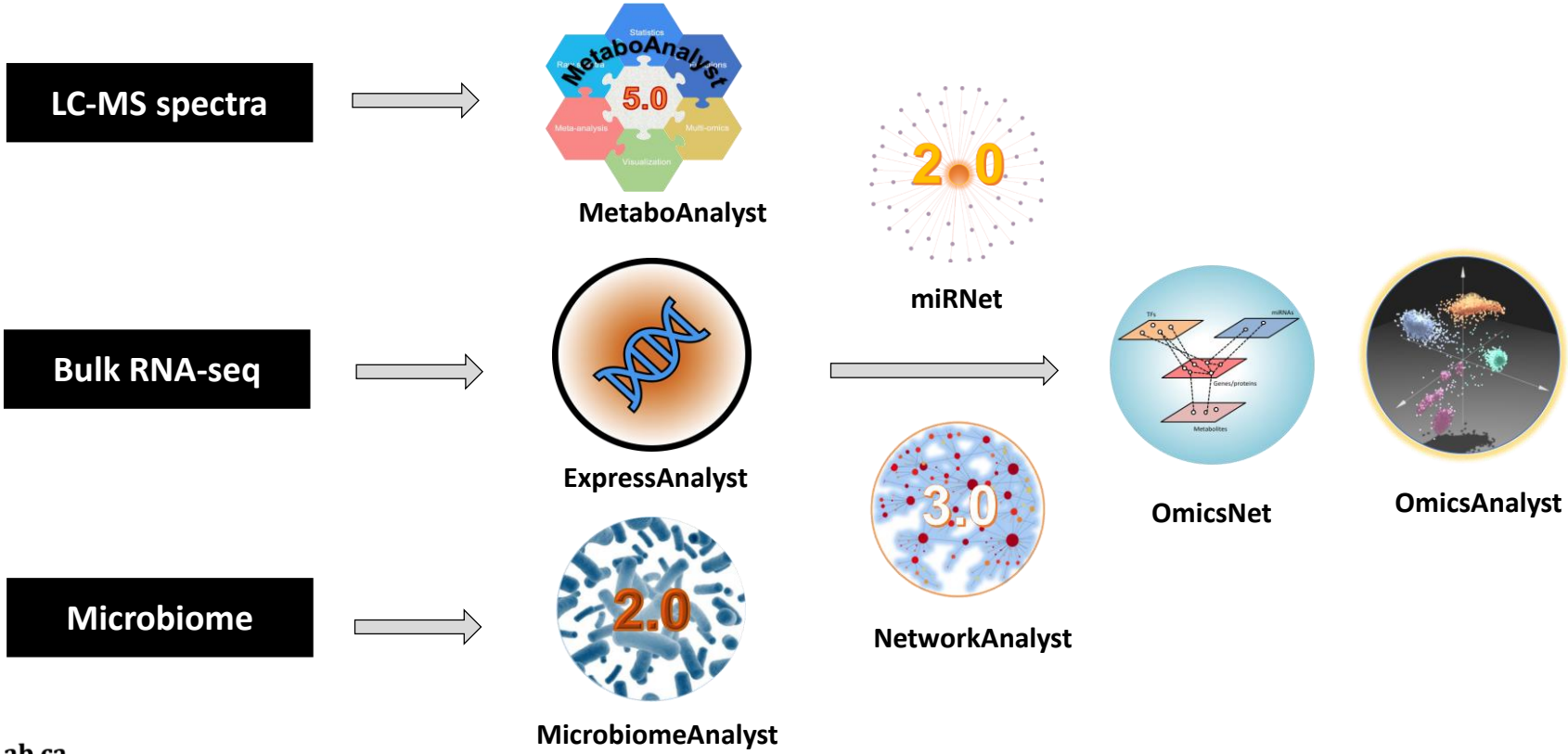
Omics Data Science Training Course

Winter 2024

Raw data ➔ statistics ➔ networks ➔ functions



Our Resources

Recordings & slides

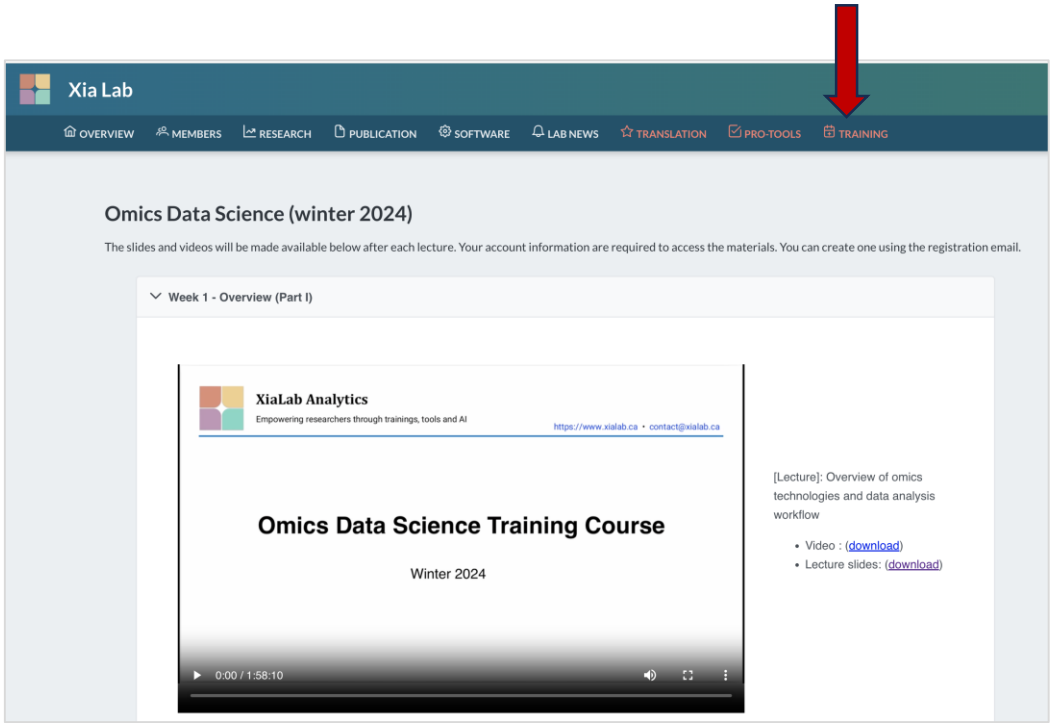
- https://www.xialab.ca under the "**Training**" tab within 3 days after lecture
- Faster with Firefox

Community tool:

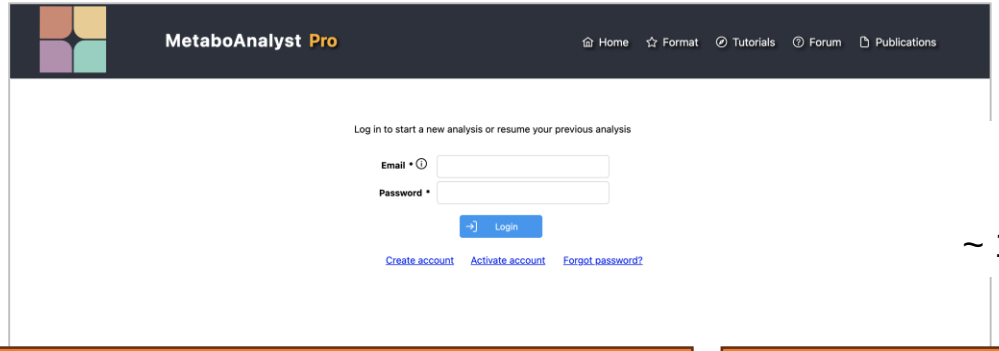
- https://www.####.ca

Pro Tools:

- https://pro.####.ca
  - You will be assigned to one of the cloud nodes



About the “pro” tools



Community version
~ 5000 users / day
~ 100s concurrent users

A total of six tools have their “pro” versions

- https://pro.metaboanalyst.ca
- https://pro.microbiomeanalyst.ca
- https://pro.expressanalyst.ca
- https://pro.omicsnet.ca
- https://pro.omicsanalyst.ca
- https://pro.mirnet.ca

1. Dedicated computing
2. Live report & project management
3. More stable
4. Prioritized support

Our Syllabus

Topic	Date	Lecture	Lab
Omics Data Science Foundations	Jan. 6	Omics data processing, statistics and visualization	--
	Jan. 13	From raw data to functional insights	--
Transcriptomics	Jan. 20	Gene expression data analysis (part I)	ExpressAnalyst & NetworkAnalyst
	Jan. 27	Gene expression data analysis (part II)	ExpressAnalyst & Seq2Fun
Proteomics, Networks, & Biomarkers	Feb. 3	Biological network analysis & gene regulatory networks	NetworkAnalyst & miRNet
	Feb. 10	Proteomics & biomarker analysis	ExpressAnalyst & MetaboAnalyst
Metabolomics	Feb. 17	Targeted metabolomics data analysis	MetaboAnalyst
	Feb. 24	LC-MS untargeted metabolomics data analysis	MetaboAnalyst
Microbiomics	Mar. 2	Marker gene data analysis	MicrobiomeAnalyst
	Mar. 9	Shotgun metagenomics data analysis	MicrobiomeAnalyst
Multi-omics	Mar. 16	Knowledge-driven multi-omics integration	OmicsNet
	Mar. 23	Data-driven multi-omics integration	OmicsAnalyst

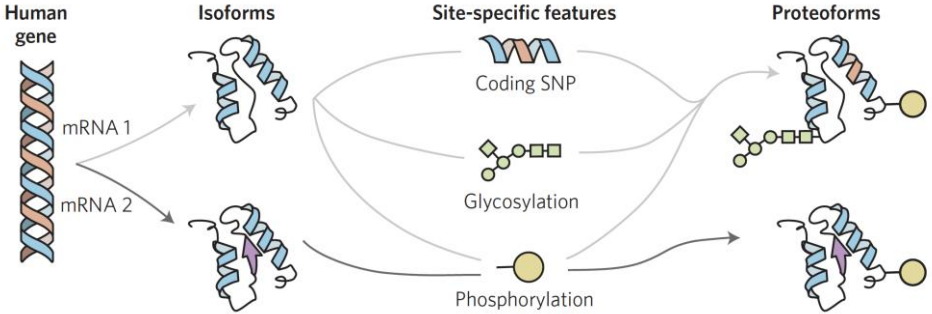
Schedule for today

Time	Topics
9:00 – 9:10	General introduction & recap
9:10 – 9:40	Proteomics data analysis workflow
9:45 – 10.15	Live demo & hands on
10:20 – 10:50	Biomarker analysis
10:55 – 11:25	Live demo & hands on
Summary and discussion	

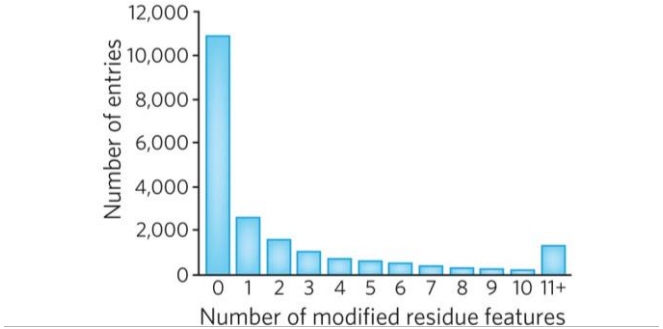
Proteomics & proteoform

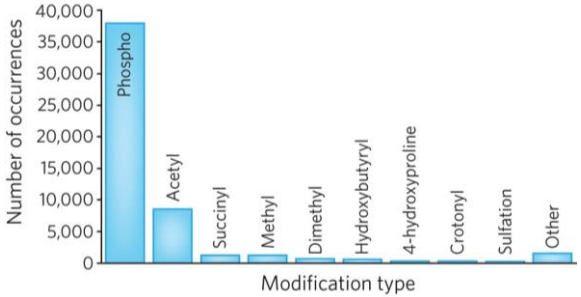
- Proteomics: study the entire set of proteins produced by a cell
  - Include functional, structural and quantitative proteomics
- A proteoform is defined by its exact amino acid sequence combined with any post translational modifications.
  - Each “protein” arising from a gene may exist in numerous different proteoforms.
- Technological platform is similar to metabolomics
  - LC-MS based shotgun proteomics
- Statistical & functional analysis are similar to transcriptomics (proteins ⇔ genes)

Protein & their proteoforms



It is estimated to have at least ~6 million proteoforms (2016)



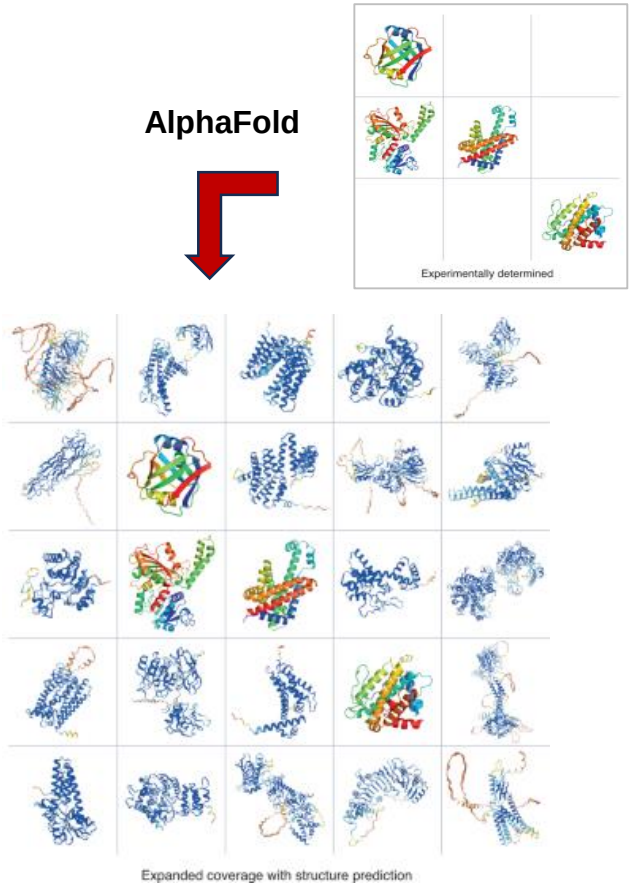


https://www.nature.com/articles/nchembio.2576

Three types of proteomics research

1. Structural proteomics (protein 3D structures)
2. Functional proteomics (large-scale PPI studies)
   - Yeast two-hybrid screening (Y2H)
   - Affinity purification coupled to mass spectrometry.
3. Quantitative proteomics (abundance & biomarkers)
   - Gel-based proteomics
   - Protein microarrays
   - MS-based proteomics (shotgun proteomics)

Structural Proteomics



nature methods

Explore content ▾ About the journal ▾ Publish with us ▾

nature › nature methods › editorials › article

Editorial | Published: 11 January 2022

Method of the Year 2021: Protein structure prediction

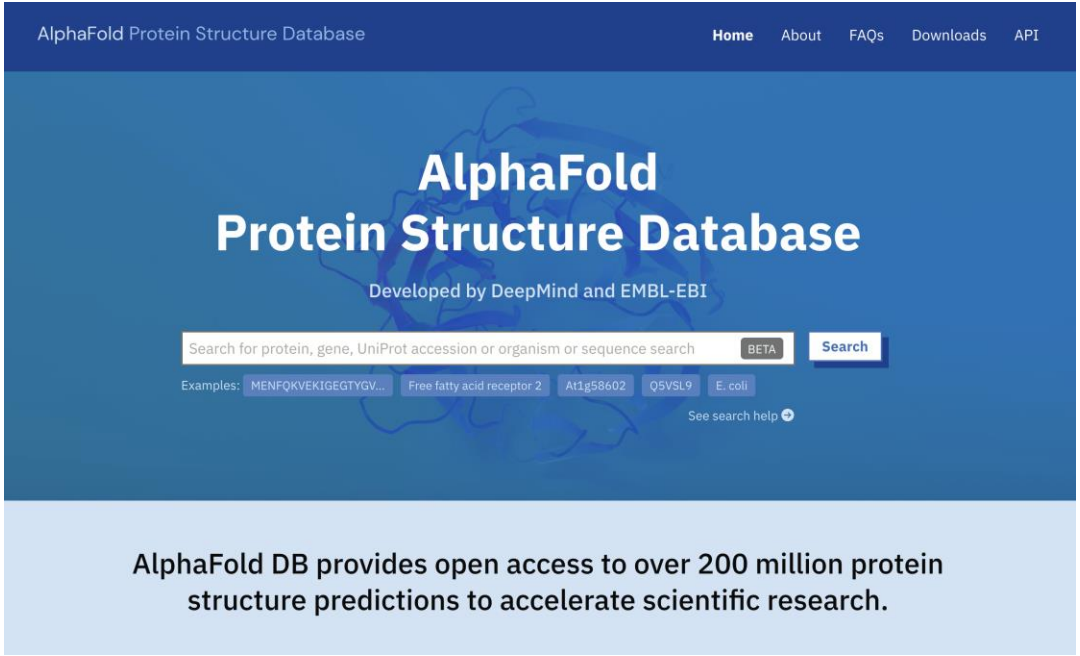
Nature Methods **19**, 1 (2022) | Cite this article

27k Accesses | **17** Citations | **428** Altmetric | Metrics

Deep Learning based approaches for protein structure prediction have sent shock waves through the structural biology community. We anticipate far-reaching and long-lasting impact.

Structural Proteomics (II) - a “solved” problem?

- AlphaFold was trained on protein chains in the PDB.
- Can make a strong prediction based on a multiple sequence alignment alone.
  - ✓ Templates are not a critical input to make an accurate prediction
  - ✓ Can ignore templates if they appear unhelpful



➔ predict the impact of single mutations on protein stability

Functional proteomics – PPI

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature › articles › article

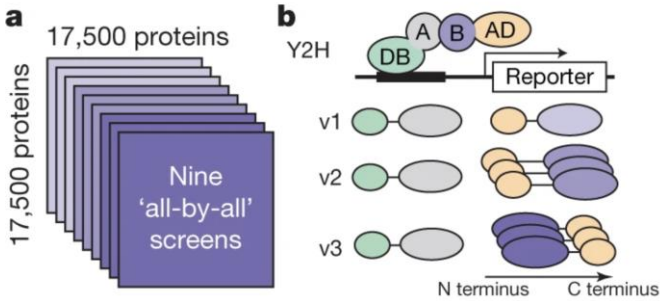
Article | Published: 08 April 2020

A reference map of the human binary protein interactome

Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J. Campos-Laborie, Benoit Charloteaux, Dongsic Choi, Atina G. Coté, Meaghan Daley, Steven Deimling, Alice Desbuleux, Amélie Dricot, Marinella Gebbia, Madeleine F. Hardy, Nishka Kishore, Jennifer J. Knapp, István A. Kovács, Irma Lemmens, Miles W. Mee, Joseph C. Mellor, … Michael A. Calderwood + Show authors

Nature **580**, 402–408 (2020) | Cite this article

68k Accesses | **534** Citations | **453** Altmetric | Metrics



HuRI

The Human Reference Interactome

PROTEINS	INTERACTIONS
9094	64006

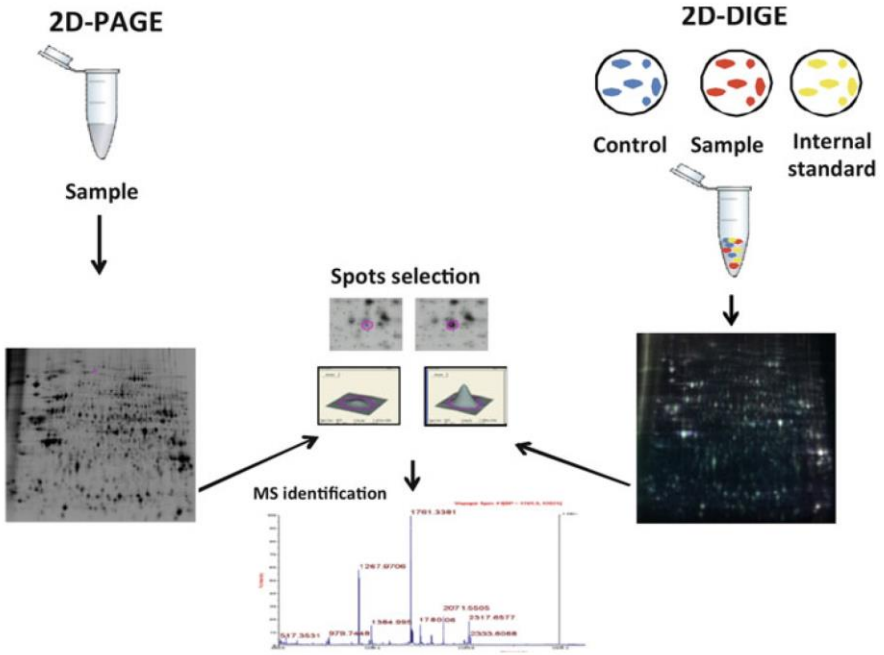
Literature Benchmark

PROTEINS	INTERACTIONS
6047	13441

http://www.interactome-atlas.org/

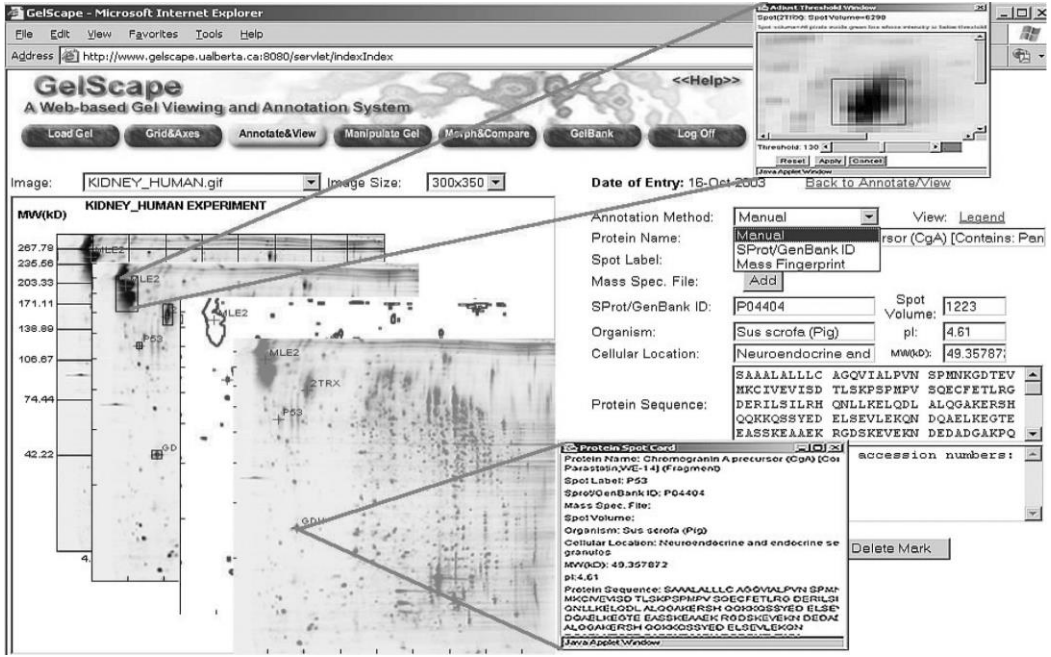
Quantitative proteomics based on gels

1) Separating proteins according to their isoelectric point (pI) and molecular weight (MW)
2) Comparison of the spot sizes reveals differences in the protein amount
3) MS identifies the proteins with significant changes



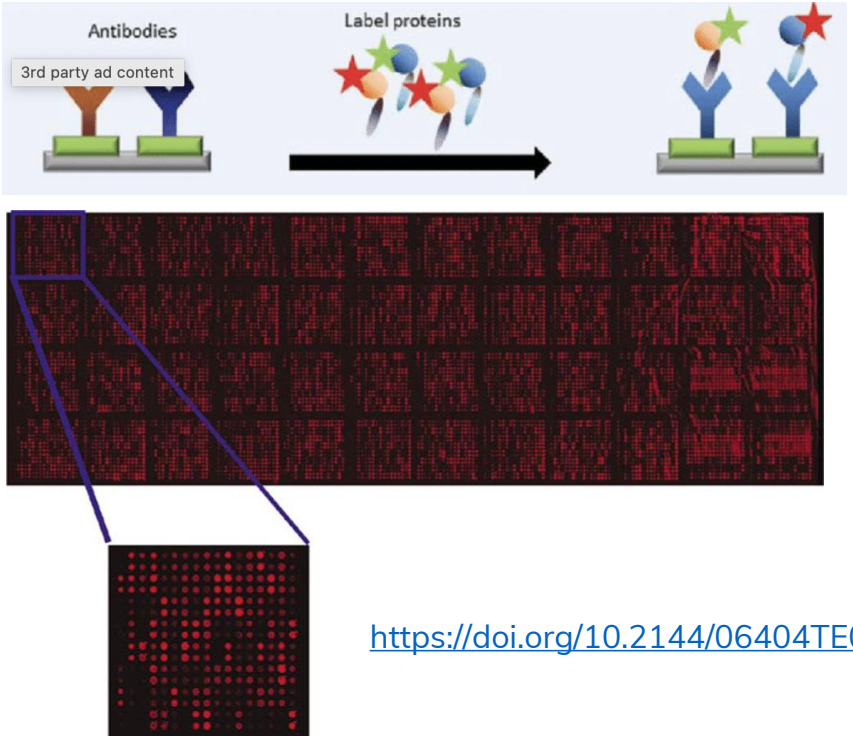
Imaging-based analysis

1. Load the gel image
2. Lay the grid
3. Annotate the spots
4. Compare with database/archived gel images
5. …..



Protein microarrays (I)

- Immobilize probes on protein chip
  - **Antibodies**
  - Aptamers
  - Affibodies
- Targeted proteins can be detected either by direct labeling or using a reporter antibody in sandwich assay format

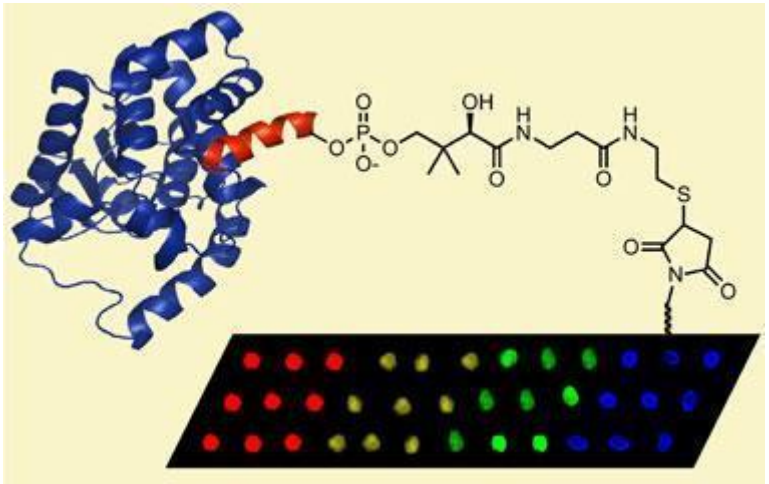


https://doi.org/10.2144/06404TE01

Protein microarrays (II) – many challenges

Production of reliable, consistent, high-throughput proteins that are correctly folded & functional are very challenging

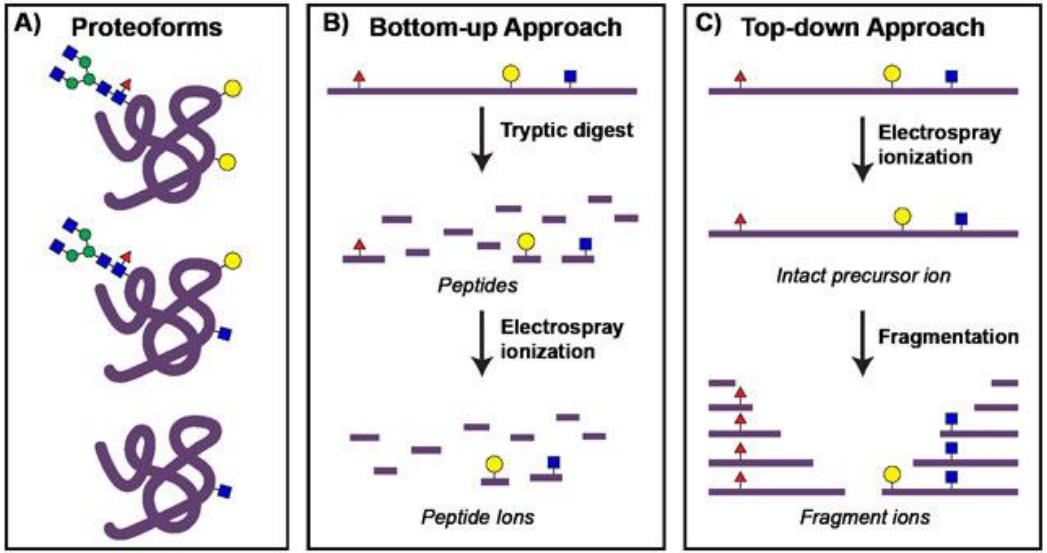
- Require a lot more steps in its creation than does a DNA chip
- Different surface chemistries
- Requiring long time in storage
- Reducing non-specific binding by the capture agents
- Complete representation of the proteome



DOI: 10.1021/ja8030278

MS-based quantitative proteomics

- Bottom-up:
  - Shotgun proteomics
  - The protein samples are first proteolytically digested into peptides before analyzing in MS
- Top-down:
  - Intact proteins are directly analyzed by MS



https://www.technologynetworks.com/proteomics/articles/a-fusion-of-proteomic-practices-the-indisputable-complementarity-of-bottom-up-and-top-down-337094

Latest breakthrough in MS-based proteomics

nature biotechnology

Article

https://doi.org/10.1038/s41587-023-02099-7

Ultra-fast label-free quantification and comprehensive proteome coverage with narrow-window data-independent acquisition

Received: 14 June 2023

Accepted: 13 December 2023

Published online: 01 February 2024

Check for updates

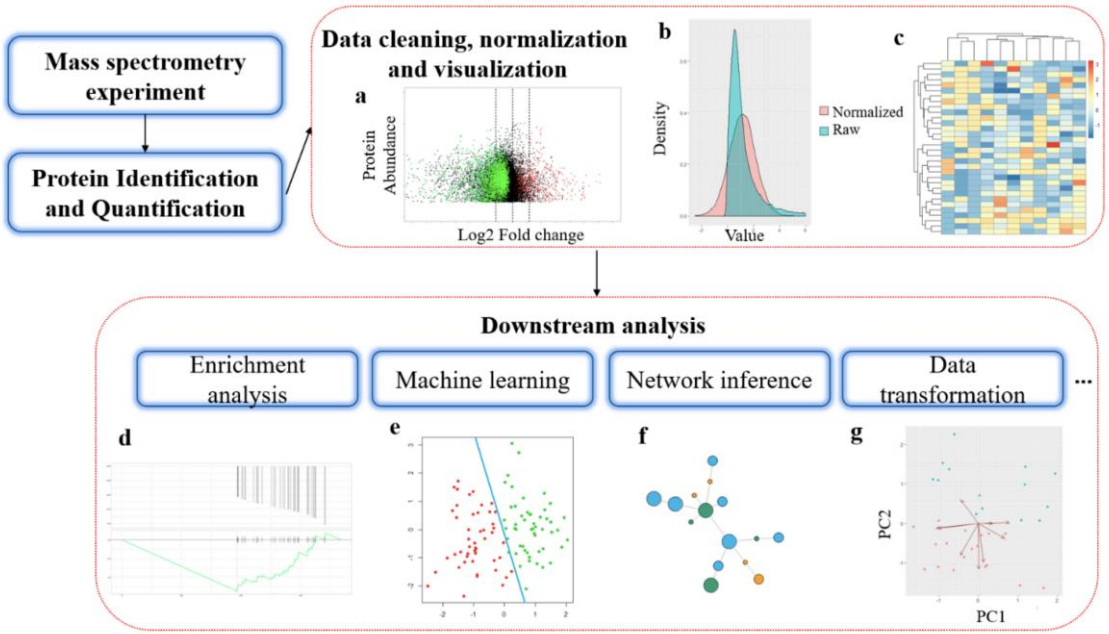
Ulises H. Guzman[1,7], Ana Martinez-Val[1,7], Zilu Ye[1,2,7], Eugen Dar
Tabiwang N. Arrey[3], Anna Pashkova[3], Santosh Renuse[4], Eduard Den
Johannes Petzoldt[3], Amelia C. Peterson[3], Florian Harking[1], Ole Øste
Rasmus Rydbirk[5], Susana Aznar[6], Hamish Stewart[3], Yue Xuan
Daniel Hermanson[4], Stevan Horning[3], Christian Hock[3], Alexander M
Vlad Zabrouskov[4] & Jesper V. Olsen[1]

Mass spectrometry (MS)-based proteomics aims to characterize comprehensive proteomes in a fast and reproducible manner. Here we present the narrow-window data-independent acquisition (nDIA) strategy consisting of high-resolution MS1 scans with parallel tandem MS (MS/MS) scans of ~200 Hz using 2-Th isolation windows, dissolving the differences between data-dependent and -independent methods. This is achieved by pairing a quadrupole Orbitrap mass spectrometer with the asymmetric track lossless (Astral) analyzer which provides >200-Hz MS/MS scanning speed, high resolving power and sensitivity, and low-ppm mass accuracy. The nDIA strategy enables profiling of >100 full yeast proteomes per day, or 48 human proteomes per day at the depth of ~10,000 human protein groups in half-an-hour or ~7,000 proteins in 5 min, representing 3× higher coverage compared with current state-of-the-art MS. Multi-shot acquisition of offline fractionated samples provides comprehensive coverage of human proteomes in ~3 h. High quantitative precision and accuracy are demonstrated in a three-species proteome mixture, quantifying 14,000+ protein groups in a single half-an-hour run.

General bioinformatics workflow

1. Data preprocessing
   - ➢ Protein ID & quantification
2. Data processing & normalization
3. Statistical analysis
4. Functional analysis

#1 and #2 are unique to omics technology. #3 and #4 are common

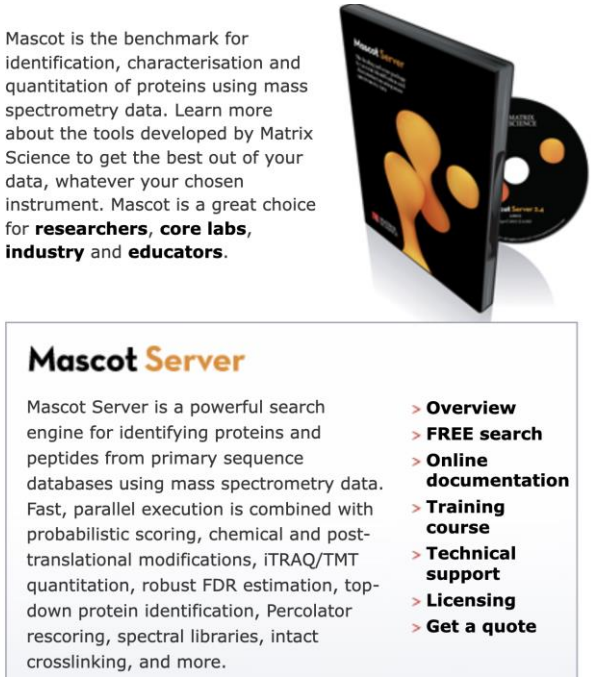


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7216093/

Data Pre-processing

MS-based protein identification

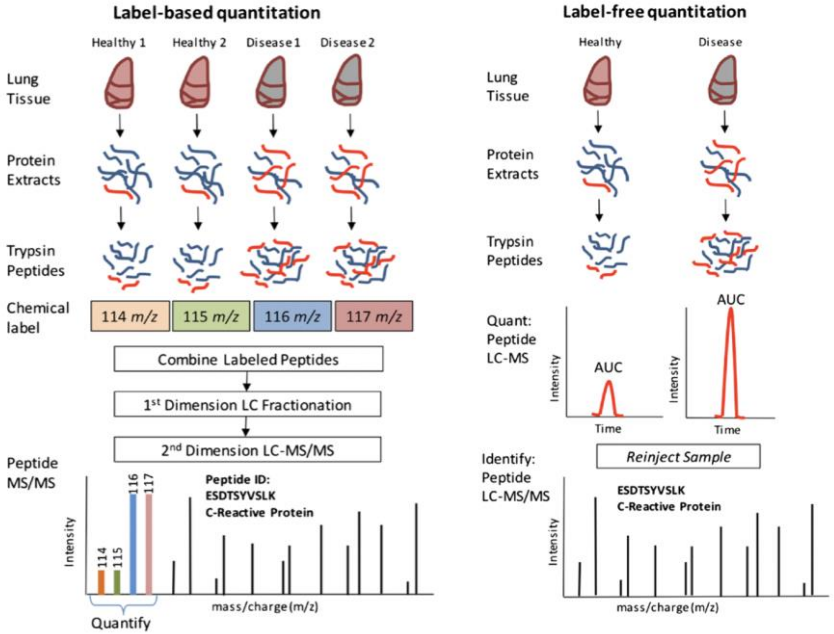
- Searching against the fragmentation spectra databases
  - A target database is established from in silico digestion of all expressed or hypothetical protein sequences. Then a peptide spectrum match score is calculated for each fragmentation spectra and all theoretical fragmentation spectra information from the target database.
    - Mascot, SEQUEST, etc
- de novo peptide sequencing
  - The peptide sequence is determined from fragmentation spectra information and fragmentation method
    - DeepNovo-DIA



https://www.matrixscience.com/

MS-based protein quantification

1. Labeling methods
   - Using isobaric stable isotope labeled chemical tags
     - Have identical mass, vary in terms of distribution of heavy isotopes in their structure
     - Distribution of the Isotope patterns allowing the relative quantifications
   - Mix samples (multiplexing)
2. Label-free methods
   - Different samples are acquired from separate LC-MS/MS experiments
   - Quantification based on the peak area (or ion intensity in LC-MS) or spectral counting (MS2)



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4883989/

MaxQuant for protein quantification

- Supporting both label-based and label-free methods

PROTOCOL UPDATE

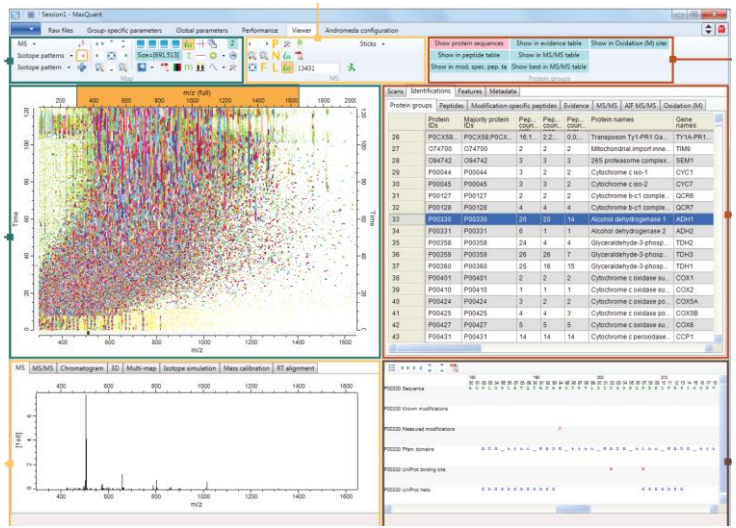
The MaxQuant computational platform for mass spectrometry–based shotgun proteomics

Stefka Tyanova[1,2], Tikira Temu[1,2] & Juergen Cox[1]

[1]Computational Systems Biochemistry, Max-Planck Institute for Biochemistry, Martinsried, Germany. [2]These authors contributed equally to this work.
Correspondence should be addressed to J.C. (cox@biochem.mpg.de).

Published online 27 October 2016; doi:10.1038/nprot.2016.136

MaxQuant is one of the most frequently used platforms for mass-spectrometry (MS)-based proteomics data analysis. Since its first release in 2008, it has grown substantially in functionality and can be used in conjunction with more MS platforms. Here we present an updated protocol covering the most important basic computational workflows, including those designed for quantitative label-free proteomics, MS1-level labeling and isobaric labeling techniques. This protocol presents a complete description of the parameters used in MaxQuant, as well as of the configuration options of its integrated search engine, Andromeda. This protocol update describes an adaptation of an existing protocol that substantially modifies the technique. Important concepts of shotgun proteomics and their implementation in MaxQuant are briefly reviewed, including different quantification strategies and the control of false-discovery rates (FDRs), as well as the analysis of post-translational modifications (PTMs). The MaxQuant output tables, which contain information about quantification of proteins and PTMs, are explained in detail. Furthermore, we provide a short version of the workflow that is applicable to data sets with simple and standard experimental designs. The MaxQuant algorithms are efficiently parallelized on multiple processors and scale well from desktop computers to servers with many cores. The software is written in C# and is freely available at http://www.maxquant.org.

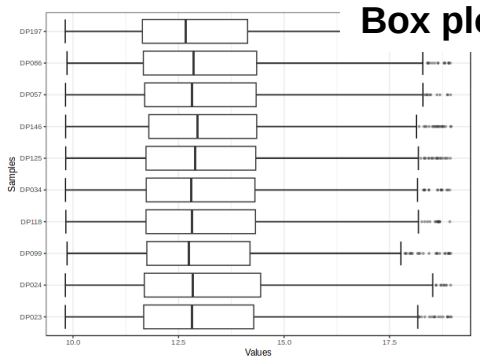


Data Processing & Normalization

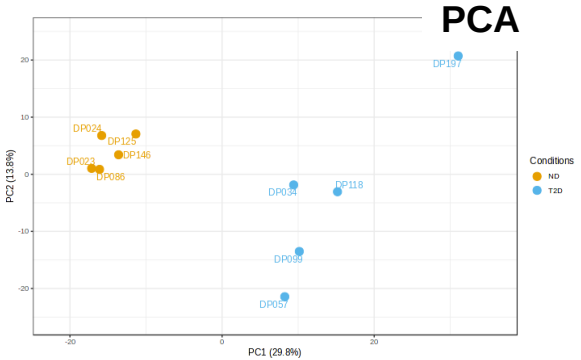
QC => Missing Values => Data Filtering
=> Data Normalization

Understanding QC plots in ExpressAnalyst

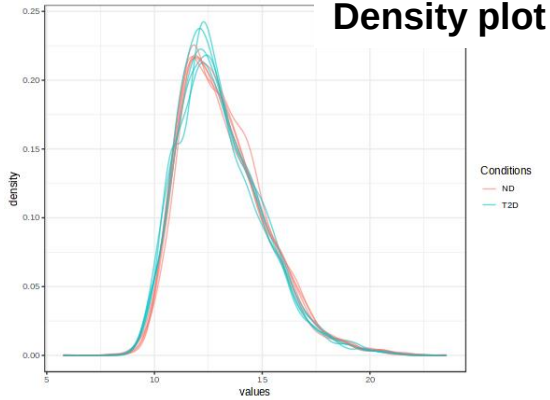
Box plot



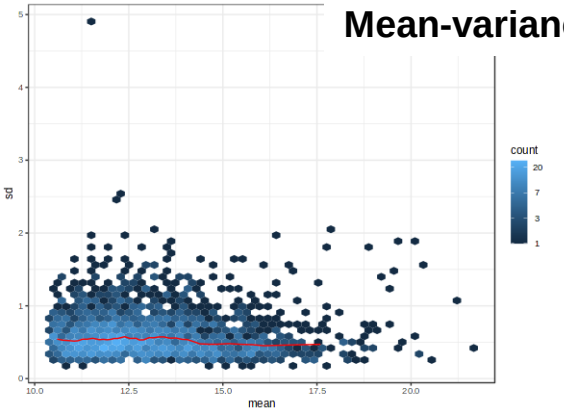
PCA



Density plot

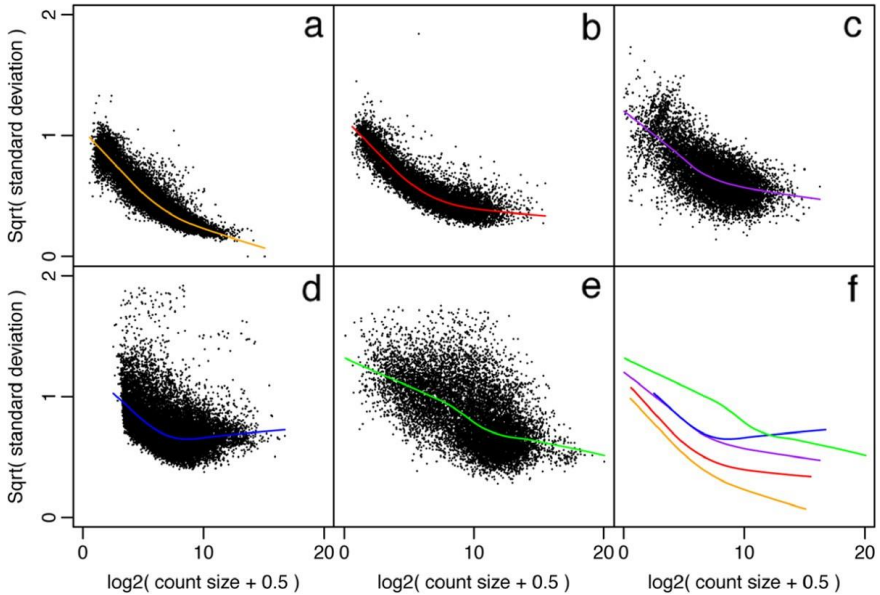


Mean-variance plot



QC based on mean-variance plot

- Showing a decreasing trend between the means and variances resulting from a combination of technical variation in the sequencing experiment and biological variation amongst replicates.
  - Experiments with high biological variation usually result in flatter trends, where variance values plateau at high expression values.
  - Experiments with low biological variation tend to result in sharp decreasing trends.

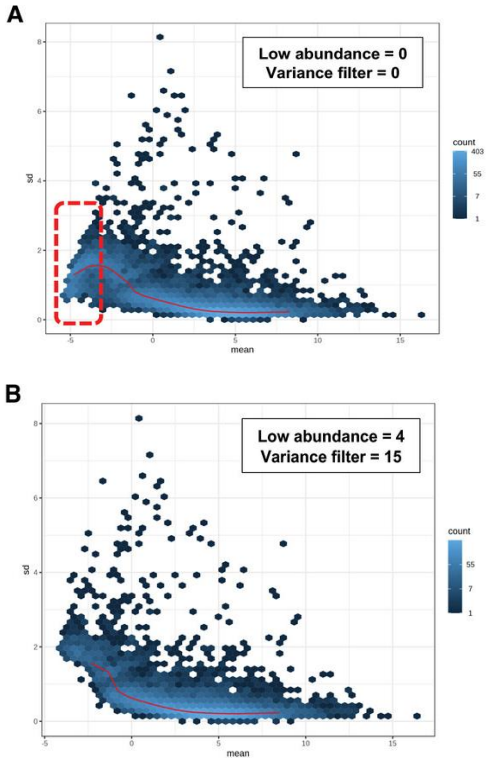


Ordered by increasing levels of biological variation in datasets

https://pubmed.ncbi.nlm.nih.gov/24485249/

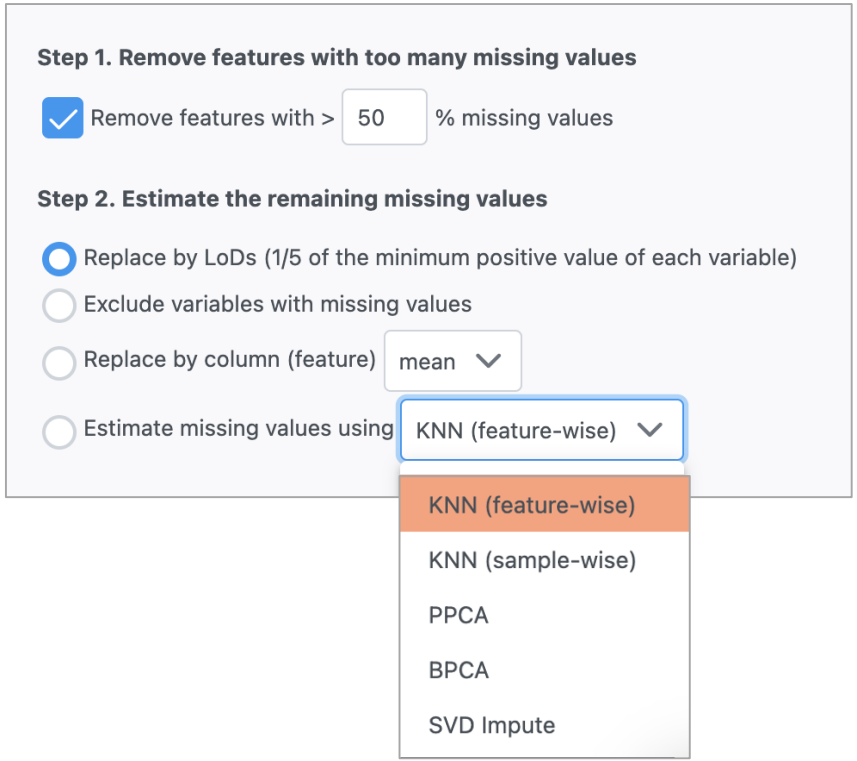
Mean-variance (MA) plot for filtering

- Visual check on the level of filtering performed upstream.
- If filtering of lowly-expressed genes is insufficient, a drop in variance levels can be observed at the low end of the expression scale due to very small counts.
- If this is observed, return to the earlier filtering step and increase the expression threshold applied to the dataset.



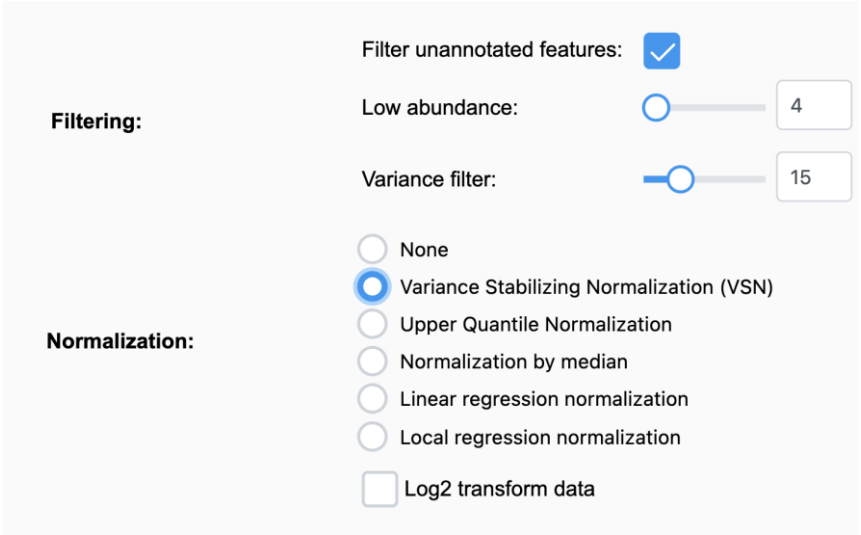
Dealing with missing data

- Missing at random: missingness can be fully accounted for by variables where there is complete information
  - Predictable using other features of the data
- Missing not at random: where there is a systematic explanation
  - Below detection limit in MS-based proteomics or metabolomics.



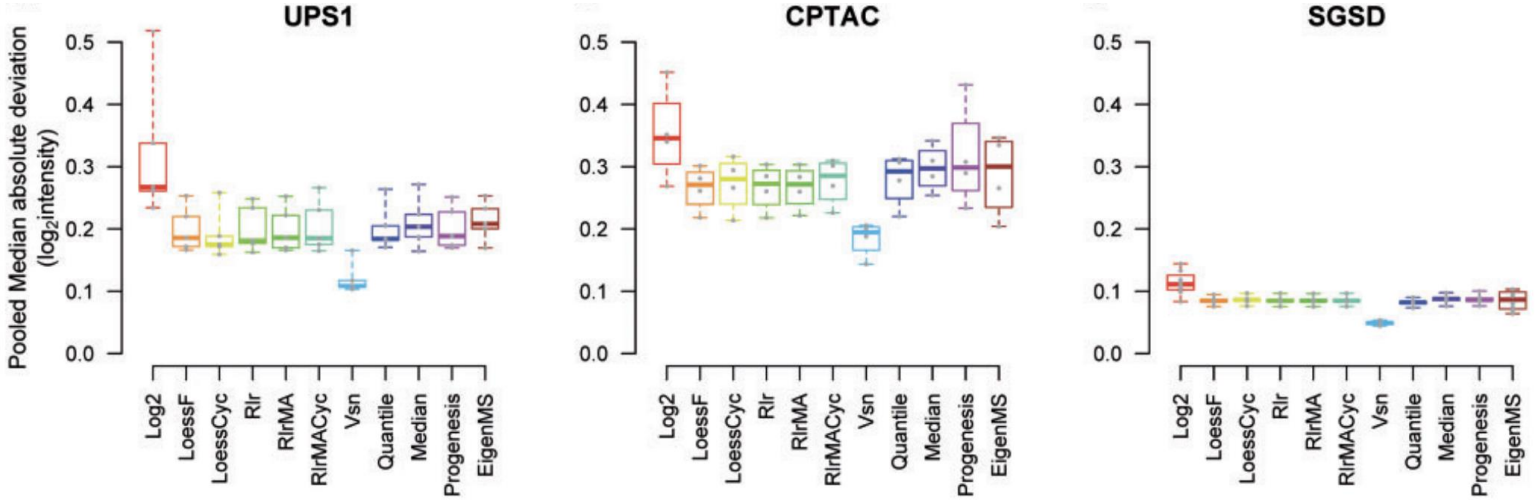
Variance stabilization normalization (VSN)

- Originally designed for microarray to overcome the limitations of log transformations by accommodating negative values and minimizing the inflated variance around low signal intensities
- Scale data from different samples into the same level through parametric transformations and maximum likelihood estimation
- Eliminate the dependency between variances and mean abundances



MS-based proteomics data normalization

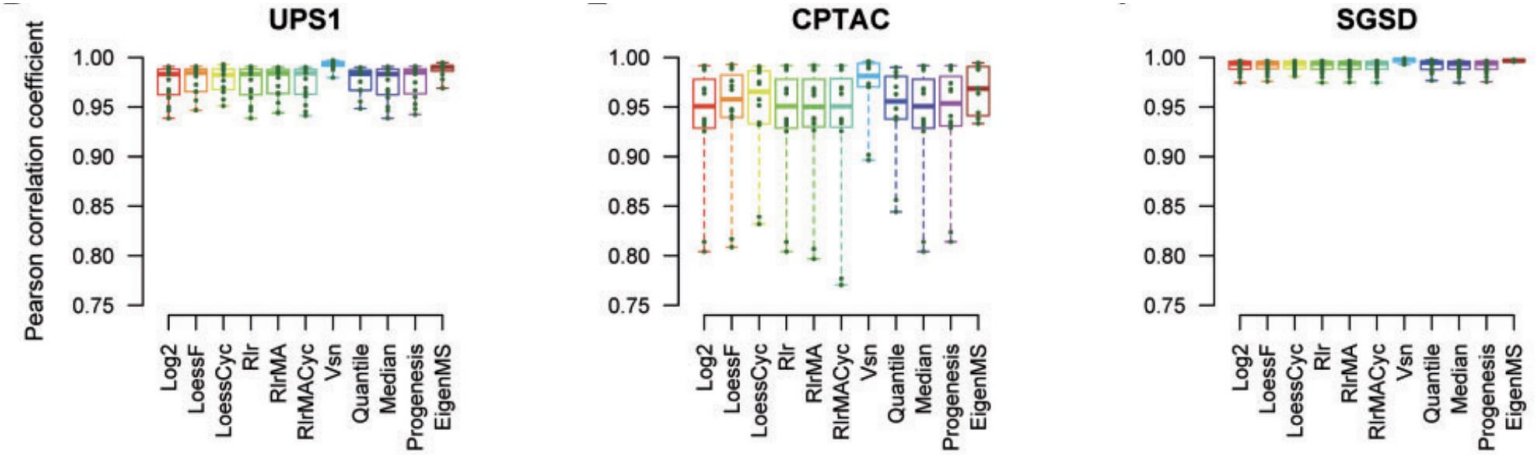
Variance among house keeping genes
(smaller => better)



https://pubmed.ncbi.nlm.nih.gov/27694351/

MS-based proteomics data normalization

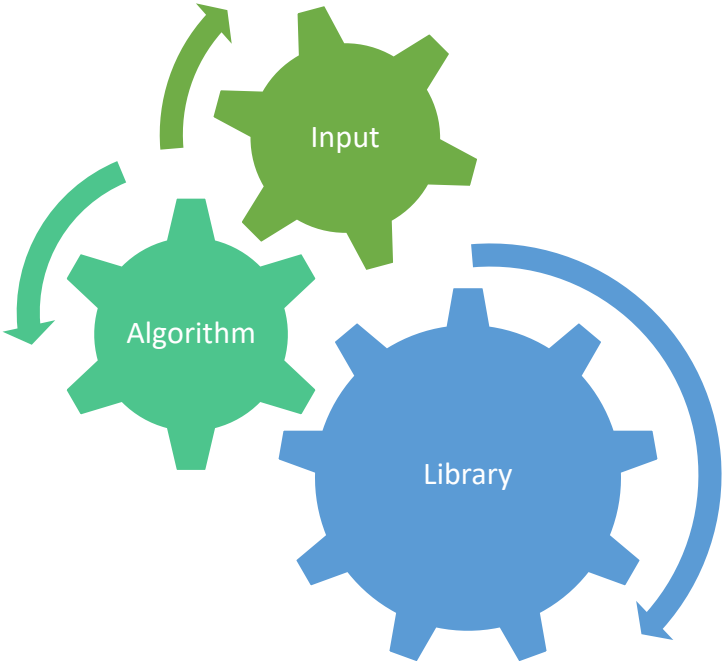
Pearson correlation coefficients among house keeping genes (higher => better)



https://pubmed.ncbi.nlm.nih.gov/27694351/

Enrichment analysis

Three components in enrichment analysis



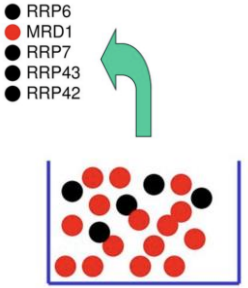
1. Input: define the signals of interest
   - **Significant features**
   - Complete ranked list
   - A data table
2. Define functions libraries:
   - Pathways, gene sets
3. Perform enrichment tests evaluate the coordinated changes of a group.
   - Over-representation analysis (ORA)
   - Permutation based approaches

Enrichment Tests

We need to calculate the null distribution - how often we can see this by random chance?

- Model driven
  - Hypergeometric distribution
- Data driven
  - Don’t know - we can use permutation to get this

$$f_X(k|N,K,n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



	Sampled	Not Sampled	Total
success	k	K-k	K
non-success	n-k	(N-K)-(n-k)	N-K
Total	n	N-n	N

Over-representation analysis (ORA)

Potential sources of bias with ORA

ORA assumption: all genes/proteins can be measured with equal reliability or opportunity. However, the realities are:

1. Technological bias
   - Platform (MS or microarray) capture a limited subset of “gene/protein universe”
2. Biological bias
   - A given cell type or tissue have specialized proteome / transcriptome
3. Sampling bias
   - Could be introduced during sample prep (selectively enrich some groups)

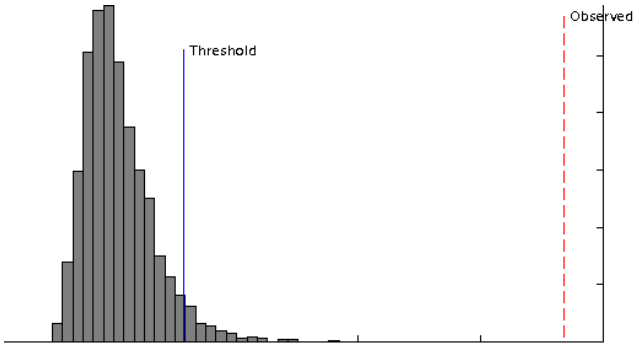
Addressing bias (I) – providing background

- Better define the “background universe” reflecting the technical and biological context of the current study
  - This information is provided in data table upload
    - i.e. all genes / proteins measured are already present in the raw abundance table)
  - For gene/protein list upload, you can optionally upload the complete list

Addressing bias (II) – using permutations

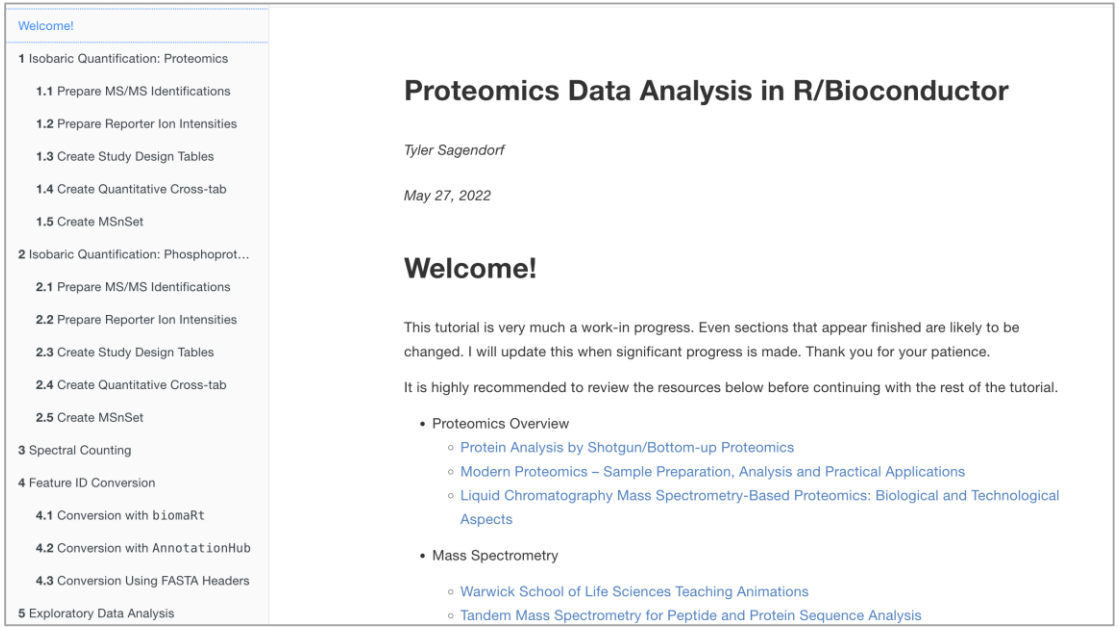
To estimate null distribution from the current data

- Transcriptomics / proteomics
  - GSEA
- miRNA-omics
  - Unbiased sampling (introduced in the previous lecture)
- Metabolomics
  - Mummichog
    - will be introduced in the next session



Learning more on computational proteomics

❖ Computational Mass Spectrometry @ Pacific Northwest National Laboratory



https://pnnl-comp-mass-spec.github.io/proteomics-data-analysis-tutorial/index.html

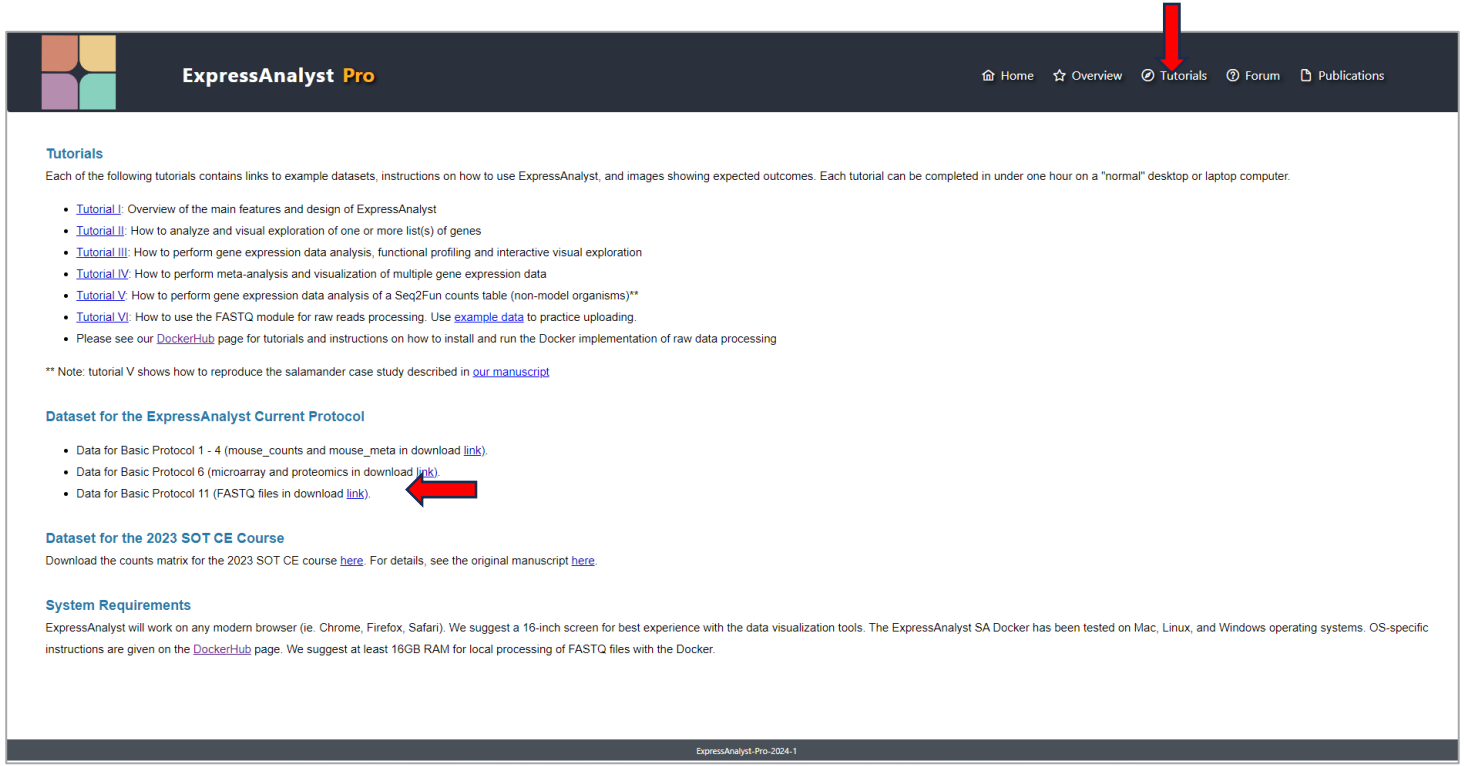
Live Demo & Hands On

Proteomics vs RNA-seq data

- Proteomics data:
  - LC-MS
  - Intensity based, has decimal values.
  - Lower coverage (a subset of whole proteome)
  - Often have missing values

- RNA-seq data
  - High-throughput sequencing
  - Count based, integer
    - In some cases, can be decimal in the case of fractional counts (depending of the algorithm used)
  - Capture most genes
  - Do not have missing values

Example dataset

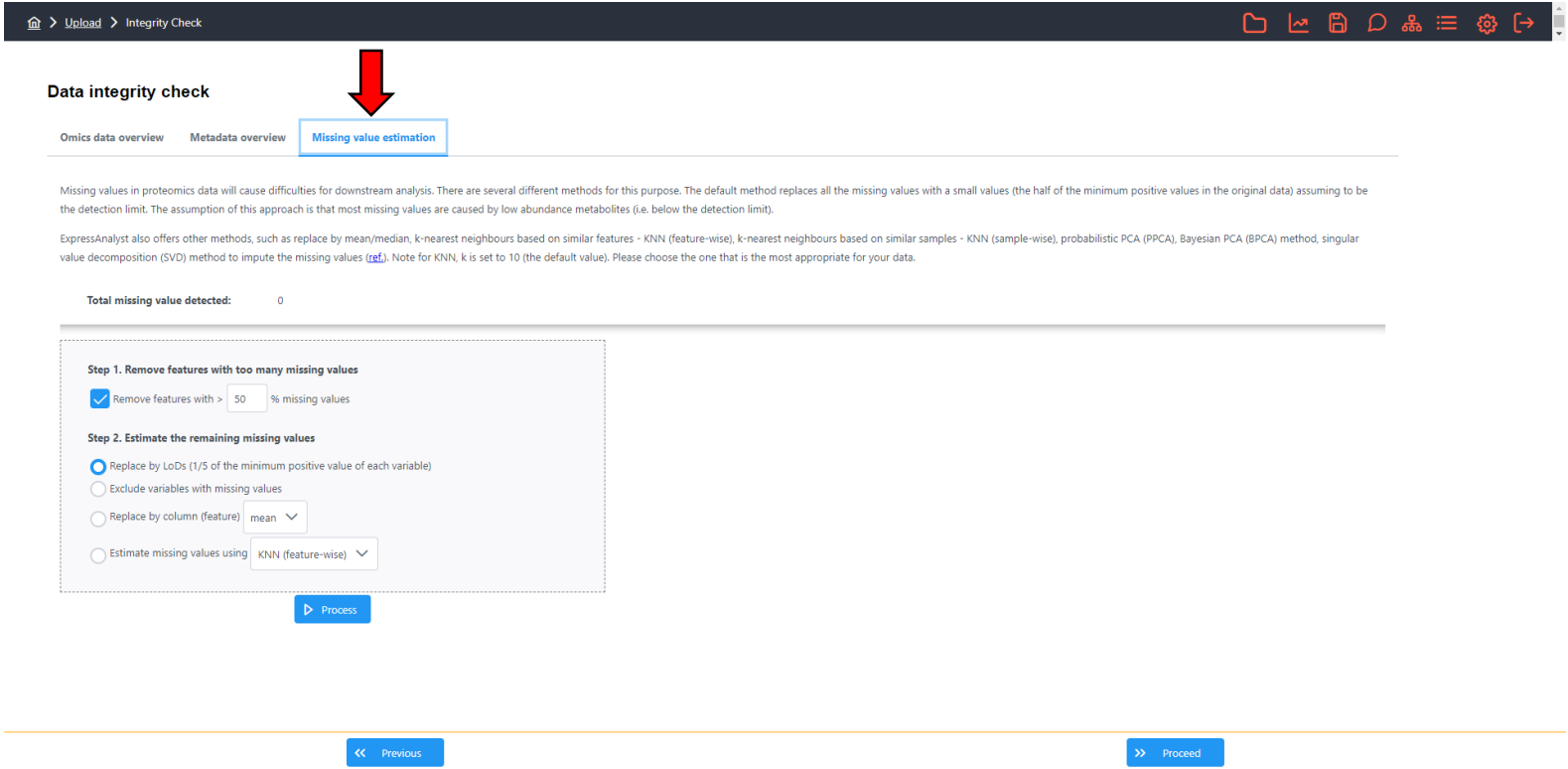


Example dataset

#NAME	D034	D057	D086	D099	D118	D125	D146	D197	D023	D024
#CLASS: Diagnosis	T2D	T2D	ND	T2D	T2D	ND	ND	T2D	ND	ND
CRYAB	12.08658823	12.64878448	12.28710506	NA	13.29318423	12.50848808	13.41511448	13.54087129	NA	11.80891996
HDLBP	16.00951903	15.85587266	16.31787505	15.38619931	15.61725193	15.78135854	15.8326181	15.08939838	16.26661951	15.76832481
EHD1	11.82545124	13.4000789	12.83612759	13.29662656	12.53035756	12.64786312	13.16280041	12.98266615	13.14742926	12.90561264
SERPINA1	NA	NA	14.35252462	NA	13.36863623	13.27338767	NA	NA	14.01216748	12.25590661

Proteomics data often have missing data (LC-MS）.
In RNA-seq, they would be simply 0 count.

Missing value imputation



Missing value imputation: assumptions

- Missing due to the abundance level is lower than detection limit:
  - Replace by 1/5 of the lowest value for that variable --> default option in ExpressAnalyst

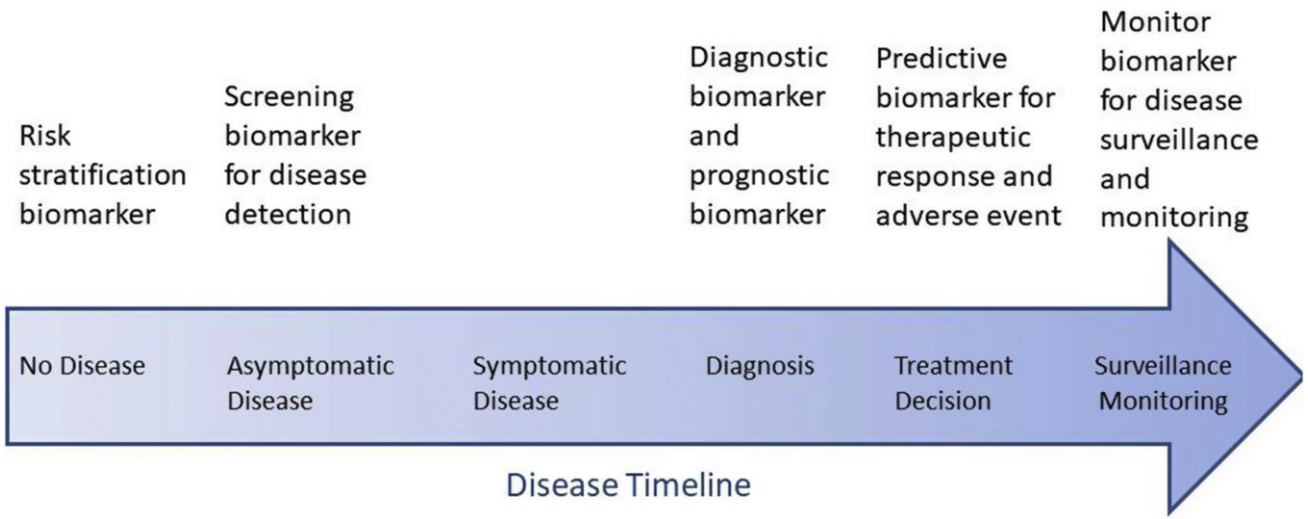
- Missing due to technical or human error:
  - Replace by mean/median

Schedule for today

Time	Topics
9:00 – 9:10	General introduction & recap
9:10 – 9:40	Proteomics data analysis workflow
9:45 – 10.15	ExpressAnalyst live demo & hands on
10:20 – 10:50	Biomarker analysis
10:55 – 11:25	MetaboAnalyst live demo & hands on
Summary and discussion	

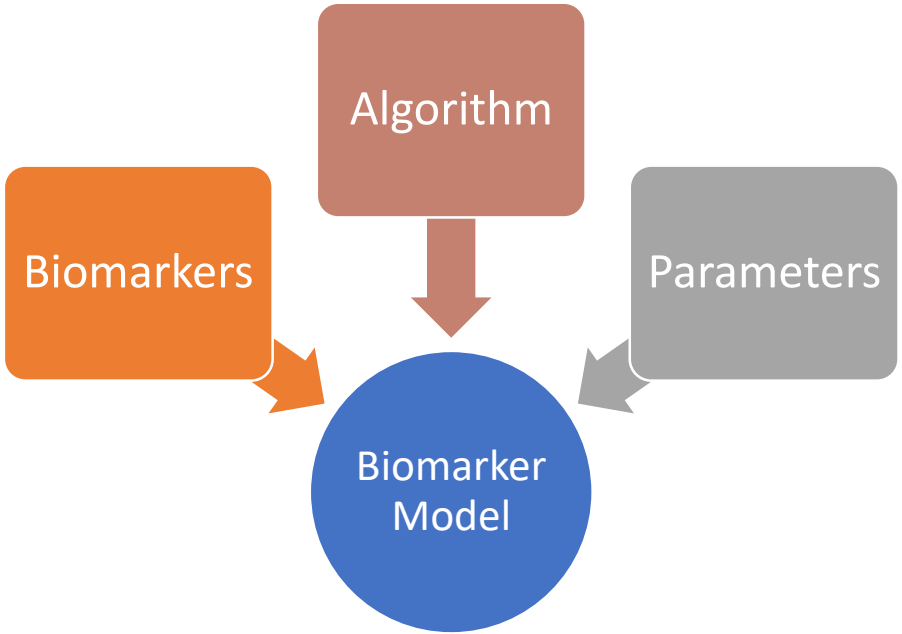
What are Biomarkers

A characteristic that is **objectively measured** and **evaluated** as an indicator of normal biological processes, pathogenic processes or pharmacological responses to a therapeutic indication.



Key components in biomarker analysis

- The main activities in biomarker analysis involve selecting biomarkers and tuning parameters based on a few well-established classification algorithms.
- Evaluating performance of biomarker models is based on their capacities to classify new samples using cross validation (CV).
- Permutation tests are often used to evaluate whether a classifier has learned anything better than random guessing (null model).



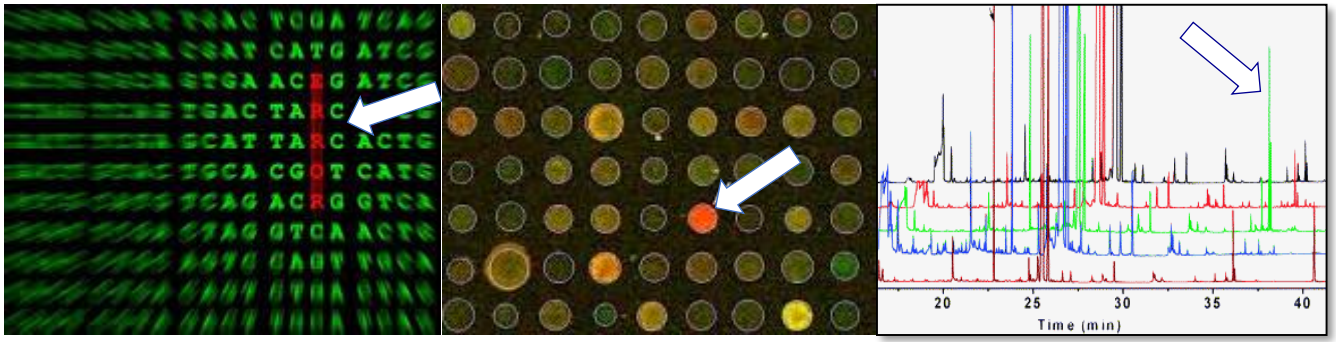
Overview of biomarker analysis

1. Biomarker selection
   - ✓ Identification of an optimal subset of features that will provide the maximal discriminating power between the diseased and healthy samples
2. Performance evaluation
   - ✓ Assessment & validation of the panel of biomarkers
3. Model creation
   - ✓ Developing a fixed mathematical equation or computer algorithm, which combines the panel of selected biomarkers into a single test score with the aim of accurately predicting a particular clinical outcome,

Measuring biomarkers

We look for biomarkers that

- Can be measured earlier
- Can be measured more easily or frequently
- Can be measured with higher precision, or less subjective



Genetic Variation

Gene Expression

Metabolite Concentration

Desirable properties of biomarkers

- Objectively measurable
  - Can be reliably measured
  - By different platforms or technologies
  - Clear identifications
  - Absolute quantification
  - Interpretable?
- High sensitivity, high specificity

Convenient & accurate

Evaluating Performances

Balanced data (control group and disease group ~same size)

- **Accuracy**
  - 9/13 correct => 69% accuracy
- **Error rate**
  - 1 – accuracy => 31%

Not suitable for imbalanced data

- In a population, cancer incidence is low: ~5 cases in 1000 people. If a classifier predict all people to be healthy, then it is 99.5% accurate (majority vote)
  - Need to develop other measurements

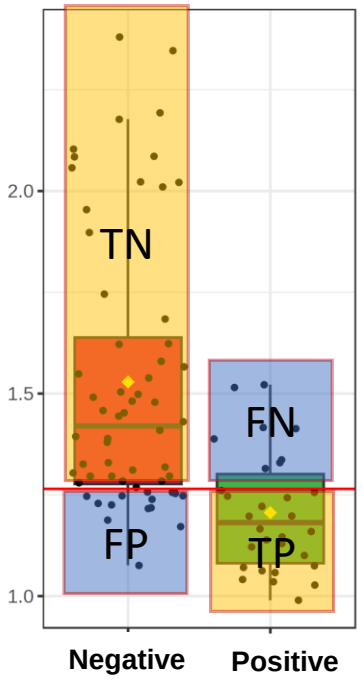
Evaluating Biomarker Performance

- Performance evaluation in clinical diagnosis (usually very unbalanced i.e., few case, most normal)
  - True positives (TP)
  - True negatives (TN)
  - False positives (FP)
  - False negatives (FN)

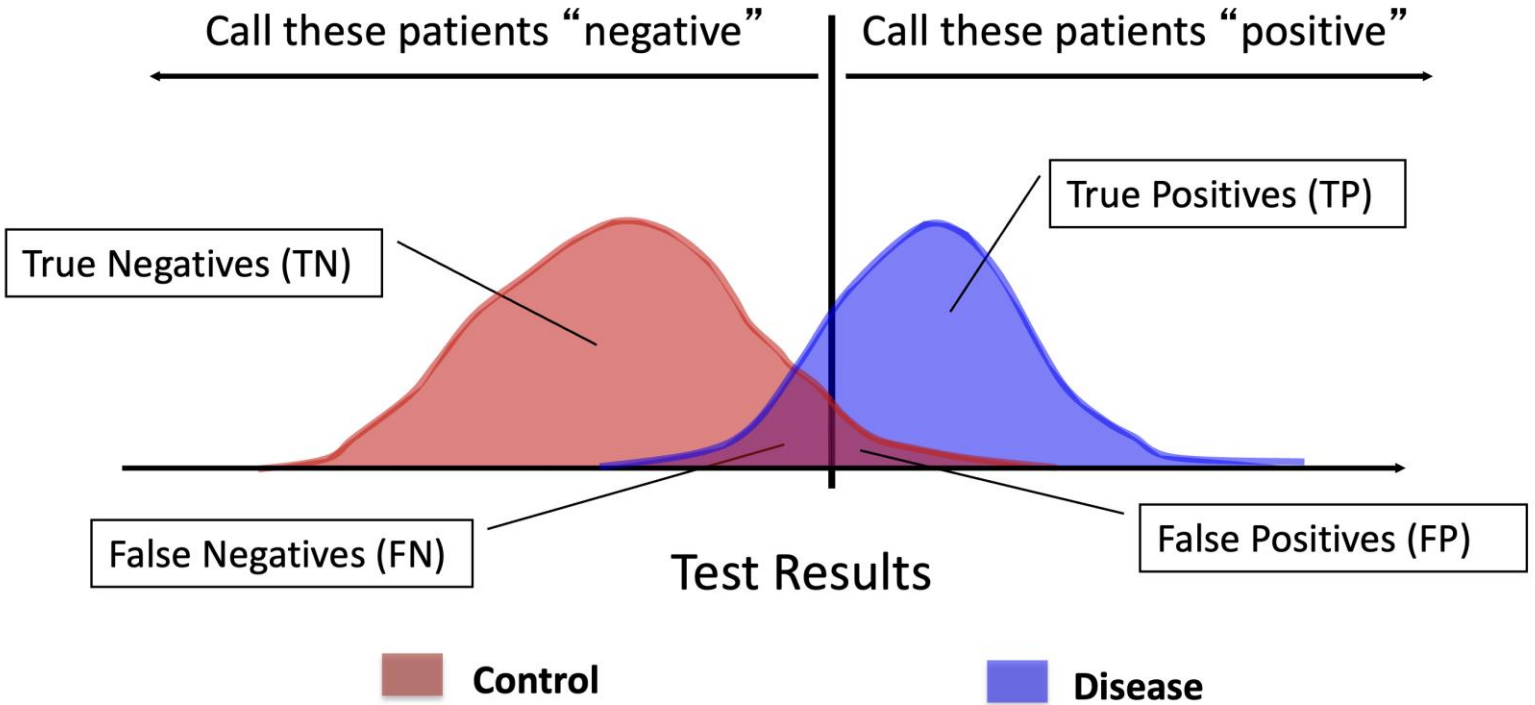
	p' (Predicted)	n' (Predicted)
P (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Evaluating Biomarkers - Sensitivity & Specificity

- Sensitivity: true positive rate
  - Sn = TP / (TP + FN)
  - The probability of a positive test result given that a subject has an actual positive outcome
- Specificity: true negative rate
  - Sp = TN/(TN + FP)
  - The probability of a negative test result given that a subject has an actual negative outcome

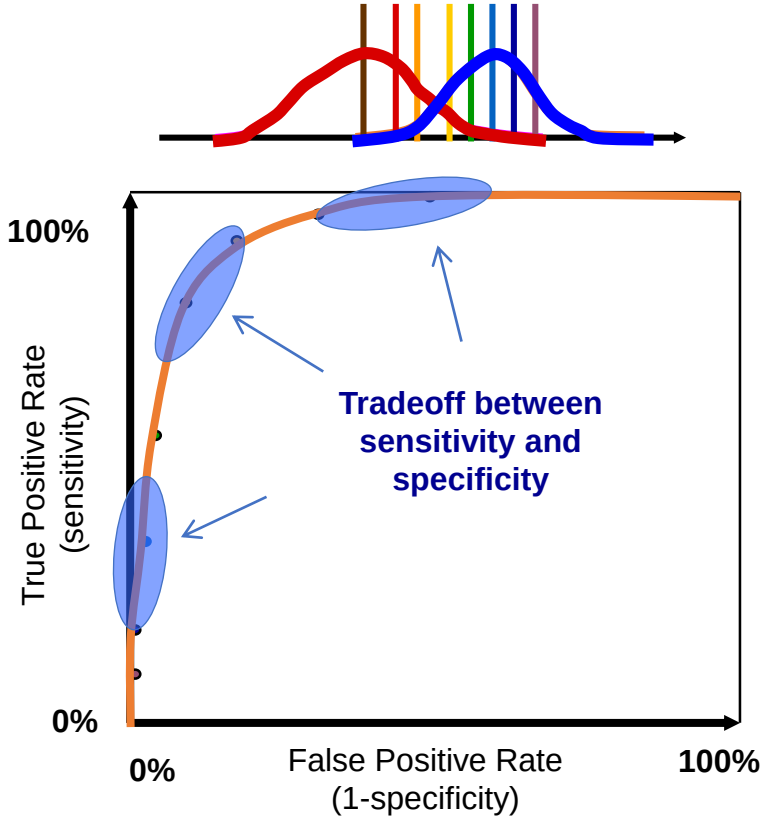


Another illustration



Receiver Operating Characteristic (ROC) curves

- A historic name from radar studies
- Very popular in biomedical applications
  - To assess biomarker performance.
  - To compare different biomarker models
- A graphical plot of the true positive rate (TPR) vs. false positive rate (FPR), for a binary classifier (i.e. positive/negative) as its cutoff point is varied



Sensitivity, Specificity & ROC curve

Two important performance measures in a diagnostic tests

- Sensitivity (true positive rate)
- Specificity (true negative rate)

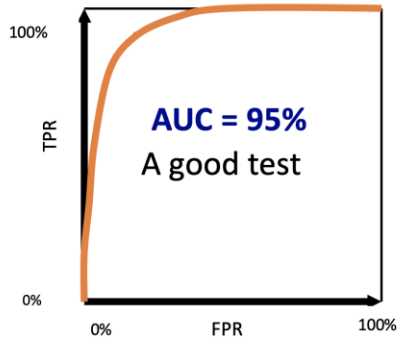
Cutoff dependent

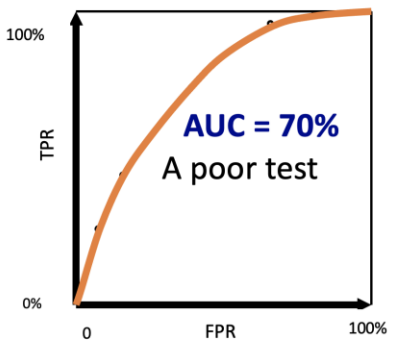
- Increase cutoff, will improve specificity, decease sensitivity

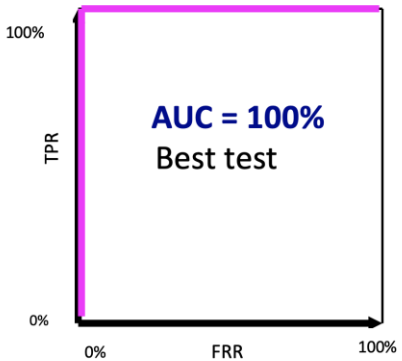
ROC curves integrate these two measures

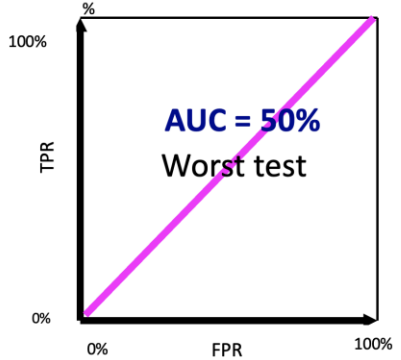
Area under an ROC curve (AUC)

- Overall measure of a test performance
- AUC is the probability that a classifier will rank a randomly chosen positive case higher than a randomly chosen negative one
- Comparisons between two tests based on differences between (estimated) AUCs









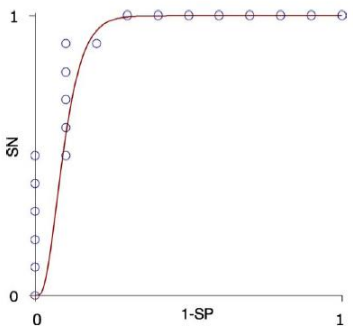
From Biomarkers to ROC curves

How to construct ROC curves

Input: a score on a univariate scale

- A test gives continuous value (i.e. blood *Glucose* level)

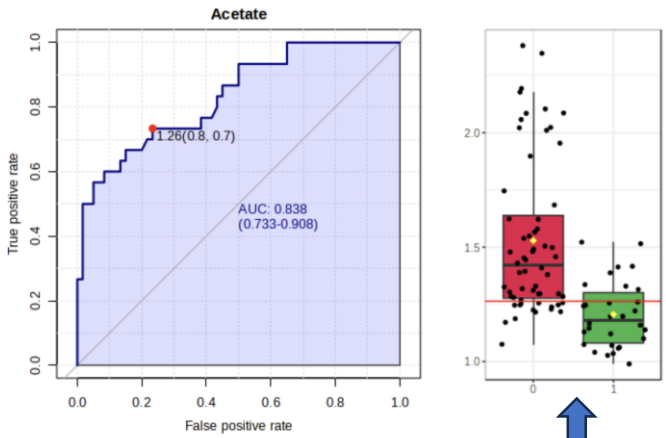
2-h plasma glucose (mmol/L)		Threshold-dependent values*					
Healthy	Diseased	*TP*	*TN*	*FP*	*FN*	1 – *SP*	*SN*
		10	0	10	0	1.00	1.00
4.86		10	1	9	0	0.90	1.00
5.69		10	2	8	0	0.80	1.00
6.01		10	3	7	0	0.70	1.00
6.06		10	4	6	0	0.60	1.00
6.27		10	5	5	0	0.50	1.00
6.37		10	6	4	0	0.40	1.00
6.55		10	7	3	0	0.30	1.00
7.29	7.29	9	8	2	1	0.20	0.90
7.82		9	9	1	1	0.10	0.90
	9.22	8	9	1	2	0.10	0.80
	9.79	7	9	1	3	0.10	0.70
	11.28	6	9	1	4	0.10	0.60
	11.83	5	9	1	5	0.10	0.50
12.06		5	10	0	5	0.00	0.50
	18.48	4	10	0	6	0.00	0.40
	18.50	3	10	0	7	0.00	0.30
	20.49	2	10	0	8	0.00	0.20
	22.66	1	10	0	9	0.00	0.10
	26.01	0	10	0	10	0.00	0.00



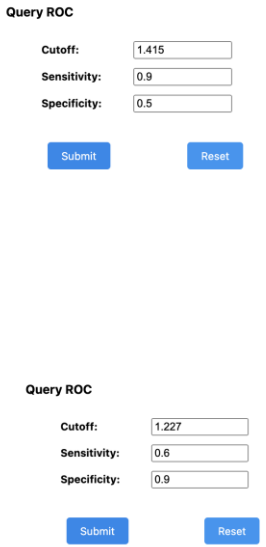
TA Lasko, et al (2005)

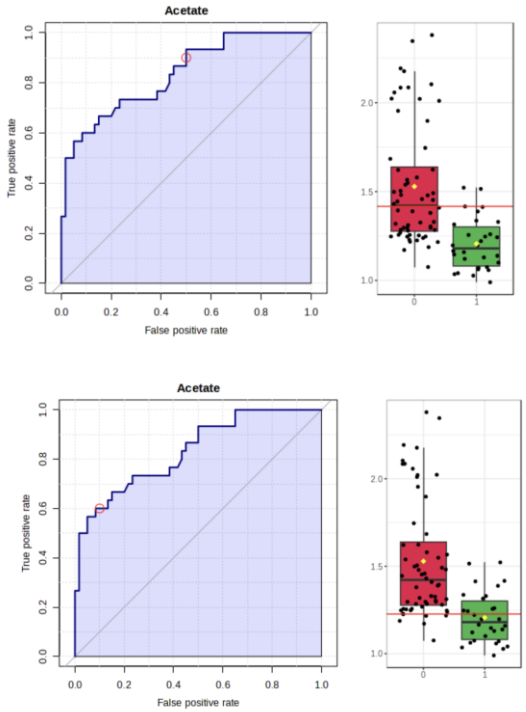
Straightforward for a single biomarker

Abundance values can be used directly to predict Positive/Negative



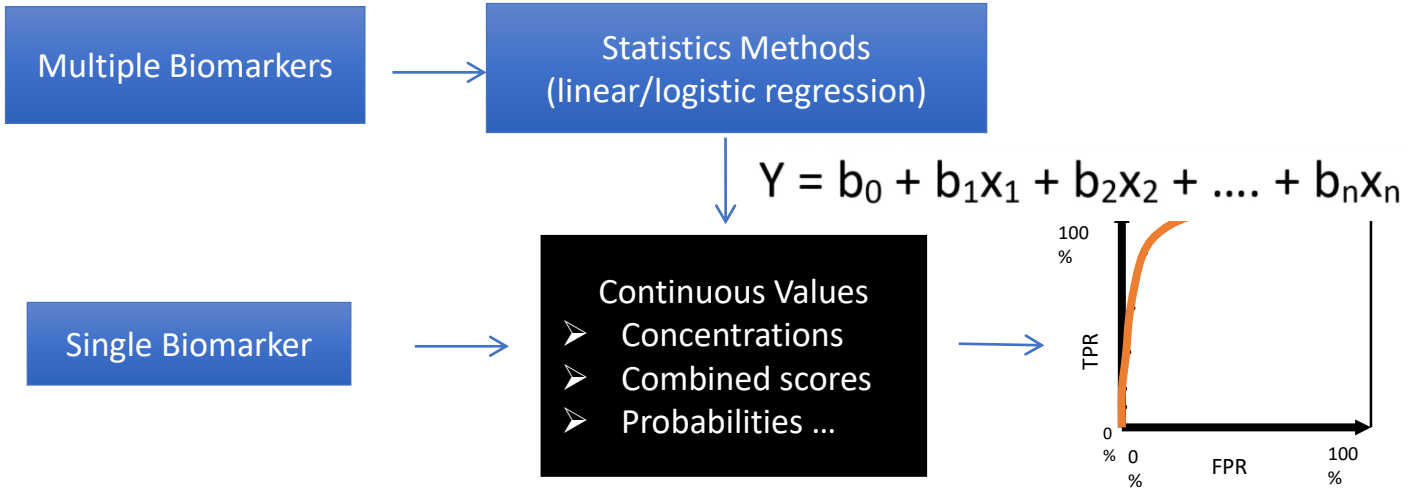
Move the cut-off (red line)





How to deal with multiple biomarkers?

- Input: a score on a univariate scale
  - A classifier that produces a continuous score (i.e. likelihood, probabilities)



The challenges of Omics Data

Omics data is usually of small sample size with large number of variables (n << p)

- SNPs (~ 1,000,000s)
- Gene expressions (~10,000s)
- Compound concentrations (~ 1000s)

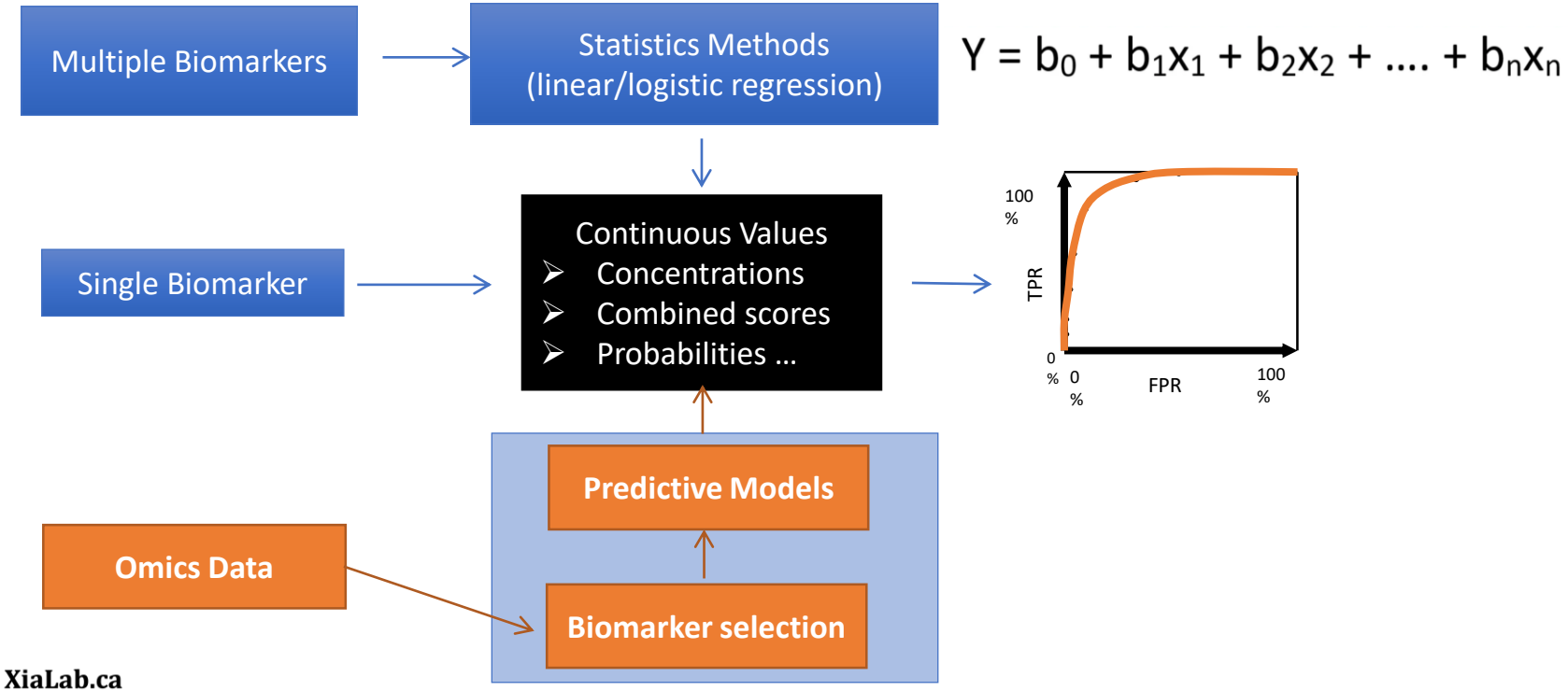
➔ Linear/logistic regression will not work (n > p)

Need new strategies

- Computationally efficient;
- Not susceptible to over-fitting;

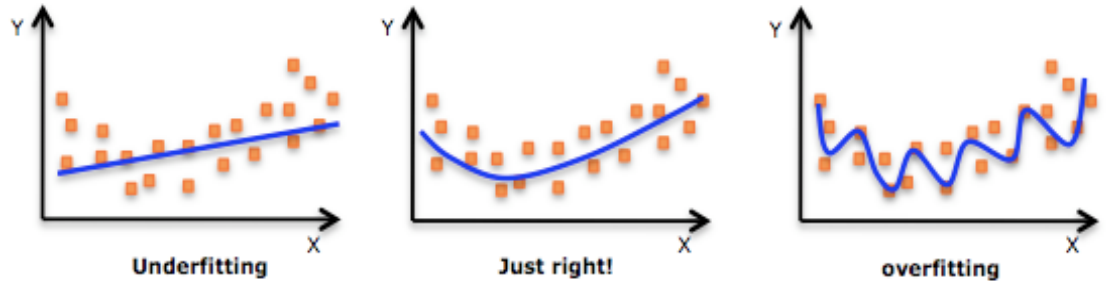
➔ Advanced machine learning approaches

Dealing with omics data



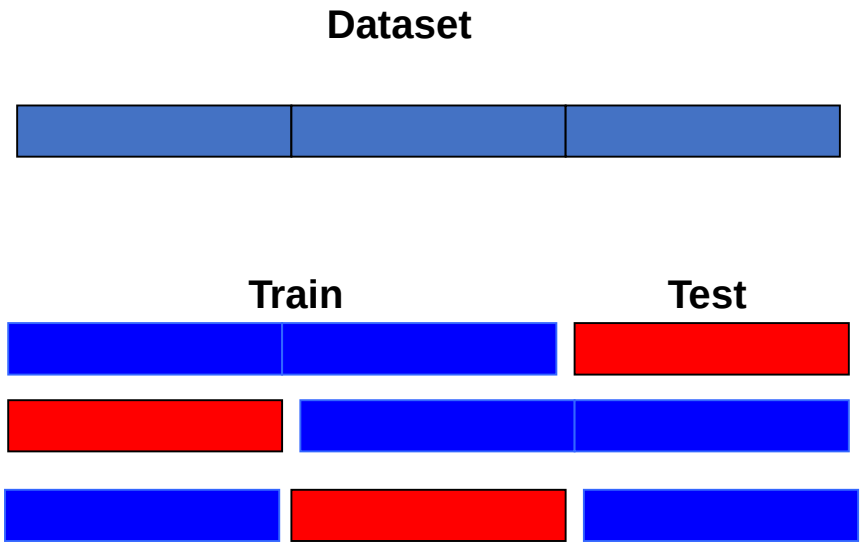
Overfitting issue

- Fitted model performs well for the current data
- Fitted model is not good for prediction of new data – prediction error is underestimated
- Model is too elaborate, models “noise” that will not be the same for new data



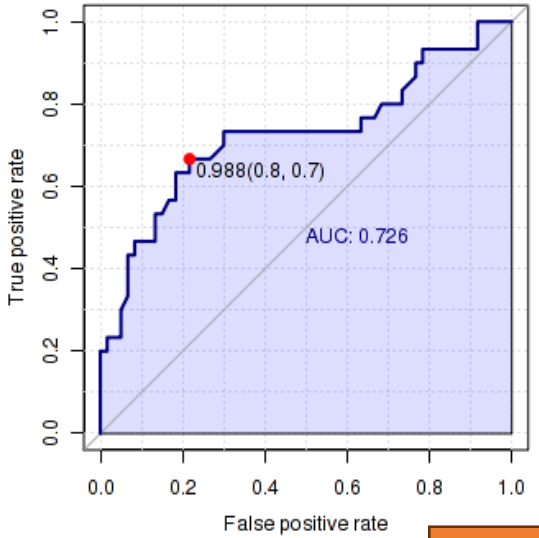
Addressing overfitting issue

- **Cross validation** – whether the model can predict on new events
  - Prediction accuracy
  - Sum of squares captured by the model (R^2)
  - Cross-validated R^2 (also known as Q^2)
- **Permutation tests** – whether the model captures real signals compared to null

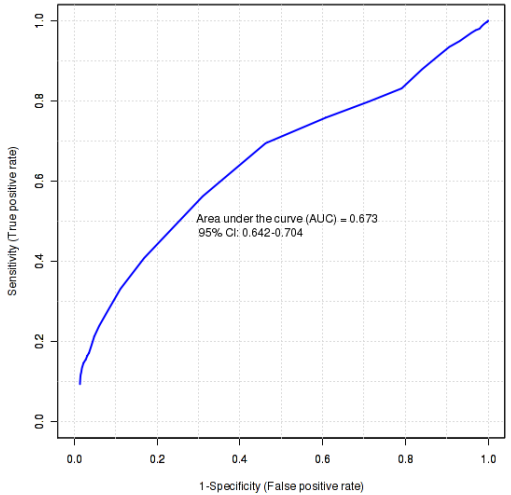


Over-fitting in classical univariate ROC curve

AUC Classical approach: 0.726



AUC CV-based: 0.673

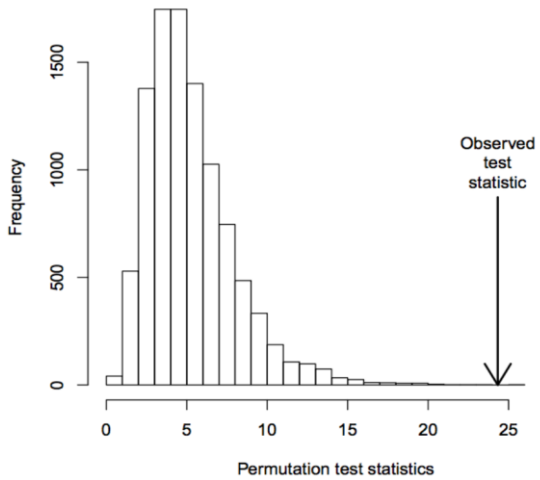


Classical univariate based on **all** data points
(no separation of training & testing)

Permutation Tests

To test whether your model is significantly different from the null models

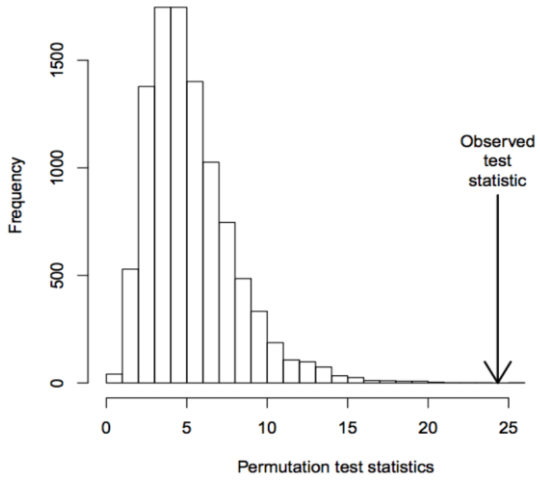
1. Randomly shuffle the class labels (y) and build the (null) model between new y and x;
2. Test whether there is still the similar performance;
3. We can compute empirical p values
   - If the result is similar as the permuted results (i.e. null model), then we can not say y and x is significantly correlated



Permutations

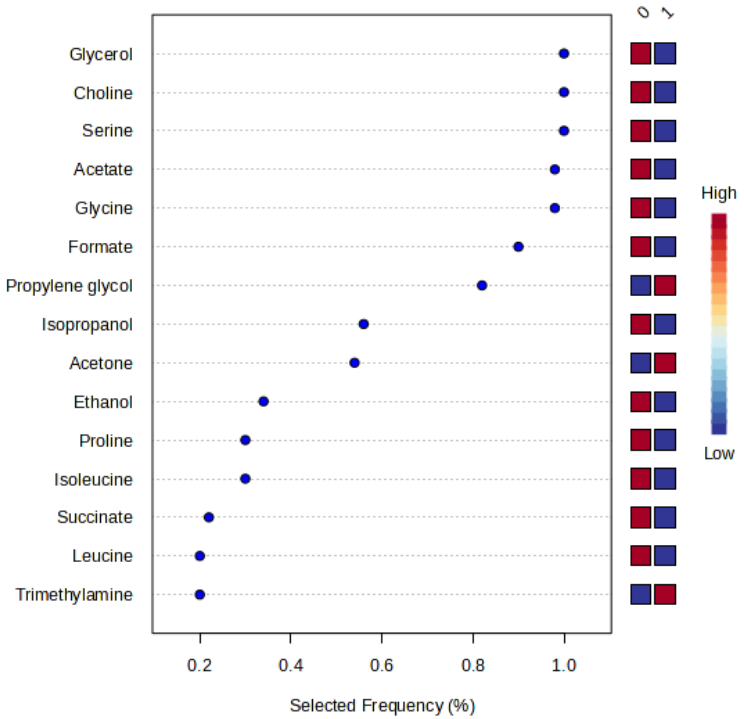
To test whether your model is significantly different from the null models

1. Randomly shuffle the class labels (y) and build the (null) model between new y and x;
2. Test whether there is still the similar performance;
3. We can compute empirical p values
   - If the result is similar as the permuted results (i.e. null model), then we can not say y and x is significantly correlated



Model instability (variance)

- Multiple biomarker models will be created during cross validation-based subset of datasets
- Each model will use the same algorithm, but “slightly” different ingredients
  - Biomarkers
  - Parameters
- More severe when sample size are small



Biomarker Module in MetaboAnalyst

Biomarker selection & model creation

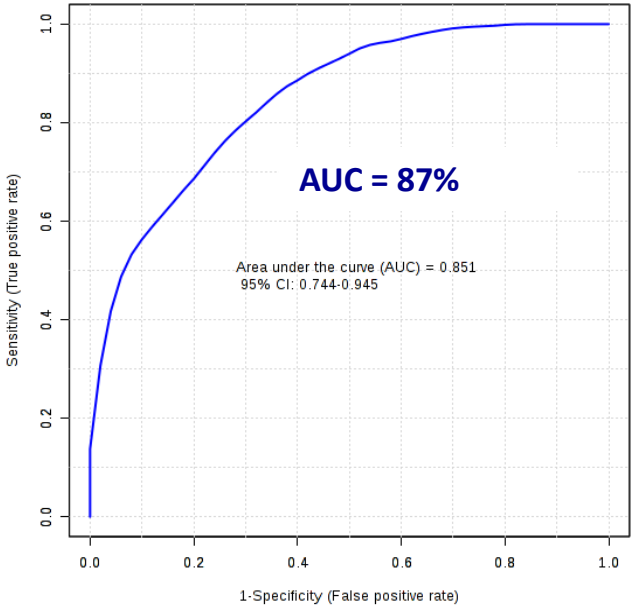
- Support vector machine (SVM);
- Random forests;
- Partial least squares;

Balanced repeated random sampling & cross validation

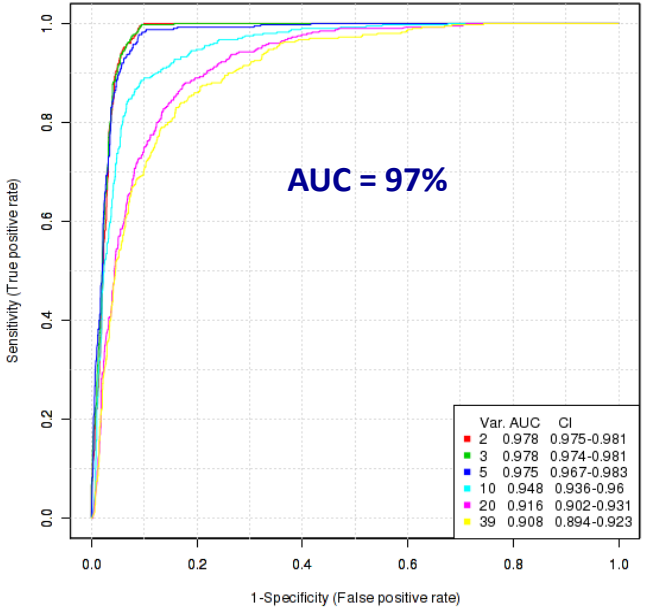
ROC curves are generated by Monte-Carlo cross validation (MCCV) using balanced sub-sampling. In each MCCV, two thirds (2/3) of the samples are used to evaluate the feature importance. The top 2, 3, 5, 10 ...100 (max) important features are then used to build classification models which is validated on the 1/3 the samples that were left out. The procedure were repeated multiple times to calculate the performance and confidence interval of each model.

- Smooth ROC curve
- Confidence intervals
- Imbalanced samples

Manual or automatic biomarker selection



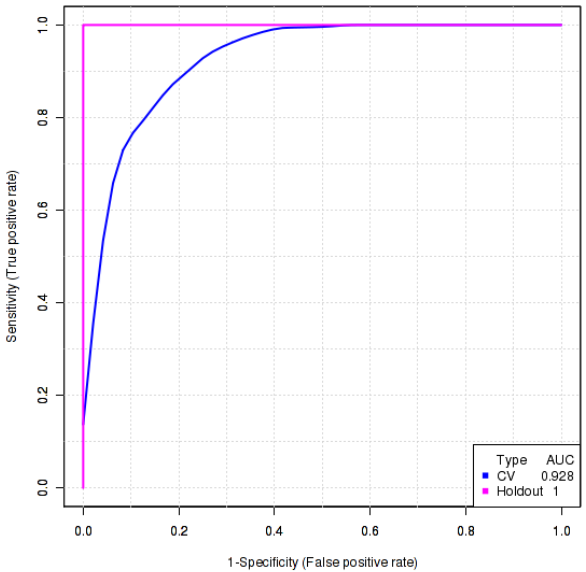
Manual selected biomarkers



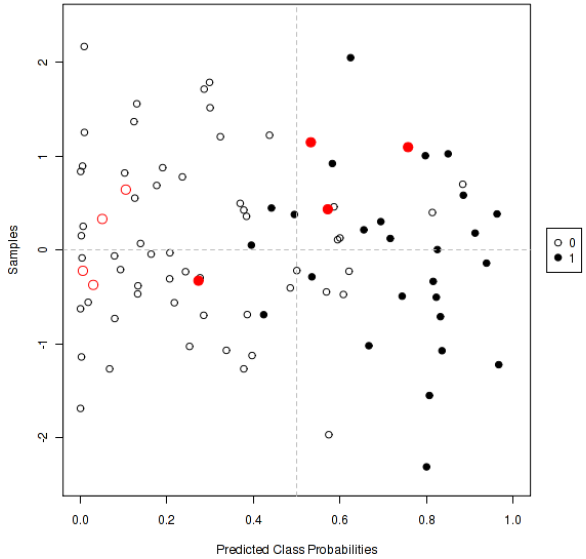
Biomarkers selected by SVM

Predicting new samples with current model

AUC = 1



Accuracy = 7/8



Final Remark

Omics data is usually used for the first-round screening for potential biomarkers

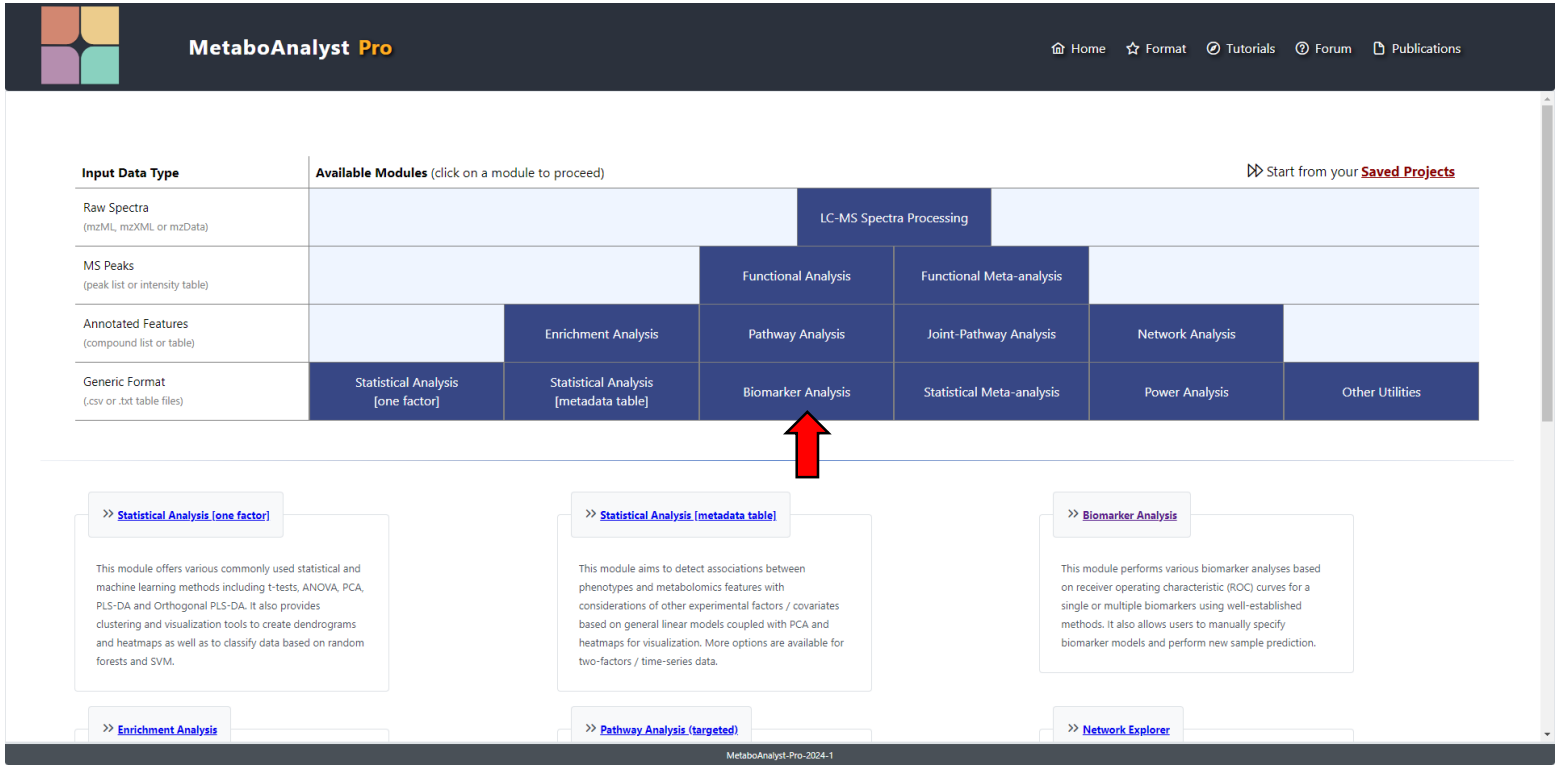
- Need independent validation

Biomarkers are not necessarily mechanistically related to the underlying disease

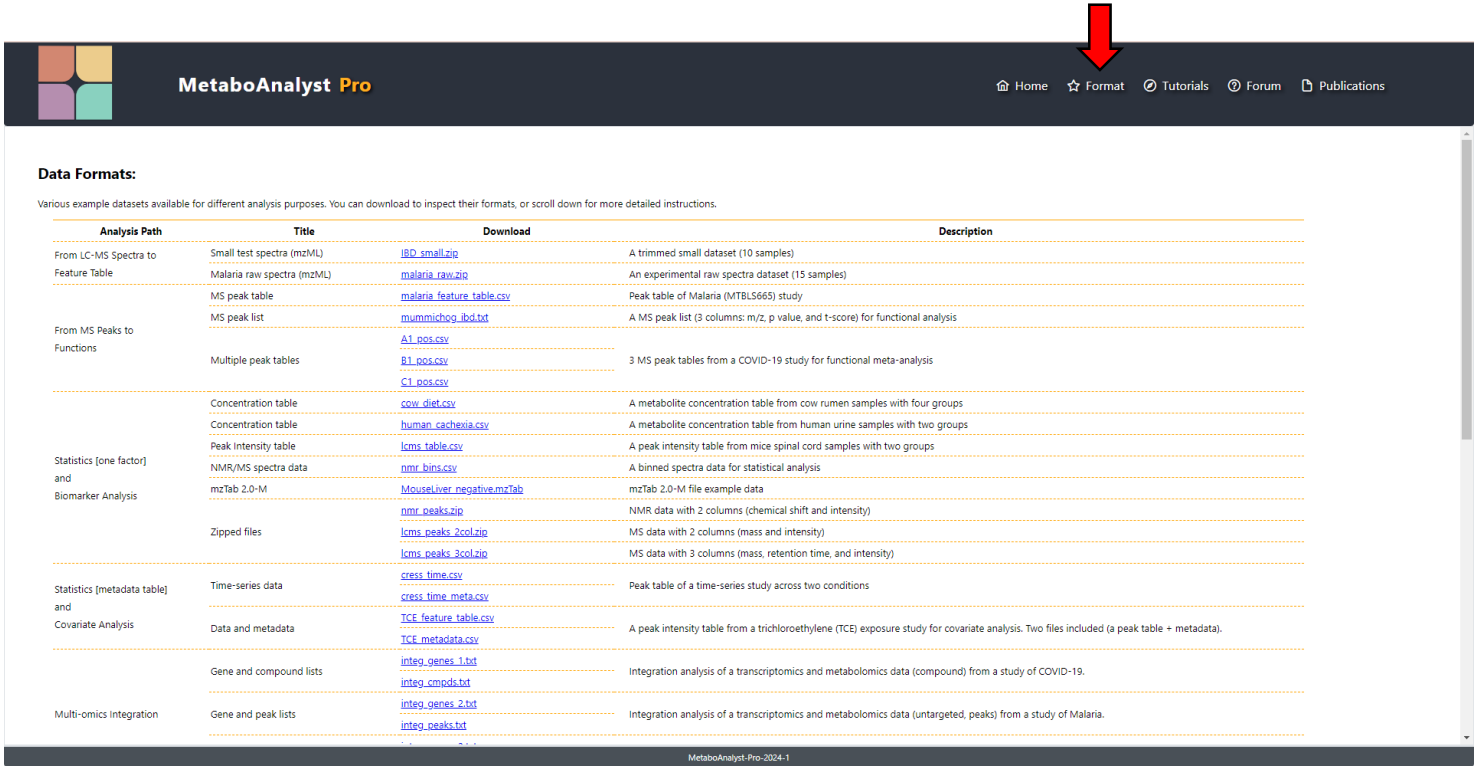
- Association ≠ causality

Live Demo & Hands On

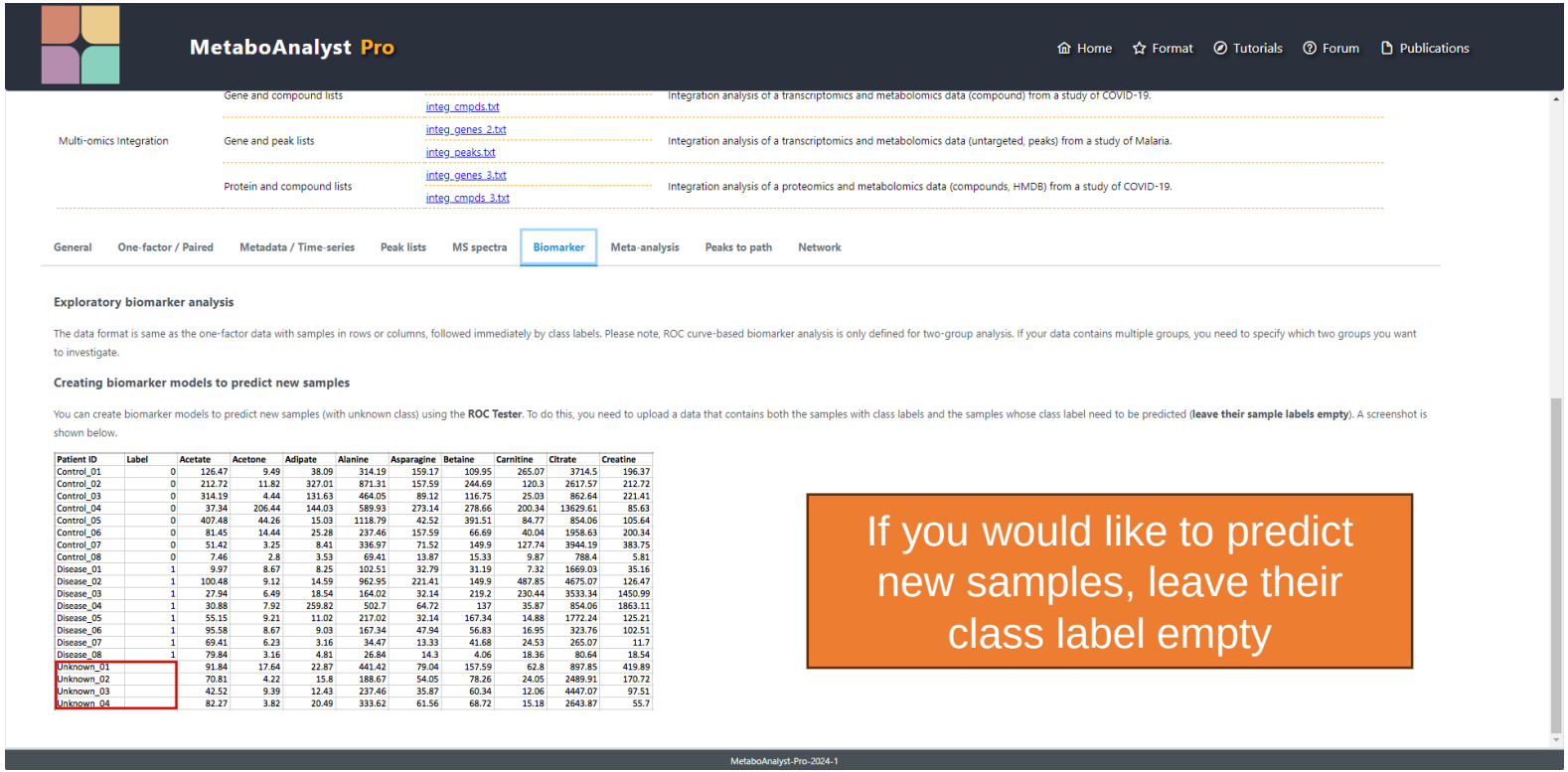
Biomarker Module



Dataset Format



Dataset format



Dataset Format

Samples can be either in rows or columns, first row/column is samples, second row/column should be metadata group. Only one metadata group is supported.

Samples

Metadata

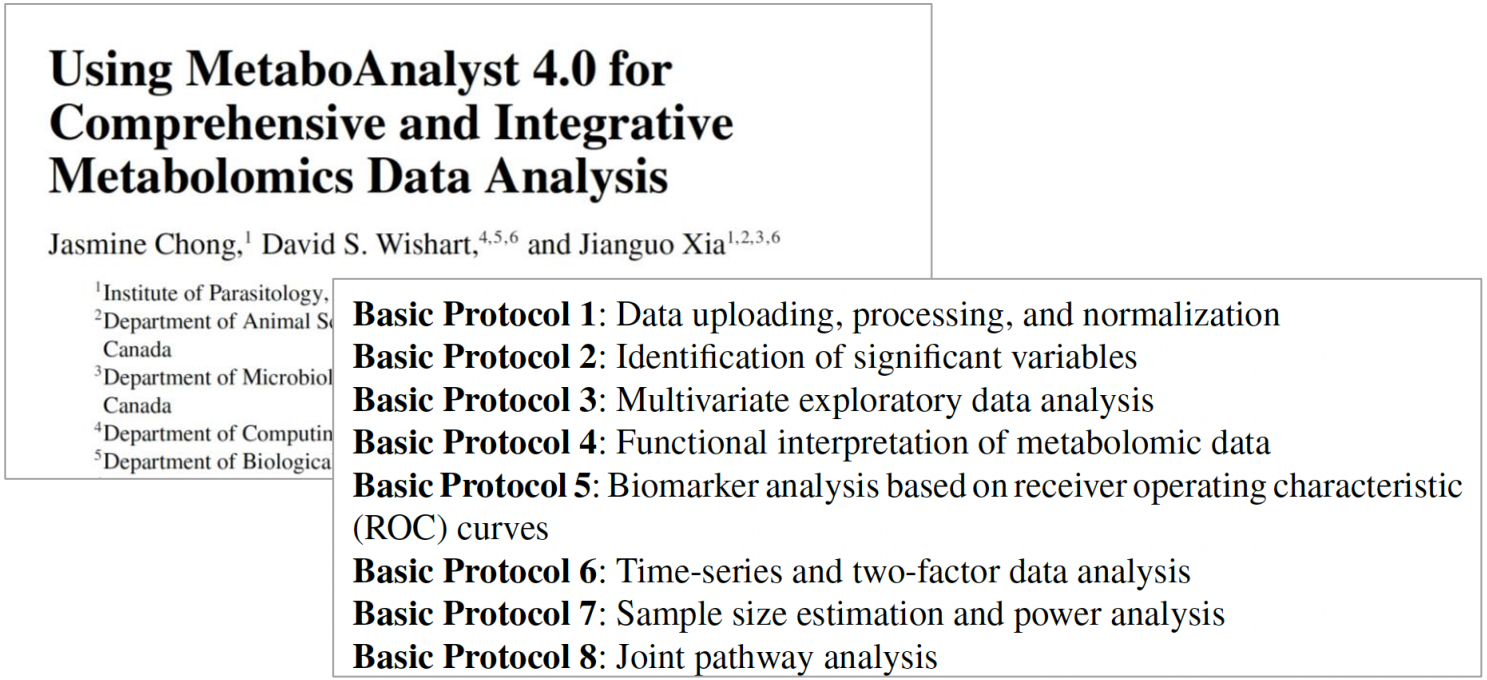
Patient ID	Label	Acetate	Acetone	Adipate	Alanine	Asparagine	Betaine	Carnitine	Citrate	Creatine
Control_01	0	126.47	9.49	38.09	314.19	159.17	109.95	265.07	3714.5	196.37
Control_02	0	212.72	11.82	327.01	871.31	157.59	244.69	120.3	2617.57	212.72
Control_03	0	314.19	4.44	131.63	464.05	89.12	116.75	25.03	862.64	221.41
Control_04	0	37.34	206.44	144.03	589.93	273.14	278.66	200.34	13629.61	85.63
Control_05	0	407.48	44.26	15.03	1118.79	42.52	391.51	84.77	854.06	105.64
Control_06	0	81.45	14.44	25.28	237.46	157.59	66.69	40.04	1958.63	200.34
Control_07	0	51.42	3.25	8.41	336.97	71.52	149.9	127.74	3944.19	383.75
Control_08	0	7.46	2.8	3.53	69.41	13.87	15.33	9.87	788.4	5.81
Disease_01	1	9.97	8.67	8.25	102.51	32.79	31.19	7.32	1669.03	35.16
Disease_02	1	100.48	9.12	14.59	962.95	221.41	149.9	487.85	4675.07	126.47
Disease_03	1	27.94	6.49	18.54	164.02	32.14	219.2	230.44	3533.34	1450.99
Disease_04	1	30.88	7.92	259.82	502.7	64.72	137	35.87	854.06	1863.11
Disease_05	1	55.15	9.21	11.02	217.02	32.14	167.34	14.88	1772.24	125.21
Disease_06	1	95.58	8.67	9.03	167.34	47.94	56.83	16.95	323.76	102.51
Disease_07	1	69.41	6.23	3.16	34.47	13.33	41.68	24.53	265.07	11.7
Disease_08	1	79.84	3.16	4.81	26.84	14.3	4.06	18.36	80.64	18.54
Unknown_01		91.84	17.64	22.87	441.42	79.04	157.59	62.8	897.85	419.89
Unknown_02		70.81	4.22	15.8	188.67	54.05	78.26	24.05	2489.91	170.72
Unknown_03		42.52	9.39	12.43	237.46	35.87	60.34	12.06	4447.07	97.51
Unknown_04		82.27	3.82	20.49	333.62	61.56	68.72	15.18	2643.87	55.7

Samples without metadata labels for prediction

Our Syllabus

Topic	Date	Lecture	Lab
Omics Data Science Foundations	Jan. 6	Omics data processing, statistics and visualization	--
	Jan. 13	From raw data to functional insights	--
Transcriptomics	Jan. 20	Gene expression data analysis (part I)	ExpressAnalyst & NetworkAnalyst
	Jan. 27	Gene expression data analysis (part II)	ExpressAnalyst & Seq2Fun
Proteomics & Biological Networks	Feb. 3	Biological network analysis & gene regulatory networks	NetworkAnalyst & miRNet
	Feb. 10	Proteomics & biomarker analysis	ExpressAnalyst & MetaboAnalyst
Metabolomics	Feb. 17	Targeted metabolomics data analysis	MetaboAnalyst
	Feb. 24	LC-MS untargeted metabolomics data analysis	MetaboAnalyst
Microbiomics	Mar. 2	Marker gene data analysis	MicrobiomeAnalyst
	Mar. 9	Shotgun metagenomics data analysis	MicrobiomeAnalyst
Multi-omics	Mar. 16	Knowledge-driven multi-omics integration	OmicsNet
	Mar. 23	Data-driven multi-omics integration	OmicsAnalyst

Tutorials



Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis

Jasmine Chong,[1] David S. Wishart,[4,5,6] and Jianguo Xia[1,2,3,6]

[1]Institute of Parasitology,
[2]Department of Animal S
Canada
[3]Department of Microbiol
Canada
[4]Department of Computin
[5]Department of Biologica

Basic Protocol 1: Data uploading, processing, and normalization
Basic Protocol 2: Identification of significant variables
Basic Protocol 3: Multivariate exploratory data analysis
Basic Protocol 4: Functional interpretation of metabolomic data
Basic Protocol 5: Biomarker analysis based on receiver operating characteristic (ROC) curves
Basic Protocol 6: Time-series and two-factor data analysis
Basic Protocol 7: Sample size estimation and power analysis
Basic Protocol 8: Joint pathway analysis

https://www.dropbox.com/s/pm6t6w2qo8q1z95/CPIB_MetaboAnalyst4.pdf?dl=0

We would like to hear your comment & feedback

contact@xialab.ca

See (most of) you next week!