Omics Data Science Training Course

Winter 2024

Schedule for today

Time	Topics
9:00 – 9:20	Introduction & logistics
9:25 – 10:00	Omics data journey
10:05 – 10:50	Statistics & performance measures
10:55 – 11:20	Visual analytics on omics data
Comments & Discussions	

You can download slides using the URLs from Zoom Chat

XiaLab Team

Jianguo (Jeff) Xia, PhD MD
Canada Research Chair in
Bioinformatics & big data analytics
jeff.xia@xialab.ca

Guangyan (Joe) Zhou, PhD
Transcriptomics, visual analytics
guangyan.zhou@xialab.ca

Zhiqiang (Qiang) Pang, PhD
Metabolomics, big data analytics
zhiqiang.pang@xialab.ca

Yao Lu, PhD Candidate
Microbiomics, multi-omics
yao.lu@xialab.ca

Our Class


23
20
3
7
Canada
Hudson Bay
Labrador Sea
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Nicaragua
Caribbean Sea
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
North Atlantic Ocean
South Atlantic Ocean
Norway
Finland
Russia
United Kingdom
Ireland
Denmark
Belarus
Poland
Germany
Ukraine
Romania
Italy
Spain
Portugal
Greece
Türkiye
Kazakhstan
Mongolia
Kyrgyzstan
Turkmenistan
Syria
Iraq
Iran
Afghanistan
Pakistan
Morocco
Tunisia
Algeria
Libya
Egypt
Western Sahara
Mauritania
Mali
Niger
Chad
Sudan
Saudi Arabia
Oman
Yemen
Gulf of Aden
Arabian Sea
Burkina Faso
Guinea
Nigeria
Ghana
Gulf of Guinea
Ethiopia
Somalia
Kenya
Gabon
DRC
Tanzania
Angola
Zambia
Mozambique
Namibia
Zimbabwe
Botswana
Madagascar
South Africa
Indian Ocean
Laccadive Sea
Nepal
India
Myanmar (Burma)
Thailand
Vietnam
Bay of Bengal
Gulf of Thailand
Malaysia
Indonesia
China
South Korea
Japan
Sea of Okhotsk
Sea of Japan
East China Sea
Philippine Sea
Banda Sea
Arafura Sea
Papua New Guinea
Australia
Great Australian Bight

Our Resources

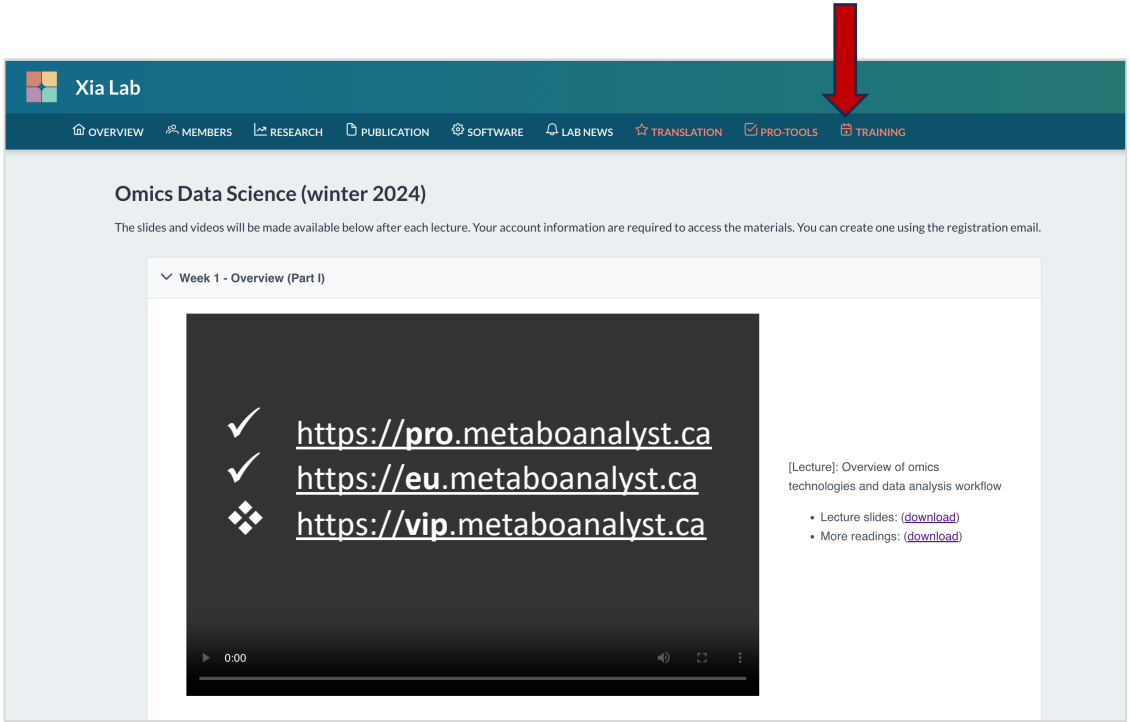
Recordings & slides

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

Pro Tools: 3 nodes

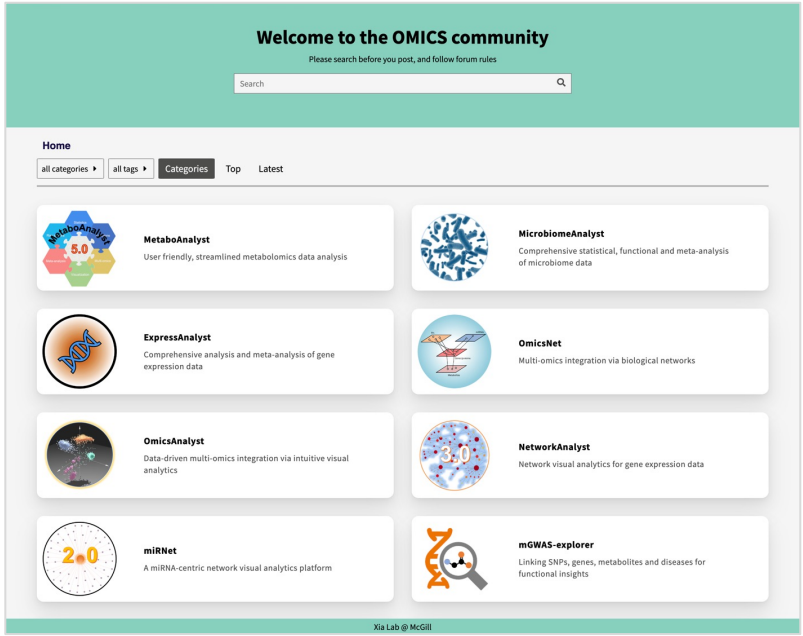
- PRO node (USA)
- EU node (Germany)
- VIP node (Canada)*

* Supporting raw data processing



Our Support

- User forum
  - https://omicsforum.ca
- Through email
  - support@xialab.ca
- Please provide sufficient information so we can reproduce the issue.
  1. Analysis goal
  2. Tool + input data
  3. Steps + parameters



www.omicsforum.ca

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	RNAseq data analysis in model species	ExpressAnalyst & NetworkAnalyst
	Jan. 27	RNAseq data analysis for non-model species	ExpressAnalyst & Seq2Fun
miRNAs & non-coding RNAs	Feb. 3	MicroRNAs, noncoding RNAs and biological networks	miRNet & NetworkAnalyst
Proteomics	Feb. 10	Proteomics data analysis and interpretation	ExpressAnalyst & NetworkAnalyst
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

Learning Objectives

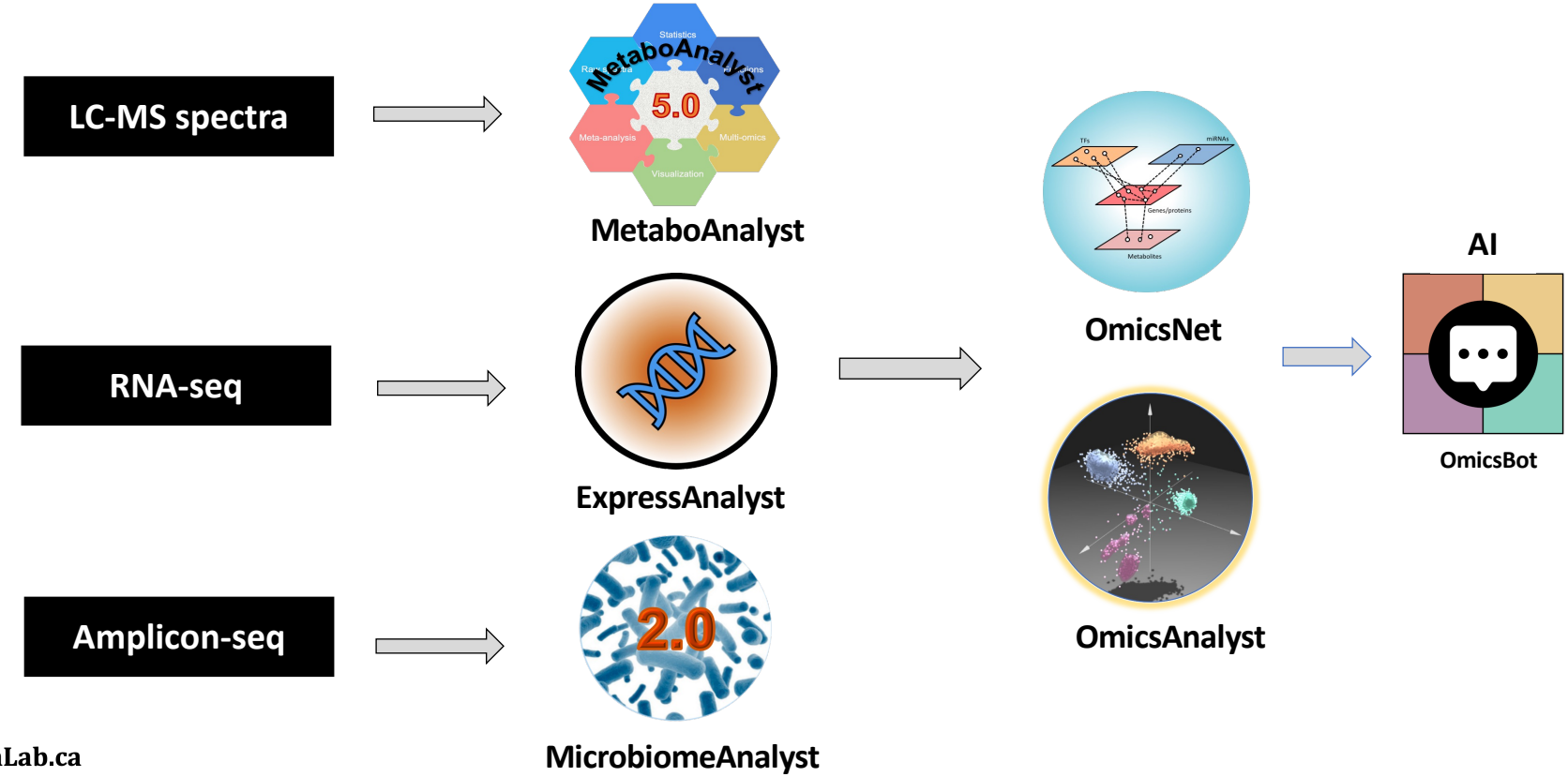
- Omics foundations
- Transcriptomics
- Proteomics
- Microbiome
- Metabolomics
- Multi-omics

→

✓ Data processing
✓ Statistics, ML
✓ Visualization
X Programming
X Computing

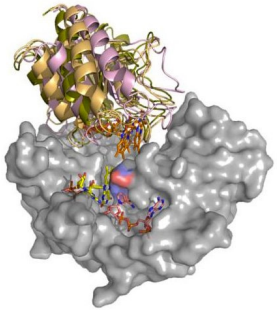
→

From raw data to multi-omics to AI

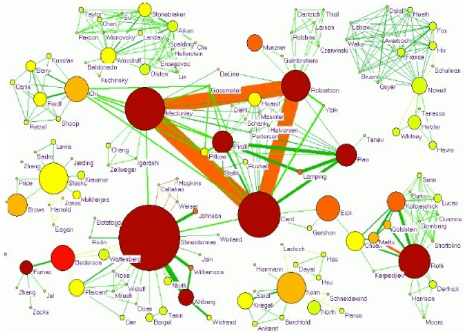


Goals of omics data analytics

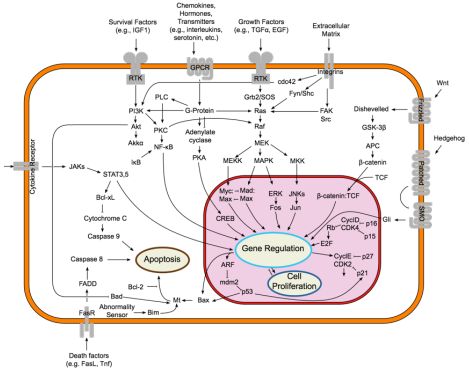
Individual Molecules



Omics & Multi-omics



Systems Biology



Mechanism ← **Patterns, functions, biomarkers, hypothesis** → Knowledge

Distinct challenges

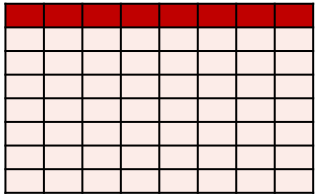
Size challenge (raw data processing)

- Raw reads, MS spectra, images
- Large (GB ~TB)
- Large storage and computing resources

Standardizable

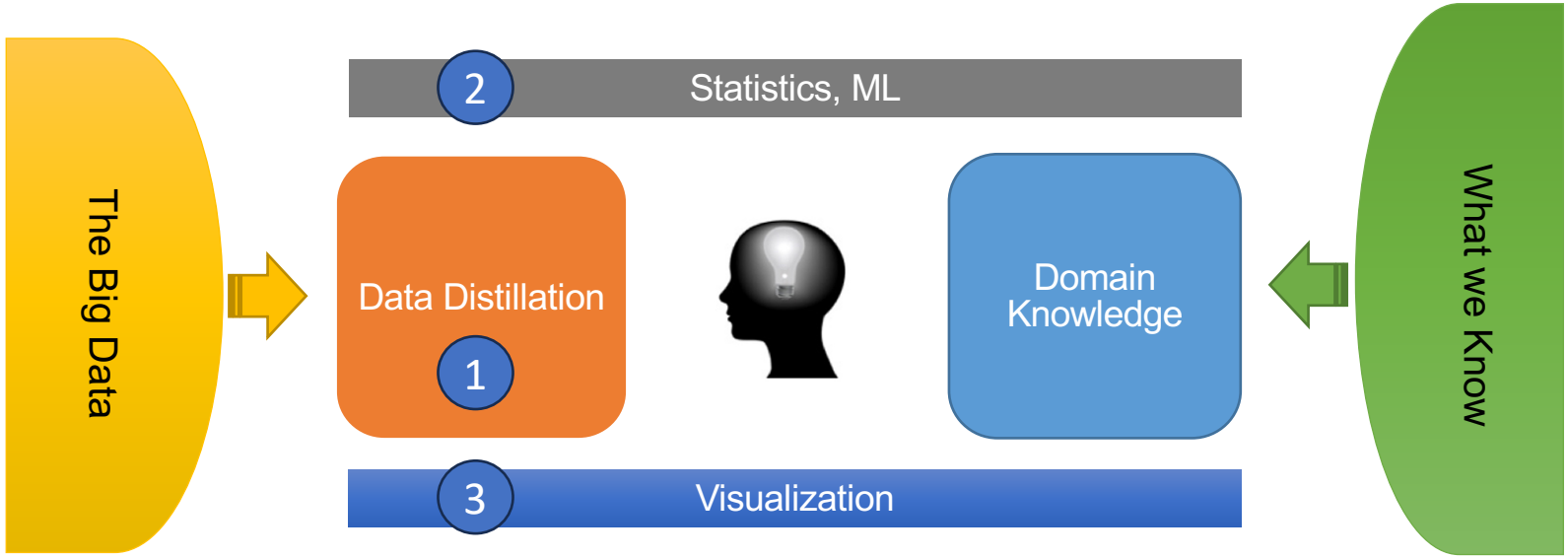
Complexity challenge (feature table)

- Feature table (abundance, intensities)
- Small (100s KB ~ MB)
- High-dimensional, heterogenous

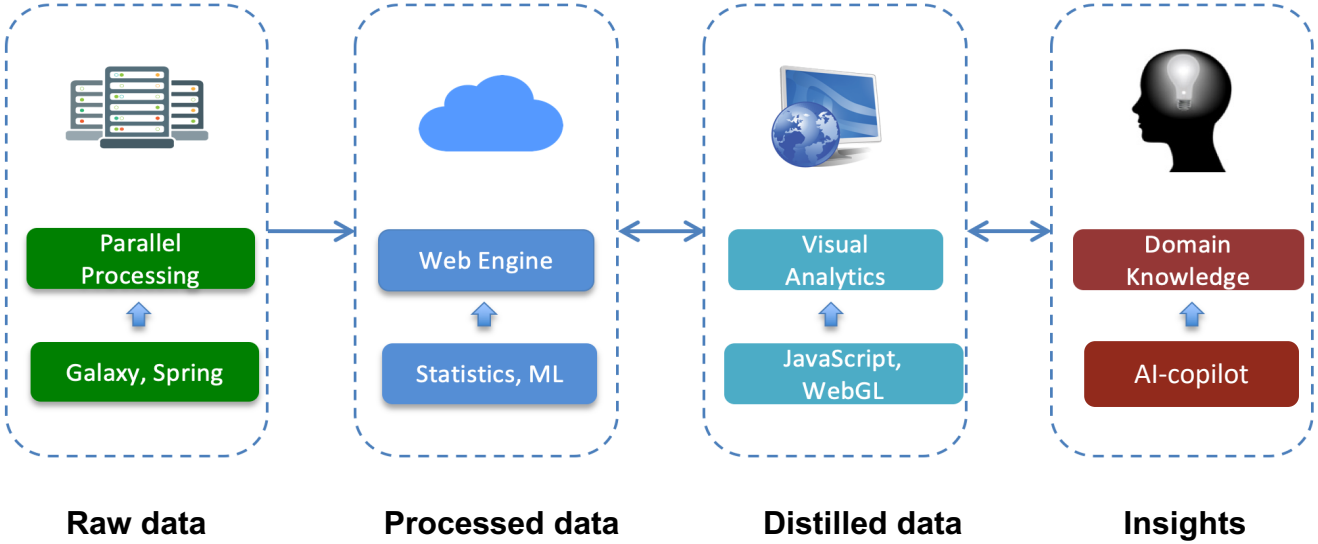


Open-ended

Dealing with omics data (conceptual design)



Omics data analytics platform (implementation)



Towards unified frameworks for omics data analysis

MetaboAnalyst 5.0 - user-friendly, streamlined metabolomics data analysis

Home
Data Formats
Tutorials
FAQs
APIs
Update History
MetaboAnalystR
Contact
User Stats
Publications
About

News & Updates

- MetaboAnalyst 5.0 is launched!
- Fixed the issue with synchronized 3D scatter plots for scores and loadings (04/12/2021);
- Upgraded to PrimeFaces 10 and fixed the issue for multi-group figure legend (04/06/2021);
- Added support for merging technical replicates with different mathematical models and QC measures (02/18/2021);
- Updated KEGG IDs in internal compound databases (02/09/2021);
- Enhanced name handling to better accommodate special characters in compound/lipid names (02/09/2021);
- Added seven tutorials introducing new features in MetaboAnalyst 5.0 (01/15/2021);
- Users can perform meta-analysis of global metabolomics data using Functional Meta-analysis module as illustrated in our COVID-19 paper (01/09/2021);
- Updated the KEGG global metabolic map for better metabolome coverage (01/08/2021);

Read more ...

nature protocols — PROTOCOL EXTENSION
https://doi.org/10.1038/s41596-022-00710-w

Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data

Zhiqiang Pang[1], Guangyan Zhou[1], Jessica Ewald[2], Le Chang[3], Orcun Hacariz[1],

2022

MicrobiomeAnalyst -- comprehensive statistical, functional and integrative analysis of microbiome data

Home | Formats | Forum | Updates | Resources

Please choose a module based on your data

Marker Data Profiling | Shotgun Data Profiling | Taxon Set Analysis

Nucleic Acids Research, 2023 1
https://doi.org/10.1093/nar/gkad407

MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data

Yao Lu[1], Guangyan Zhou[2], Jessica Ewald[2], Zhiqiang Pang[2], Tanisha Shiri[2] and Jianguo Xia[1,2,3,*]

2023

Home | Tutorials | Forum | Databases | ExpressAnalystR | Updates | Contact

ExpressAnalyst

-- a unified platform for gene expression data analysis

Start Here

nature communications

Article — https://doi.org/10.1038/s41467-023-38785-y

ExpressAnalyst: A unified platform for RNA-sequencing analysis in non-model species

Received: 20 October 2022
Accepted: 16 May 2023
Published online: 24 May 2023
Check for updates

Peng Liu[1,2], Jessica Ewald[1,2], Zhiqiang Pang[1], Elena Legrand[1], Yeon Seon Jeon[1], Jonathan Sangiovanni[1], Orcun Hacariz[1], Guangyan Zhou[1], Jessica A. Head[1], Niladri Basu[1] & Jianguo Xia[1]

The increasing application of RNA sequencing to study non-model species demands easy-to-use and efficient bioinformatics tools to help researchers quickly uncover biological and functional insights. We developed ExpressAnalyst (www.expressanalyst.ca), a web-based platform for processing, analyzing, and interpreting RNA-sequencing data from any eukaryotic species.

2023

Evolution of data analytics

Analytics 1.0

Elite Analytics

- Programming skills

Analytics 2.0

Visual Analytics

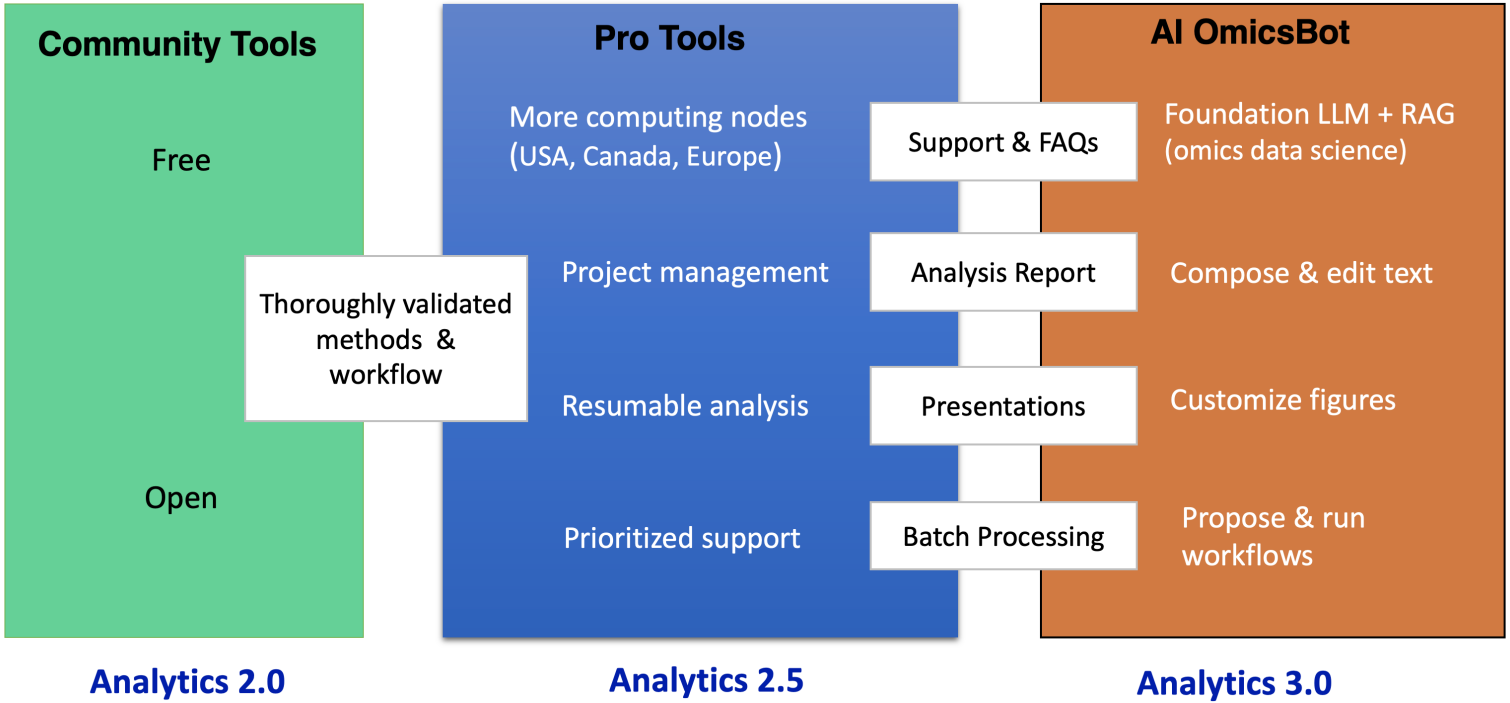
- Cloud/web technologies

Analytics 3.0

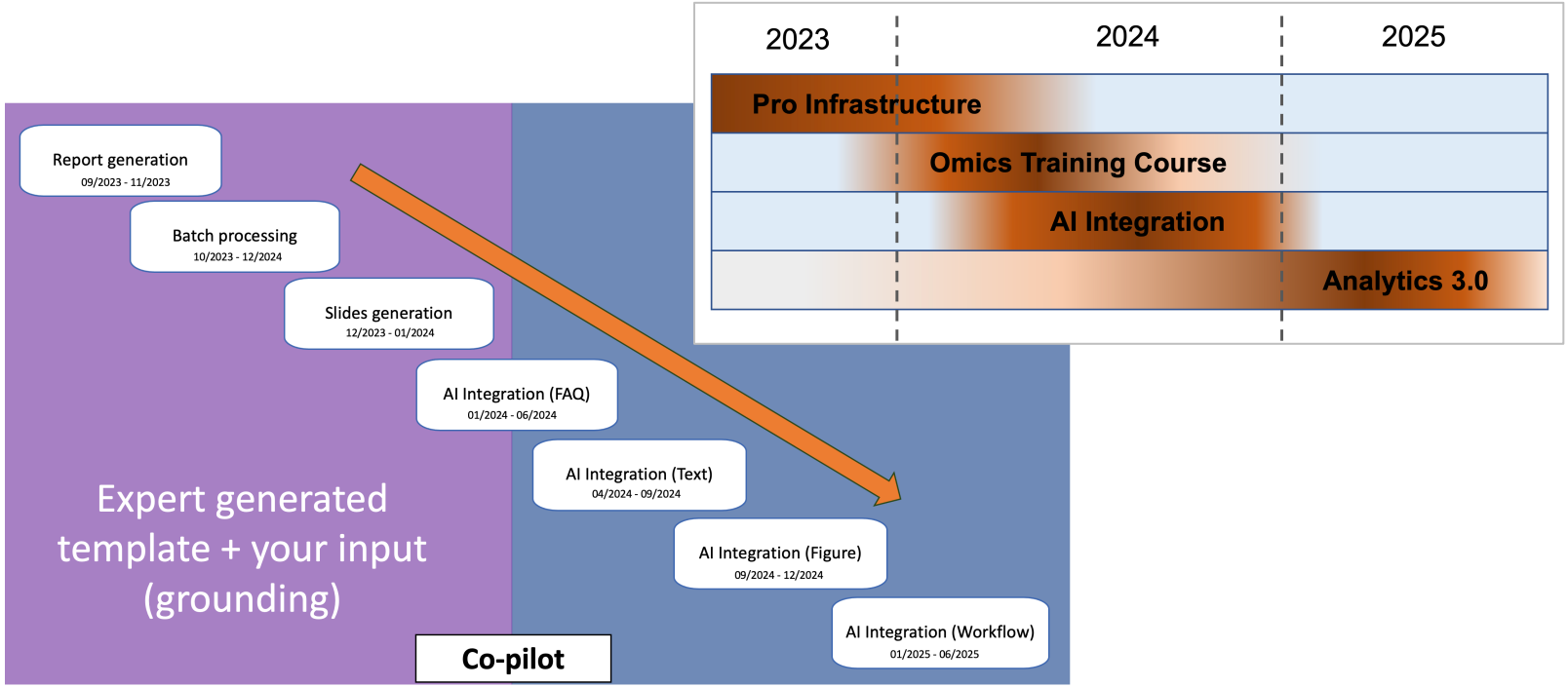
Conversational Analytics

- AI-augmented

AI will greatly accelerate data analytics



Roadmap (next 18 months)



Data ➔ Analysis Report (automate) ➔ Draft (iterate)

Upload > Data check > Normalization > Statistics > Correlations

Project Home | Edit | HTML | PDF | Slide Deck

Metabolomic Data Analysis with MetaboAnalyst Pro

1. Overview
2. Data Processing
3. Exploratory Statistical Data Analysis

Appendix: R Command History

Metabolomic Data Analysis with MetaboAnalyst Pro

Completed: Thu Jan 4 12:01:48 2024

1. Overview

The *Statistics [one factor]* module is designed for general-purpose exploratory analysis of metabolomics data with simple study design with one experimental factor (two group or multiple group). A variety of statistics, visualization and machine learning methods are supported. It accepts a data table from targeted metabolomics (i.e., concentration table) or untargeted metabolomics (i.e. peak intensity or binned spectral table). You can also upload peak list files from NMR or GC/LC-MS. For raw LC-MS spectra data, please use the *LC-MS Spectra Processing* module to first obtain a peak intensity table. For complex study design with

Live report, PDF report, slides, AI co-pilot, project storage, resumable analysis …

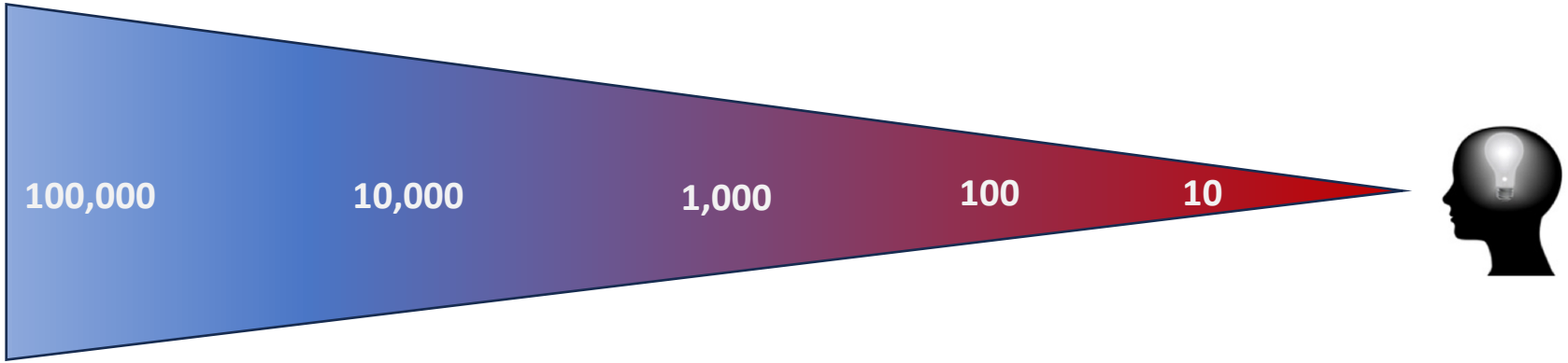
depending on your data type. For instance, missing value estimation and data filtering are primarily for untargeted metabolomics data which typically contain a lot of missing values and noises.

Schedule

Time	Topics
9:00 – 9:20	Introduction & logistics
9:25 – 10:00	Omics data journey
10:05 – 10:50	Pitfalls & performance measures
10:55 – 11:20	Visual analytics on omics data
Comments & Discussions	

==== 5 minutes break ====

Omics data journey

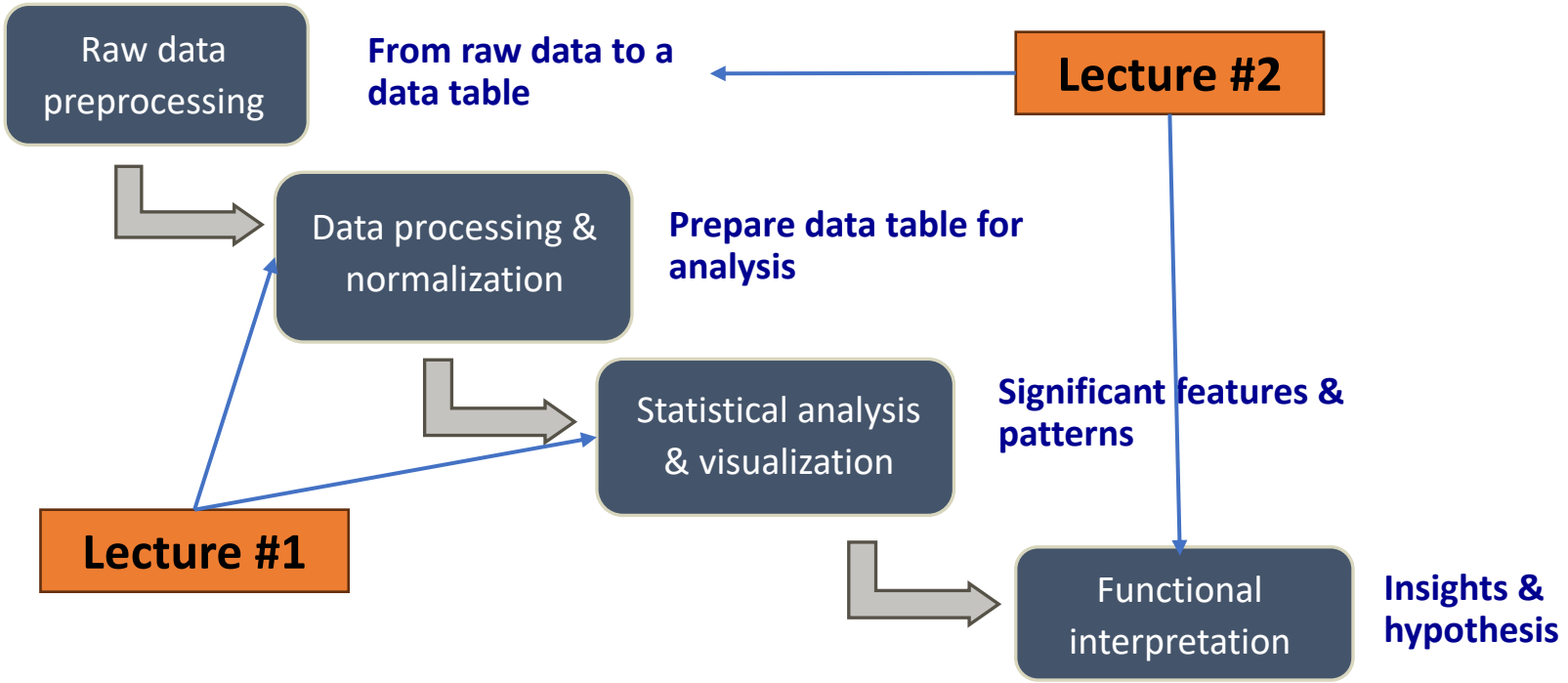


Omics data analysis aims to remove noise and identify main signals of interest

Keys concepts in omics data analytics

1. Data distillation
   - Not all data are useful
   - Big data contains big noise
2. Coupling statistics with visualization
   - P-values are not sufficient
   - Supporting user exploration
3. Integrating with domain knowledge
   - Data interpretation & hypothesis generation

Omics data analysis in a nutshell



The “shapes” of data at different stages

```
@M01380:62:000000000-B547W:1:1102:20819:1013 1:N:0:MWI006 NGCCTCTTI1INCTGCATAI1
NGTAGAGTTTGATTCTGGCTCAGGATGAACGCTGACAGAATGCTTAACACATGCAAGTCTACTTGATCCTTCGGGTGATGGTGGCGGACGGGTGAGTAACGCGTAAAGAACTTGCCCTGCAGTCTGGGACAACATTTGGAAACGAA
TGCTAATACCGGATATTATGCGAACTTCGCATGTAGCTCGTATGAAAGCTATATGCGCTGCAGGATAGCTTTGCGTCCTATTAGCTAGTTGGTGAGGTAACGGATCACCAAGGCCATGATCGGTAGCCGGGCTGAGTGTGTGAACG
GCCGCAAGG
+
#8BCCGGGGGGGGGGGGGFGGDFGFFGGFCFGGGDGFF8CEAFGFGGGGEDFGGGGGGFFGGGGGGGGGGCFAF7C<+DDFGGGD8@EFFFGGGGFGGGGCCFGGDCGDD?,B?ECG?A=FGDFGGGGGGGFF8FGGFGGG9EFF7B
FFFFFFDGCFG7CEFAF@FG,3FGGGG,+FCECGG=CC9:CCFFGGF9>CFFCGGFGGGC*6<@@,97FC@FG@EC88E?9F?F6>76+>AFC5C5EFAC6C**//@ZA=EGFEE437>:+1***122)/)/7*)9*:**)01
*)87}4),)-1:

@M01380:62:000000000-B547W:1:1102:16288:1015 1:N:0:MWI006 NGCCTCTTI1INCTGCATAI1
NTACGTAGGGTTCGATCCTGGCTCAGGATGAACGCTAGCTACAGGCTTAACACATGCAAGTCGAGGGGCAGCATCATCAAAGATTGCTTTGATGGATGGCGACCGGCGCACGGGTGAGTAACACGTATCCAACCTGCCGACAACAC
TGGGATAGCCTTTCGAAAGAAAGATTAATACCGGATGGCATAATTATTACGCATGGGATAATTATTAAAGAATTTCGGTGGCCGATGGGGGTGCGTTACATTAGGCAGATGGCGGGGGAAAGGCCTACCAAAACAACGACGGATAG
GGTGTGTGG
+
#8BACGG@BEFF8?EFFFFF8BCFGGFG,EECCF,CF:,,F=FECCFFDFGFGGFDCCEFFFGEGGG:@FCCDF8FFFGFGG8,9@,,?<C<CFGGEFF8FCCEEC?=7FFCG+8+AE<CBEGFEFF:BFFGFC8,,BF7@?CE8B
=FAB8,5,,7@FAE**>=@,FCCFA@FFCC;,>11*5*>FGFG9,@C9,6=CEGG88+29+37C+23+49<=9+7BFD8***3==/:=*;**/*1:C**+2+0:+3<C**+76==7*))*2979C**2)2)9)*)*.1>)87:.,9
*.,*4).4(
```

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<mzXML xmlns="http://sashimi.sourceforge.net/schema_revision/mzXML_3.2"  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://sashimi.sourceforge.net/schema_revision/mzXML_3.2 http://sashimi.sourceforge.net/schema_revision/mzXML
 <msRun scanCount="2535" startTime="PT0.09556S" endTime="PT359.981S">
  <parentFile fileName="file:///C:\Users\Admin\Desktop\EMP\Raw\Plate2/LC_Blank_10.raw"
          fileType="RAWData"
          fileSha1="f88f42711b213c6a4c6f08bb2d5a9f69d77a3341"/>
  <msInstrument msInstrumentID="1">
   <msManufacturer category="msManufacturer" value="Thermo Scientific"/>
   <msModel category="msModel" value="Q Exactive"/>
   <msIonisation category="msIonisation" value="electrospray ionization"/>
   <msMassAnalyzer category="msMassAnalyzer" value="quadrupole"/>
   <msDetector category="msDetector" value="inductive detector"/>
   <software type="acquisition" name="Xcalibur" version="2.5-204201/2.5.0.2042"/>
  </msInstrument>
  <dataProcessing centroided="1">
   <software type="conversion" name="ProteoWizard software" version="3.0.9935"/>
   <processingOperation name="Conversion to mzML"/>
   <software type="processing" name="ProteoWizard software" version="3.0.9935"/>
   <comment>Thermo/Xcalibur peak picking</comment>
  </dataProcessing>
```

Raw data

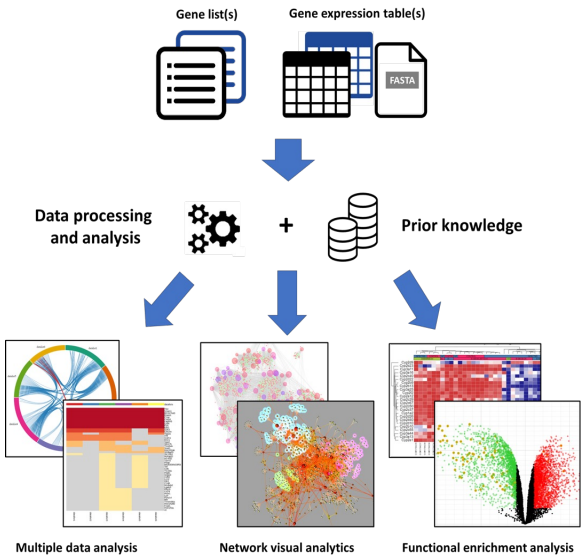
	PIF_178	PIF_087	PIF_090	NETL_005_V1
1,6-Anhydro-beta-D-glucose	40.85	62.18	270.43	154.47
1-Methylnicotinamide	65.37	340.36	64.72	52.98
2-Aminobutyrate	18.73	24.29	12.18	172.43
2-Hydroxyisobutyrate	26.05	41.68	65.37	74.44
2-Oxoglutarate	71.52	67.36	23.81	1199.91
3-Aminoisobutyrate	1480.3	116.75	14.3	555.57
3-Hydroxybutyrate	56.83	43.82	5.64	175.91
3-Hydroxyisovalerate	10.07	79.84	23.34	25.03
3-Indoxylsulfate	566.8	368.71	665.14	411.58
4-Hydroxyphenylacetate	120.3	432.68	292.95	214.86
Acetate	126.47	212.72	314.19	37.34
Acetone	9.49	11.82	4.44	206.44
Adipate	38.09	327.01	131.63	144.03
Alanine	314.19	871.31	464.05	589.93
Asparagine	159.17	157.59	89.12	273.14
Betaine	109.95	244.69	116.75	278.66
Carnitine	265.07	120.3	25.03	200.34
Citrate	3714.5	2617.57	862.64	13629.61
Creatine	196.37	212.72	221.41	85.63
Creatinine	16481.6	15835.35	24587.66	20952.22
Dimethylamine	632.7	607.89	735.1	1064.22
Ethanolamine	645.48	487.85	407.48	820.57
Formate	441.42	252.14	249.64	468.72
Fucose	336.97	198.34	186.79	407.48
Fumarate	7.69	18.92	7.1	96.54
Glucose	395.44	8690.62	1352.89	862.64
Glutamine	871.31	601.85	301.87	1685.81
Glycine	2038.56	1107.65	620.17	5064.45
Glycolate	685.4	651.97	141.17	70.81

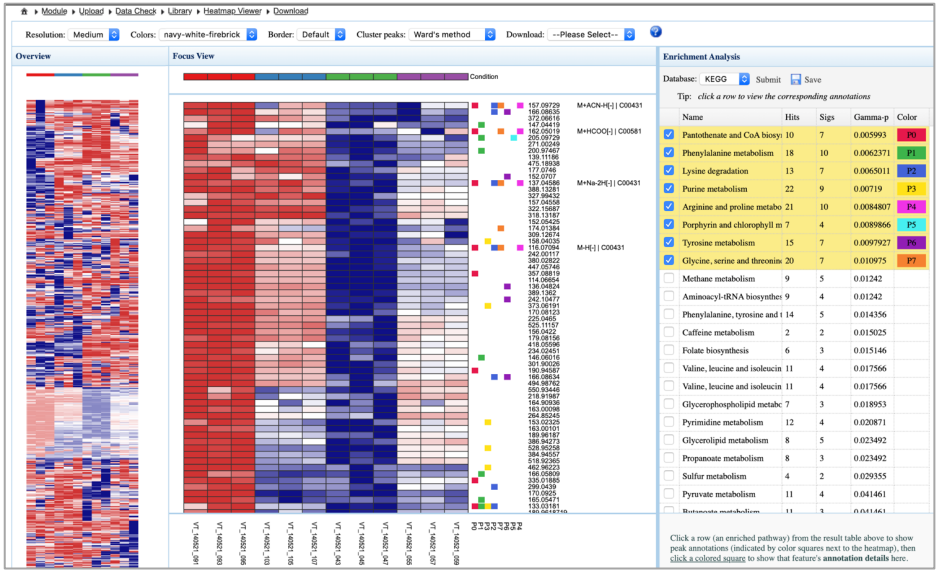
Processed data

Feature	P.Value
173.9855@96.75	2.20E-09
319.227@344.95	2.93E-09
347.2582@379.62	1.56E-08
343.2271@340.17	6.74E-08
254.2202@430	3.08E-07
335.2194@429.98	4.28E-07
406.132@281.13	4.72E-07
522.2822@375.26	5.91E-07
204.9804@118.14	6.31E-07
677.4839@504.04	9.47E-07
612.5063@496.65	1.06E-06
199.0061@210.94	2.01E-06
357.1785@488.19	2.07E-06
335.2944@537.41	2.32E-06
560.4736@486.66	3.01E-06
366.2259@488.17	3.45E-06
308.2675@497.04	3.76E-06
336.2978@537.42	4.25E-06
536.4643@478.91	4.64E-06
309.2794@532.6	4.85E-06
310.2826@532.64	5.41E-06
309.2708@497.09	7.21E-06
255.3947@479.04	7.27E-06
518.4835@478.95	8.99E-06
237.2218@479.02	9.02E-06
559.4702@486.62	9.68E-06
281.24@446.92	9.69E-06
533.455@479.03	1.02E-05

Distilled data

Need “contextual support” to understand

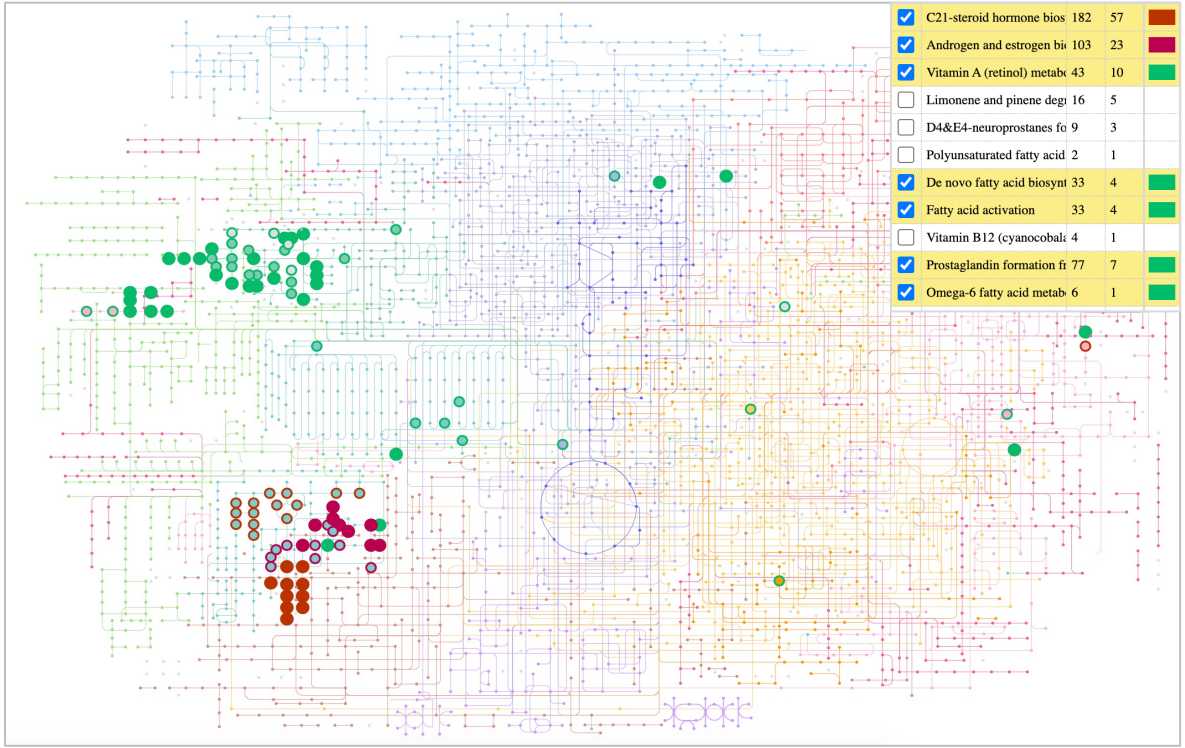




Visual Analytics

A very recent example

1. Raw data processing
   - **~30,000** features
2. Data cleaning
   - **~10,000** features
3. Statistical analysis
   - **~500** sig. features
4. Pathway analysis
   - **15** sig. pathways
5. Network analysis
   - Converging to **2** functional hotspots

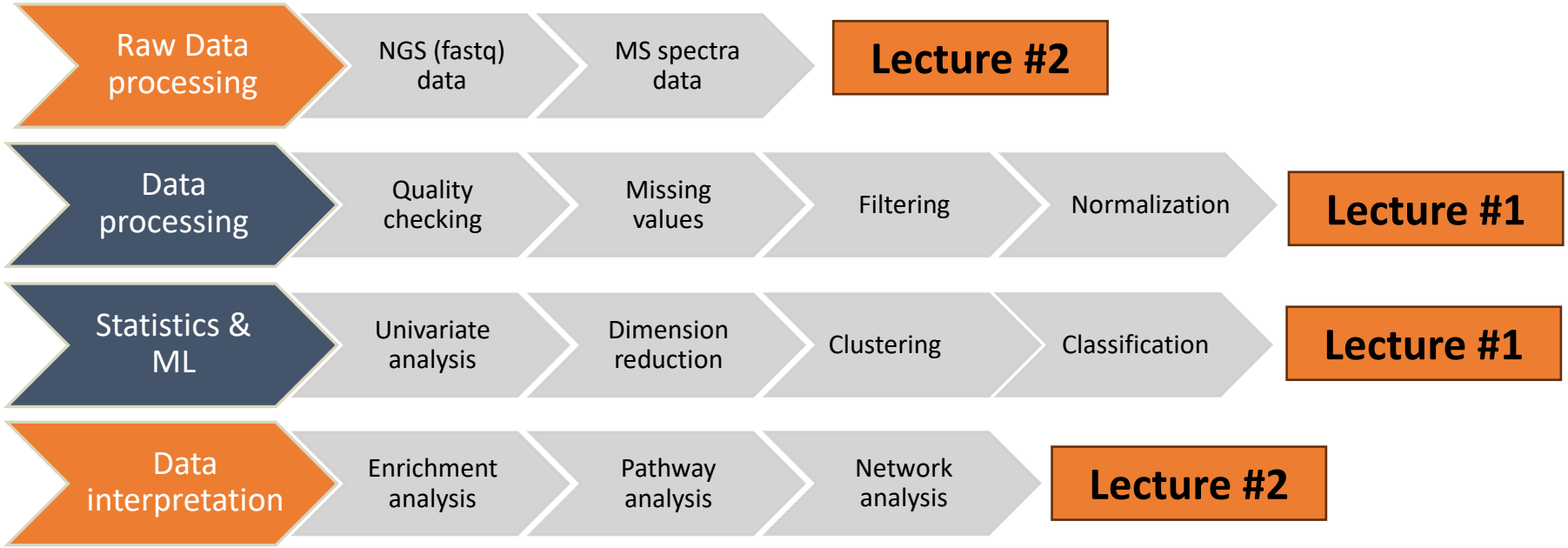


Two main types of raw omics data

1. Next-generation sequencing (NGS) data
   - Genomics
   - Transcriptomics
   - Microbiome
2. Mass spectrometry (MS) data
   - Proteomics
   - Metabolomics
   - Exposomics

Raw data processing is “largely” solved and standardized

Common steps in omics workflow



Data input & processing

Omics data & metadata (i.e. study design)

Metadata describes sample groups (i.e., treatment or control). It can be integrated into the omics data table (simple design), or uploaded as a separate file (complex design)

Sample	Label	Acetate	Acetone	Alanine	Betaine	Carnitine	Choline	Citrate	Creatine
Control_01	0	189.07	24.24	266.27	298.95	304.62	305.61	3969.16	366.7
Control_02	0	386.52	26.29	612.02	313.5	122.06	78.46	2075.6	435.99
Control_03	0	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.88	745.5	83.39
Control_04	0						0	10797.57	77.69
Control_05	0						7.98	1016.97	83.39
Disease_01	1						22.4	1486.6	152.5
Disease_02	1	325.39	13.29	295.53	85.1	89.36	44.93	3327.06	263.41
Disease_03	1	231.59	12.91	226.63	0	0	35.26	758.18	50.36
Disease_04	1	285.55	0	217.43	0	77.79	0	609.23	0
Disease_05	1	353.51	15.87	699.81	98.7	458.81	112.21	3415.49	229.79

Sample in rows

Sample	ko15	ko16	ko18	ko19	wt15	wt16	wt18	wt19
Label	KO	KO	KO	KO	WT	WT	WT	WT
200.1/2926	147887.53	451600.71	65290.38	56540.93	175177.08	82619.48	51951.61	69198.22
205/2791	1778569	1567038	1482796	1039130	1950287	1466781	1572679	1275313
206/2791	237993.6					2	211717.7	186850.9
207.1/2719	380873					5	277990.7	220972.4
219.1/2524	235544.92					5	72029.38	75096.99
231/2516	117649.77	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	96841.46	240261.21
233/3023	399145.3	356951.3	410550.7	198416.5	397107.8	271252.1	334459.9	181901.3
234/3024	76880.87	99526.27	97493.76	53461.71	65215.64	55952.44	73781.01	45211.66
235.1/2695	171995.22	128945.16	155442.48	115286.25	199981.49	30028.6	156968.3	52596.48
236.1/2524	252282.04	206031.93	71763.79	73602.47	253791.07	187225.65	79389.63	90012.64

Sample in columns

Simple design

#NAME	BPA5F1	BPA1F1	BPA4F2	BPA4F3	CON12F1	CON12F2	CON8F1	CON9F2	BPA4M2	BPA5M1	BPA1M2	BPA1M3	CON11M2	CON8M1	CON8M2	CON9M2
Plekhg2	109	174	80	76	83	88	52	76	99	76	76	75	57	52	67	46
Plekhg3	1281	1227	1263	990	546	687	561	661	866	799	895	925	930	498	579	744
Plekhg1	485	462	467	562	347	296	458	487	261	349	338	355	267	288	269	289
Plekhg6	126	172	206	198	137	164	110	152	247	193	223	251	159	184	190	185
4930592J03R	7	5	9	7	2	7	0	7	4	3	9	6	4	4	2	5
Plekhg4	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
Plekhg5	117	150	138								78	74	35	39	86	48
Nsa2	1540	1472	1577								1275	1109	1448	1074	1244	1426
Nampt	2882	3103	3487								2349	2684	3536	2406	2895	2691
Vmo1	43	68	51								51	52	25	29	21	32
Man2a1	11979	13268	15988								12304	12370	9859	8451	10191	9810
F930015N05	15	17	17	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11	19	5	3	7	6
4931406P16I	955	1038	971	934	514	810	559	874	526	572	708	720	547	584	603	538
Vwf	289	329	362	284	208	211	170	181	151	246	244	231	132	75	186	108
Mir302d	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Mier1	1262	1667	1510	1597	534	1086	558	1135	976	1084	1154	1004	1174	583	999	1095
Mier2	152	245	184	166	114	141	104	126	180	142	180	139	80	91	77	115
Mier3	949	1066	1220	1214	780	1177	808	1179	829	937	834	860	723	847	977	807
Aqp11	361	357	420	431	184	307	238	247	343	439	384	402	305	263	262	298
Aqp12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1810010H24	71	74	92	87	97	135	89	132	68	105	74	81	68	93	80	73

Omics data

#NAME	Treatment	Sex	liver_TG	liver_TC	liver_UC	liver_PC	plasma_TG	plasma_TC	plasma_HDL	plasma_UC	plasma_FFA	plasma_LDL	plasma_VLDL
BPA5F1	BPA	Female	77.5	2.82	2.35	15.26	25	63	46	15	70	12	5
BPA1F1	BPA	Female	77.5	2.82	2.35	15.26	31	103	76	13	74	20.8	6.2
BPA4F2	BPA	Female	81.65	3.22	2.47	14.85	18	79	55	8	61	20.4	3.6
BPA4F3	BPA	Female	98.92	3.24	2.27	16.38	19	79	63	8	65	12.2	3.8
CON12F1	Control	Female	180.19	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	66	11	48	20.4	1.6
CON12F2	Control	Female	134						56	7	61	18.4	1.6
CON8F1	Control	Female	137.74						55	7	75	15	8
CON9F2	Control	Female	160.54						70			18.2	12.8
BPA4M2	BPA	Male	88.97						81	13	55	21.8	7.2
BPA5M1	BPA	Male	63.03	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	93	15	59	23.4	3.6
BPA1M2	BPA	Male	99.23	2.77	2.31	14.77	48	116	90	14	66	16.4	9.6
BPA1M3	BPA	Male	123.62	3.84	3.2	24.72	55	135	89	14	64	35	11
CON11M2	Control	Male	84.46	2.29	2.11	14.6	93	129	96	22	86	14.4	18.6
CON8M1	Control	Male	73.19	2.12	1.93	12.71	106	125	90	23	78	13.8	21.2
CON8M2	Control	Male	91.94	2.83	2.36	16.27	110	120	85	18	73	13	22
CON9M2	Control	Male	110.68	2.72	2.14	15.15	93	125	84	24	84	22.4	18.6

Metadata

Complex design

General steps in data processing

1. (Samples) Quality checking
2. (Features) Missing value estimation
3. (Features) Data filtering
4. (Both) Normalization

Sample Space

Feature Space

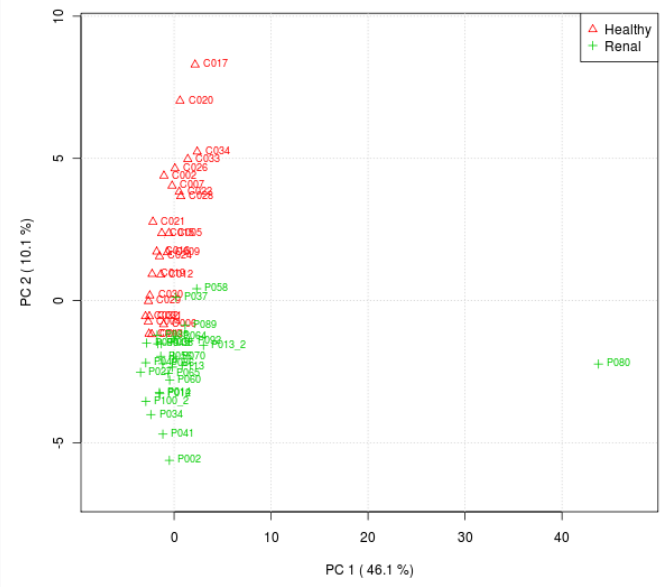
	PIF_178	PIF_087	PIF_090	NETL_005_V1
1,6-Anhydro-beta-D-glucose	40.85	62.18	270.43	154.47
1-Methylnicotinamide	65.37	340.36	64.72	52.98
2-Aminobutyrate	18.73	24.29	12.18	172.43
2-Hydroxyisobutyrate	26.05	41.68	65.37	74.44
2-Oxoglutarate	71.52	67.36	23.81	1199.91
3-Aminoisobutyrate	1480.3	116.75	14.3	555.57
3-Hydroxybutyrate	56.83	43.82	5.64	175.91
3-Hydroxyisovalerate	10.07	79.84	23.34	25.03
3-Indoxylsulfate	566.8	368.71	665.14	411.58
4-Hydroxyphenylacetate	120.3	432.68	292.95	214.86
Acetate	126.47	212.72	314.19	37.34
Acetone	9.49	11.82	4.44	206.44
Adipate	38.09	327.01	131.63	144.03
Alanine	314.19	871.31	464.05	589.93
Asparagine	159.17	157.59	89.12	273.14
Betaine	109.95	244.69	116.75	278.66
Carnitine	265.07	120.3	25.03	200.34
Citrate	3714.5	2617.57	862.64	13629.61
Creatine	196.37	212.72	221.41	85.63
Creatinine	16481.6	15835.35	24587.66	20952.22
Dimethylamine	632.7	607.89	735.1	1064.22
Ethanolamine	645.48	487.85	407.48	820.57
Formate	441.42	252.14	249.64	468.72
Fucose	336.97	198.34	186.79	407.48
Fumarate	7.69	18.92	7.1	96.54
Glucose	395.44	8690.62	1352.89	862.64
Glutamine	871.31	601.85	301.87	1685.81
Glycine	2038.56	1107.65	620.17	5064.45
Glycolate	685.4	651.97	141.17	70.81

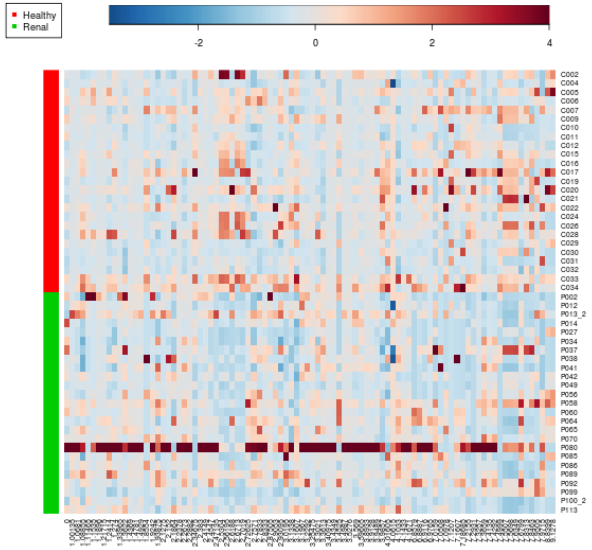
Quality checking

- The first & most critical step before analysis
  - ✓ Garbage in and garbage out
- Depending on
  - ✓ Good study design
  - ✓ Good laboratory practice
- Pay attention to
  - ✓ Sample outliers
  - ✓ Batch effects

Sample outliers

- Relative to the major grouping patterns

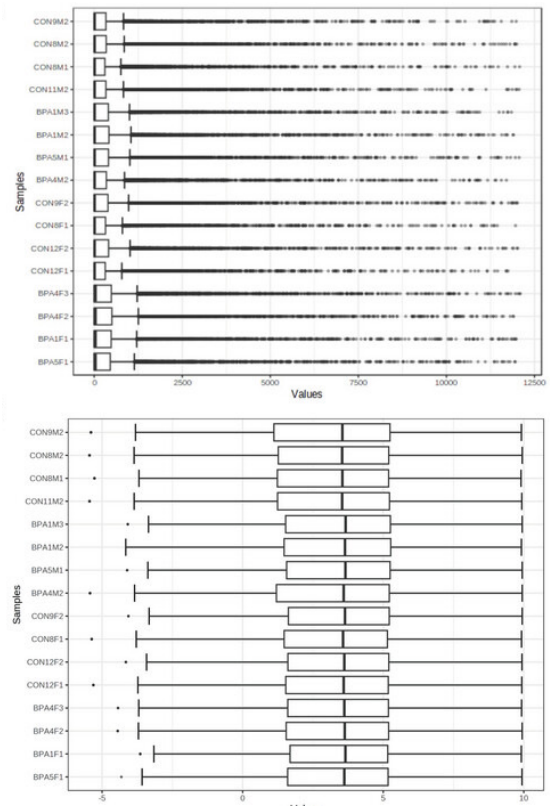




Feature outliers

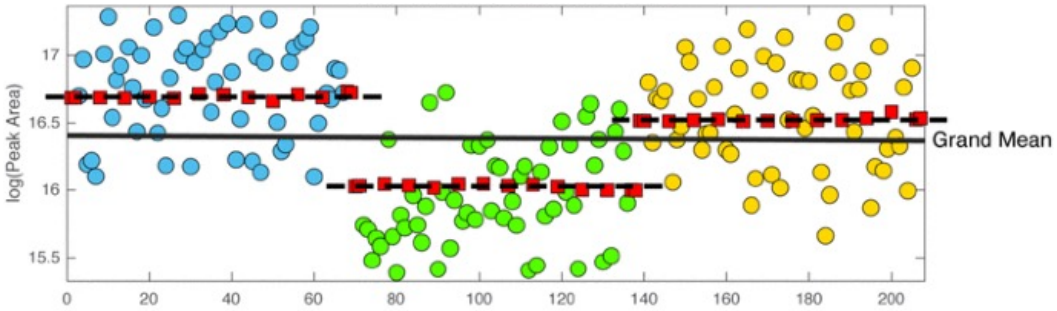
Outliers mainly concerning **sample outliers**, not about features

- ✓ Features are measured *en masse* in omics experiments
- ✓ Data filtering & normalization can address feature outliers (next steps)
- ✓ Statistical feature outliers could be our target of interest



Batch effects

- Display overall or systematic differences
- Technical (not biological) reasons
  - For instance, when samples are measured across multiple days, multiple labs, different lots, …..



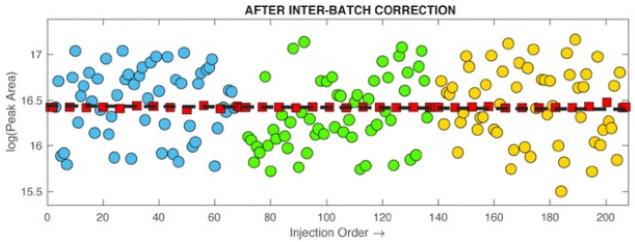
Addressing batch effects

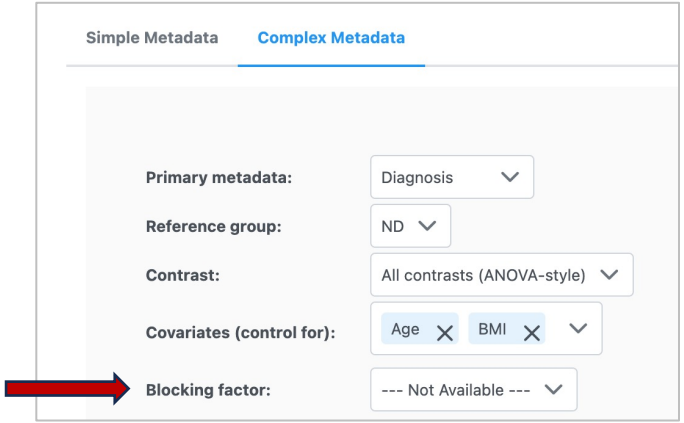
Batch effect removal methods perform batch effect correction <u>prior to statistical analysis</u>;

✓ Internal standards
✓ Computational estimation (i.e. ComBat)

Batch effect adjustment methods include batch variables within the model <u>during statistical analysis</u>, such that differences associated with batch are accounted for in analysis.

- For instance, as a “Blocking” factor





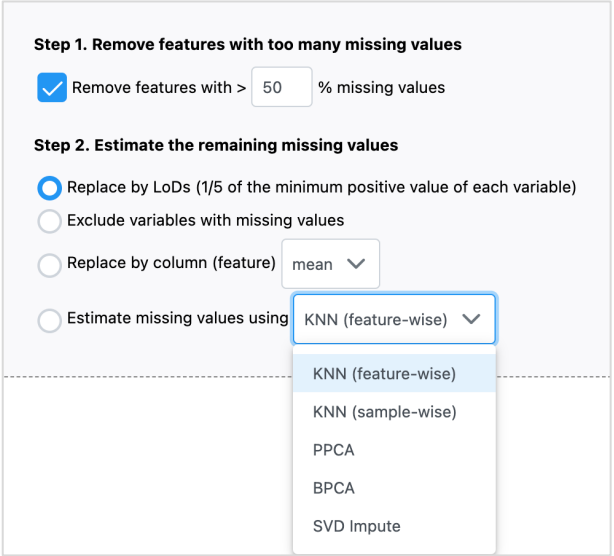
Missing values

- Common in omics data. Can be introduced during data collection, or by algorithms during raw data pre-processing (i.e. peak picking)
- Most statistical methods will complain if input contains missing values

1.4781	2	1.05	1.84	0.89	1.33	1.94	1.43	0.85	1.52	1.48	2.52	2.22
1.4929	2.03	1.06	1.86	0.88	1.33	1.13	1.46	0.88	0.97	1.47	1.88	2.19
1.8554	NA	0.47	0.83	NA	1.31	NA	NA	NA	NA	0.6	NA	0.79
1.9242	1.82	1.59	1.45	1.73	3.13	1.91	1.79	1.58	1.77	1.99	3.26	3.35
1.93875	NA	0.76	1.1	0.83	2.62	0.8	1.53	1.45	0.94	1.8	NA	3.36
2.1275	1.19	0.72	NA	NA	2.88	NA	1.68	NA	0.94	1.1	NA	3.16
2.152	NA	2.25	2.9	1.25	8.75	5.02	1.09	1.91	1.33	3.13	2.84	4.18
2.1864	NA	1.3	2.7	0.8	3.47	2.84	NA	0.9	NA	1.98	2.05	2.35
2.2378	1.58	0.61	1.03	1.75	0.84	1.51	0.72	0.77	1.15	0.85	0.79	0.96

Dealing with missing values

- When a large portion of features are missing, missing value estimation could have a large impact on the downstream analysis
- Why missing?
  - Missing completely at random
  - Missing at random
  - Missing not at random
- Estimating missing values
  - Detection limits (missing due to low abundance)
  - Mean/median
  - Weighted average values from similar samples

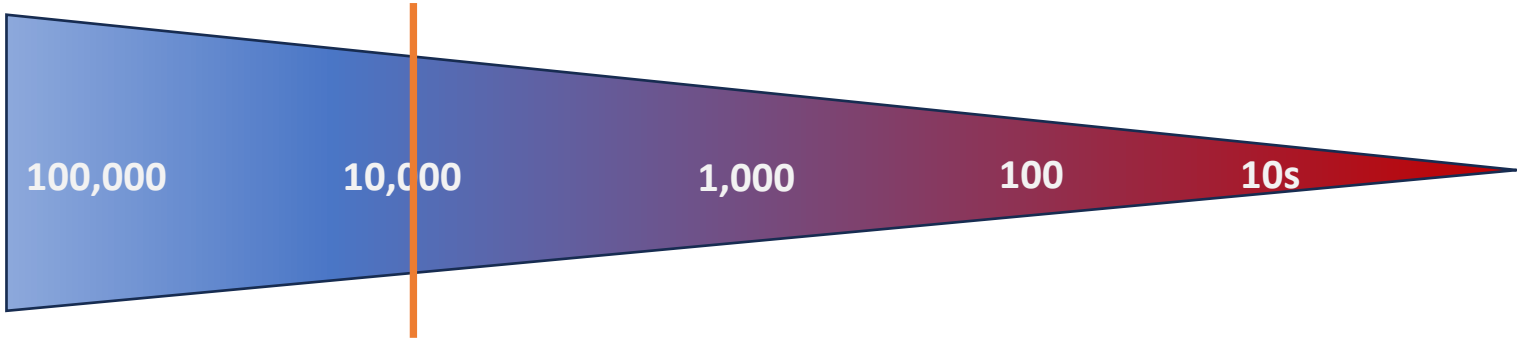


Motivation for feature filtering

- Big data contains big noises
- Not all features will respond to the treatment
  - A small percentage (5~10%) in most cases
- There are redundancies in omics data for most features
- Filtering non-informative features before statistical analysis can often <u>significantly</u> improve the power

What proportion to filter?

- Transcriptomics (microarray & RNAseq)
  - **Over 50%** of features can be removed with improved results
    - Better classification performance, more stable biomarkers
- Untargeted metabolomics (many peaks are redundant)
  - We recommend to filtering 30~50% peaks for data generated from high-resolution LC-MS instrument.



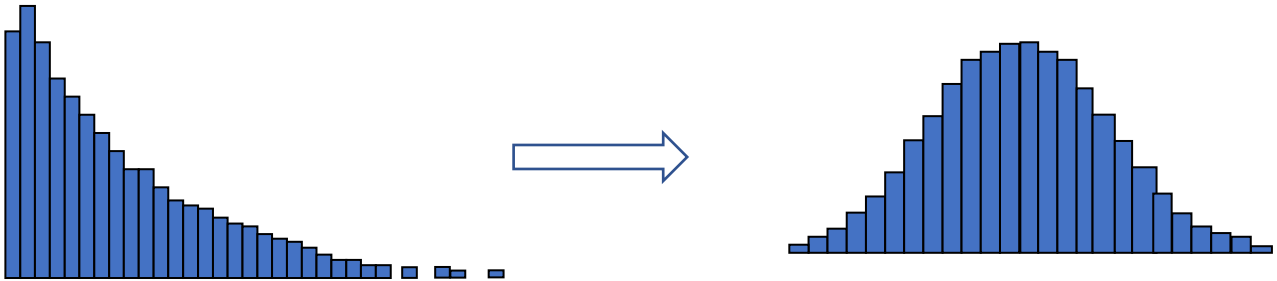
Feature filtering criteria

- **Low variance**
  - Variables that are near-constant values throughout the experiment conditions (housekeeping or homeostasis)
- **Low abundance**
  - Variables of very small values (close to baseline or detection limit).
- **Low quality**
  - Too many missing values
  - Hard to measure: low repeatability based on technical repeats

DO NOT filter features based on their p-values or fold changes

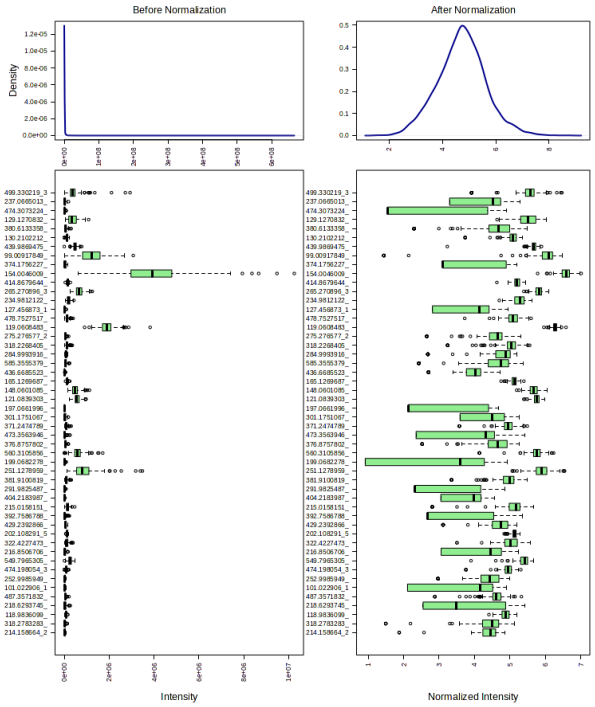
Normalization

The goal is to make omics profiles more comparable (across samples and/or across features) and to transform them to be more suitable for statistical analysis and visualization.



Why normalization?

- ❖Most statistical methods assume (or work best) when features are normally distributed
- ❖Feature abundance levels can vary across several magnitudes - the changes of abundant variables will dominate analysis (i.e., PCA) if unadjusted
  - More abundant features do not mean they are more important
- ❖Adjust other effects:
  - Dilutions, tissue volumes, etc



Common normalization methods

1. Centering
2. Scaling
3. Transformation

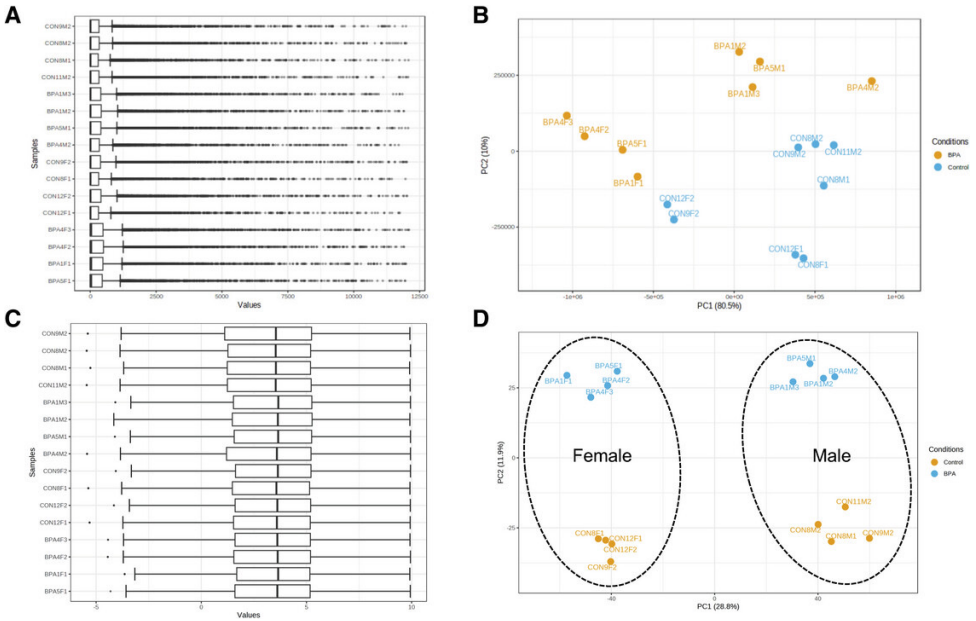
There is NO guarantee of global normal distribution in omics data

Method	Formula	Unit	Goal	Advantages	Disadvantages
Centering	$\tilde{x}_{ij} = x_{ij} - \bar{x}_i$	0	Focus on the differences and not the similarities in the data	Remove the offset from the data	When data is heteroscedastic, the effect of this pretreatment method is not always sufficient
Autoscaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$	(-)	Compare metabolites based on correlations	All metabolites become equally important	Inflation of the measurement errors
Range scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\left(x_{i_{\max}} - x_{i_{\min}}\right)}$	(-)	Compare metabolites relative to the biological response range	All metabolites become equally important. Scaling is related to biology	Inflation of the measurement errors and sensitive to outliers
Pareto scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{s_i}}$	0	Reduce the relative importance of large values, but keep data structure partially intact	Stays closer to the original measurement than autoscaling	Sensitive to large fold changes
Vast scaling	$\tilde{x}_{ij} = \frac{\left(x_{ij} - \bar{x}_i\right)}{s_i} \cdot \frac{\bar{x}_i}{s_i}$	(-)	Focus on the metabolites that show small fluctuations	Aims for robustness, can use prior group knowledge	Not suited for large induced variation without group structure
Level scaling	$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_i}{\bar{x}_i}$	(-)	Focus on relative response	Suited for identification of e.g. biomarkers	Inflation of the measurement errors
Log transformation	$\tilde{x}_{ij} =^{10} \log\left(x_{ij}\right)$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	Log 0	Correct for heteroscedasticity, pseudo scaling. Make multiplicative models additive	Reduce heteroscedasticity, multiplicative effects become additive	Difficulties with values with large relative standard deviation and zeros
Power transformation	$\tilde{x}_{ij} = \sqrt{\left(x_{ij}\right)}$ $\hat{x}_{ij} = \tilde{x}_{ij} - \bar{\tilde{x}}_i$	ν0	Correct for heteroscedasticity, pseudo scaling	Reduce heteroscedasticity, no problems with small values	Choice for square root is arbitrary.

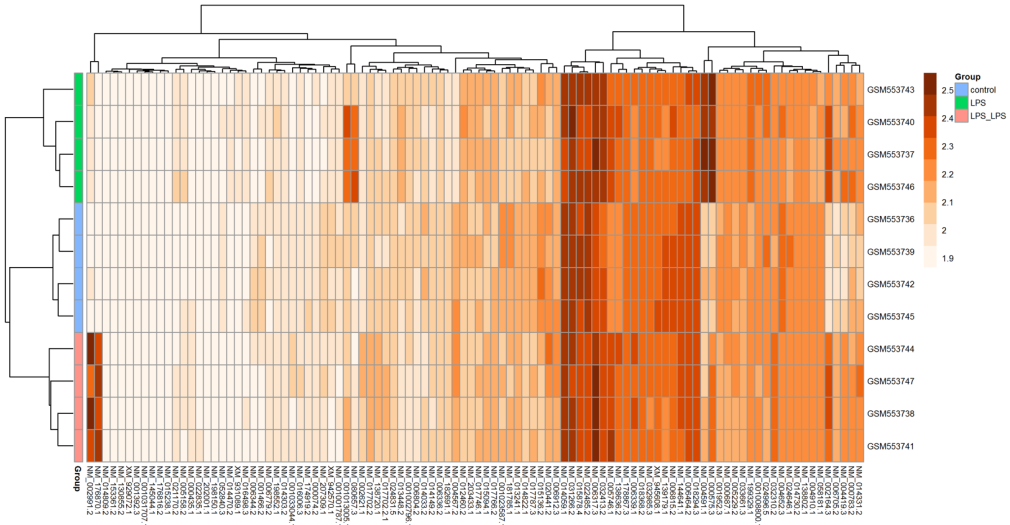
BMC Genomics. 2006; 7: 142

Normalization considerations

- Try simple methods first
  - Complex methods make data difficult to interpret
  - Most physiological measures are **log**-normal
  - Auto-scale (unit transformation, or **Z-score**)
- Some methods have their built-in normalization methods
  - EdgeR, DEseq, etc.
- Normalization often bring out meaningful patterns
  - QC samples are tighter
  - Other meaningful groups (i.e., gender)

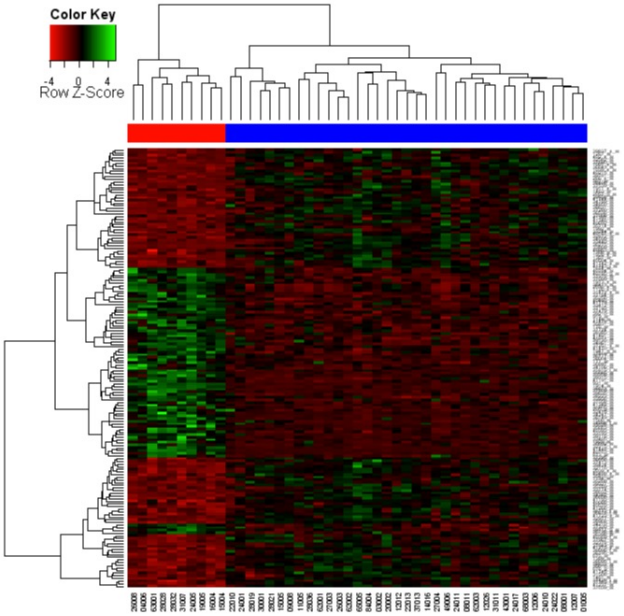


Normalization for visualization



Log Transformation

Comparing expression levels <u>across features</u>



Auto-scaling

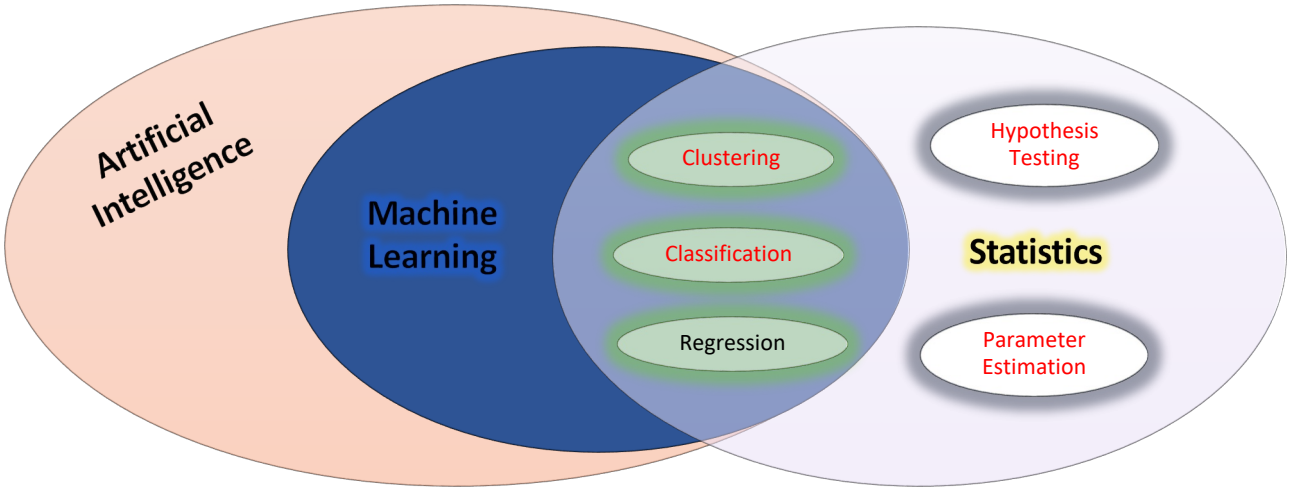
View changes <u>across conditions</u>

Schedule

Time	Topics
9:00 – 9:20	Introduction & logistics
9:25 – 10:00	Omics data journey
10:05 – 10:50	Pitfalls & performance measures
10:55 – 11:20	Visual analytics on omics data
Comments & Discussions	

==== 5 minutes break ====

Landscape of analysis methods



Analysis strategies on “cleaned” omics data

1. **Statistica**l analysis to identify significant features
   - Univariate
   - Multivariate
2. **Machine learning**
   - Unsupervised learning (clustering analysis): explore the data to find some intrinsic structures in them, without considering whether they are related to the class labels or not
   - Supervised learning (classification): identify features or patterns in the data that are associated with group labels.

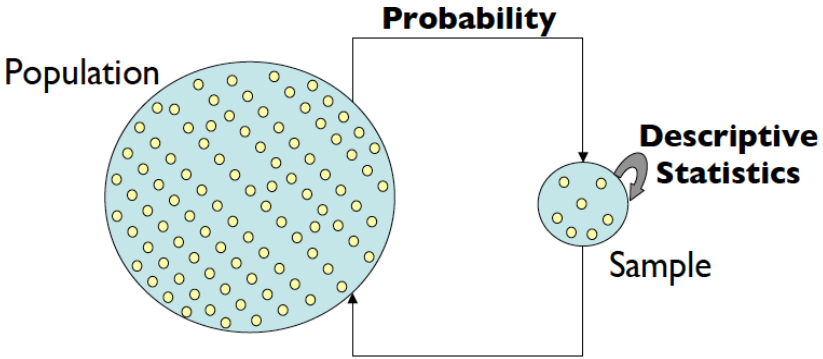
How do we evaluate the results for decision making?

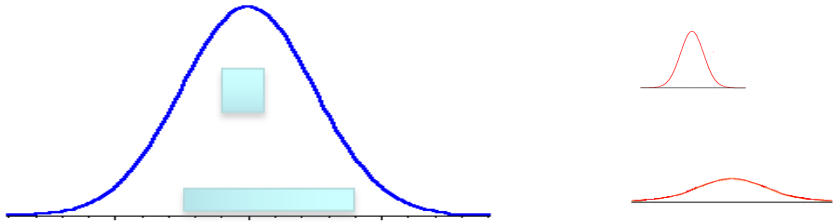
- P-values
  - Is the result likely due to random chance?
  - Commonly used for statistics
- Performance measures
  - How good is it?
  - Commonly used supervised methods
- Visual analytics
  - Does the result make sense based on prior knowledge?
  - Commonly employed for clustering methods

Statistics

- ✓ Focus on hypothesis testing & p-values
- ✓ Primarily univariate

Uncertainties in parameter estimation





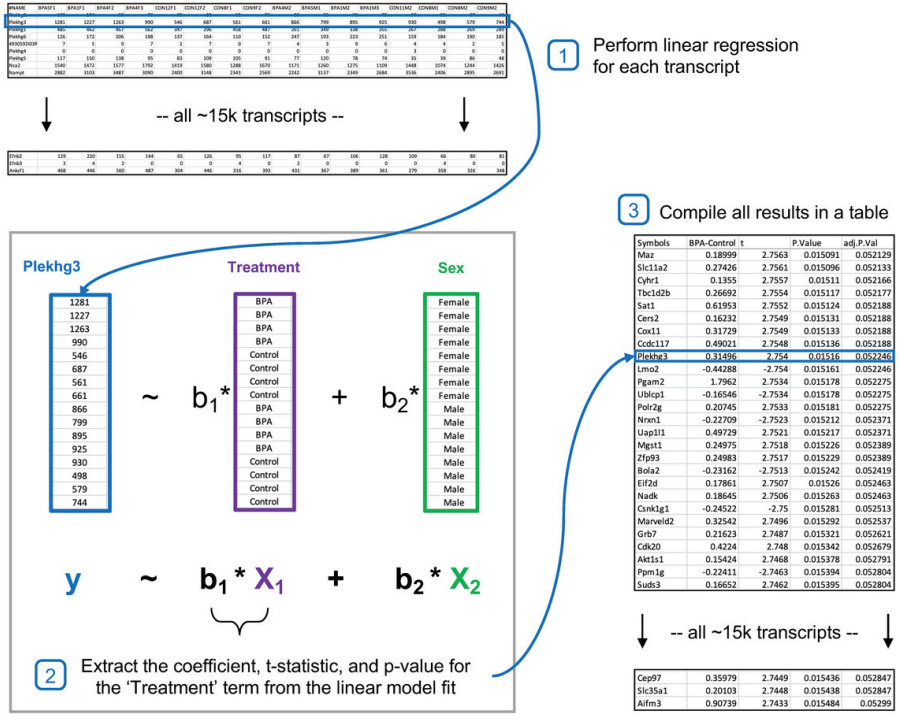
Sampling Bias

Using p-values for decision making

- How do we know whether the effect observed in our sample was genuine?
  - **We don't**
- We use p-values to indicate our level of uncertainty that our results represent a genuine effect present in the whole population
- P values = the probability that the observed result was obtained by chance
  - i.e. when the null hypothesis (H_0: *i.e.* no effect) is true
- If that probability (p-value) is small, it suggests the observed result cannot be easily explained by chance
  - i.e. it is considered significant

Most widely used practice

- Univariate statistics
  - T-tests / ANOVA
  - Linear modeling (more complex design)
- First analyze a single features, and then apply the procedure to all features, finally do multiple testing adjustment



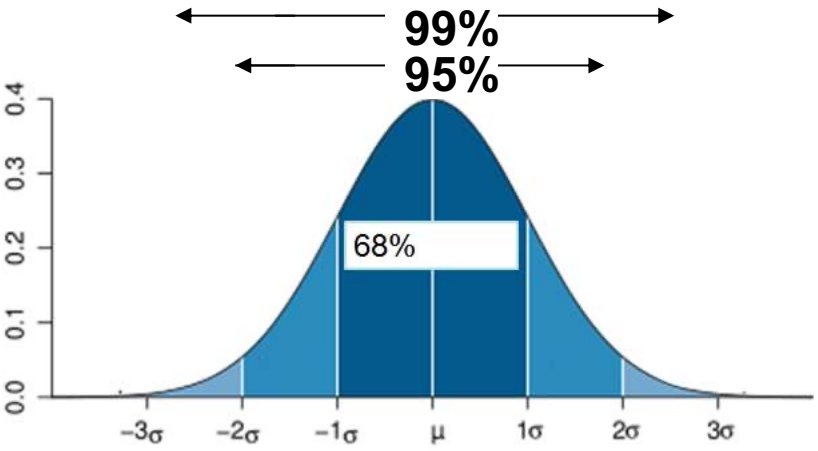
Ewald, et al. Current Protocols 2023

P-values & multiple testing adjustment

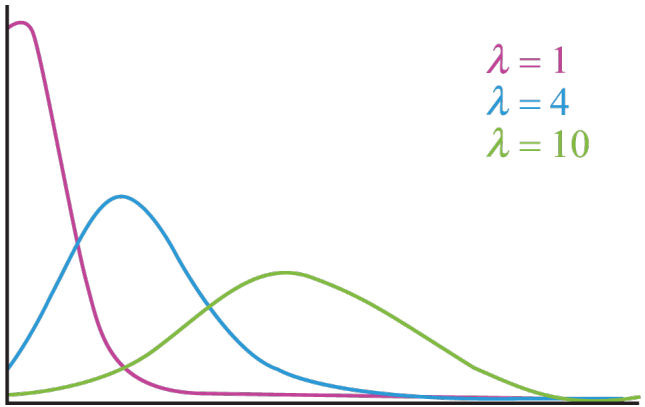
- Lots of hypothesis tests, with each one we accept a small chance.
- Performing T-tests on “cleaned” omics data might result in performing 10000 separate hypothesis tests. If we use a standard p value cut-off of 0.05, we would see 500 (10000*0.05) genes to be deemed “significant” by chance alone!
  - Family Wise Error Rate (FWER) - e.g. Bonferroni corrections
    - ✓ Corrected P-value= p-value * n (number of genes in test) <0.05. If testing 1000 genes at a time, the corrected p-value is 0.00005 (0.05/1000).
  - False Discovery Rate (FDR) - e.g. Benjamini-Hochberg
    - ✓ A FDR of 0.05 means that 5% among the significant metabolites are expected to be false positives

How do we compute p-value?

We can compute p values at any data points when distribution is known



Gaussian distr.



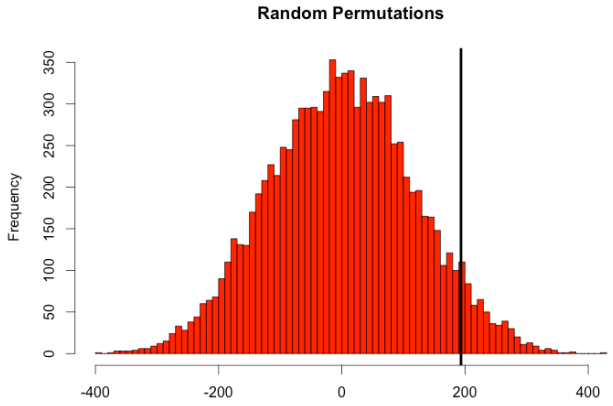
Poisson distr.

What if we don’t know the distribution?

- The only thing we know is that the data does not follow a normal distribution
- Poor performance using a model based on normal distribution
  - Try to perform normalization, or
  - We can find out the null distribution (i.e. no effect) from the data itself using a technique called **permutation**, then calculate the p-value
    - Also known as <u>empirical p-values</u>

Permutation tests to estimate null distribution

If all data (control and treatment data) are from the same population (i.e. no effect), then the data statistics calculated based on permuted group labels will be very similar to the statistics calculated from the original data labels



Computing empirical p-values

1. Under the null model (i.e. no effect), all data comes from the same distribution
   - ➔ Any subgroup/samples will be similar to each other
   - ➔ group labels are not actually meaningful (b/c they are same under the null model)
2. We calculate the original statistic (i.e. mean difference) based on the original groups
3. We then <u>shuffle (permute) the group labels</u> and recalculate the mean difference of the new groups (null stats)
4. Repeat Step #3 many many times (i.e. typically from 1000s to million times)
5. Find out where the original statistic lies in comparison to the null statistics distribution

An Example

Case-Control data, and we want to find out if there is <u>a difference in **mean**</u>

Mean difference .541

	case		control
1	-0.49274	10	1.471227
2	-0.30228	11	0.612679
3	0.093007	12	-0.47886
4	0.715722	13	0.746045
5	1.272872	14	0.871994
6	-1.37599	15	0.985237
7	-0.14798	16	-0.44421
8	-1.22195	17	0.246393
9	1.2812	18	0.68246
Mean	-0.01979		0.52144

Permutation #1

Note how the different **labels** have been swapped for the permutation

Mean difference = .329

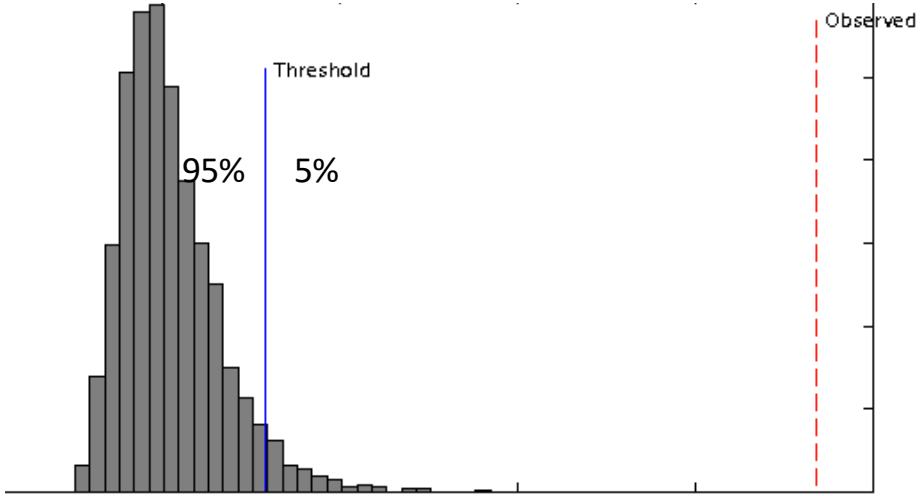
	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
15	0.985237	4	0.715722
16	-0.44421	6	-1.37599
1	-0.49274	2	-0.30228
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Permutations

	case		control
9	1.2812	11	0.612679
3	0.093007	18	0.68246
17	0.246393	14	0.871994
7	-0.14798	5	1.272872
10	1.471227	12	-0.47886
13	0.746045	8	-1.22195
Mean	0.415354		0.086295

Repeat many many many many times
(and then repeat many more times)

Compute an empirical p-value



In 1000 times permutations

- There are three times the permuted data given large difference

  ➔ **p = 0.003**

- None of the permuted mean difference is bigger than the original one

  ➔ **p < 0.001 or (1/1001) # prevent p value equal to zero**

General advantages

- Does not rely on distributional assumptions
  - Corrects for hidden selection
  - Corrects for hidden correlation
- Can be applied to almost universally – as long as there is a meaningful summary statistics can be computed
- Disadvantage??
  - Computationally intensive
  - Relatively large samples to have enough different shuffling combinations (> 40)

Clustering & dimension reduction

Identify patterns

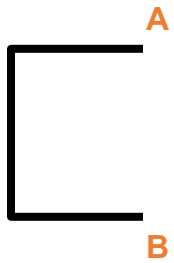
Clustering

- Organize omics data into smaller blocks
- Each block are more homogenous, we can treat these blocks as a unit

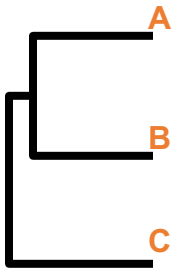
Dimensionality reduction

- Reduce the high-dimensional data into low-dimension i.e. from 1000s of variables to 2 ~ 3 variables
  - Principal component analysis (PCA)
- Play important roles in multi-omics integration

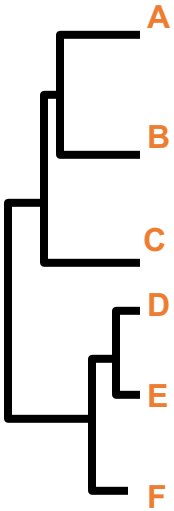
Hierarchical clustering & heatmaps



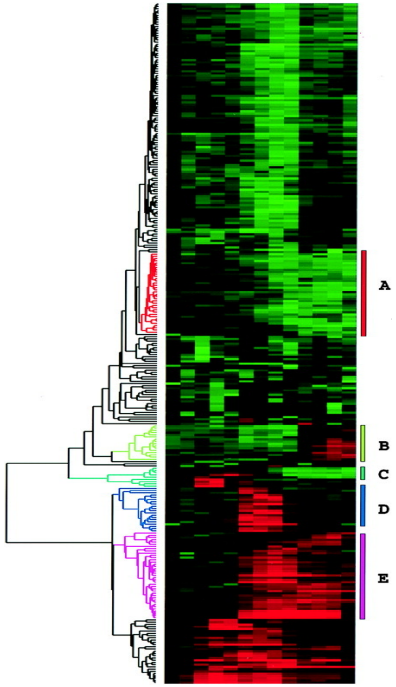
Find two most similar metabolite expression levels or curves



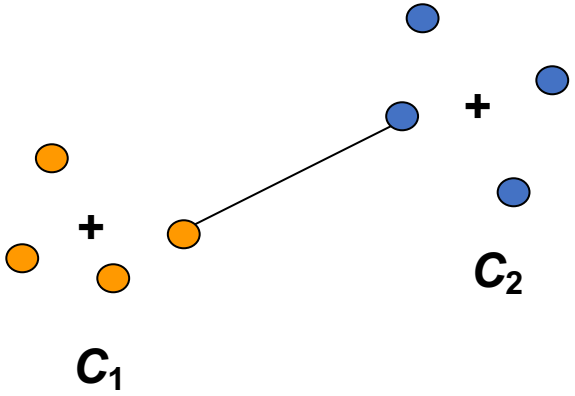
Find the next closest pair of levels or curves



Iterate



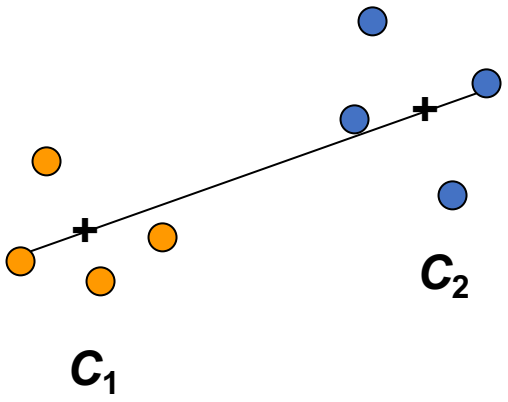
Single linkage clustering



Dissimilarity between two clusters = minimum dissimilarity between the members of two clusters

Tend to generate "long chains"

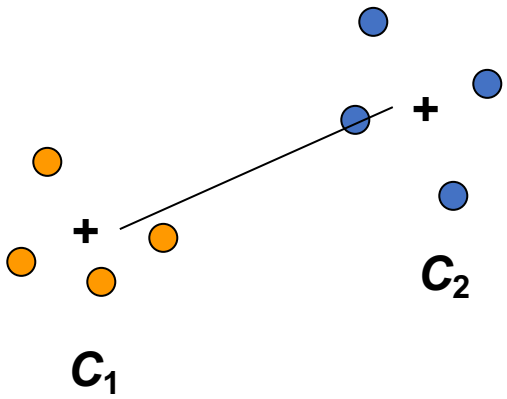
Complete linkage clustering



Dissimilarity between two clusters = Maximum dissimilarity between the members of two clusters

Tend to generate "clumps"

Average linkage clustering

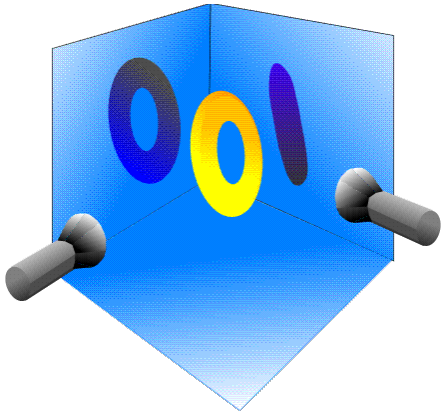


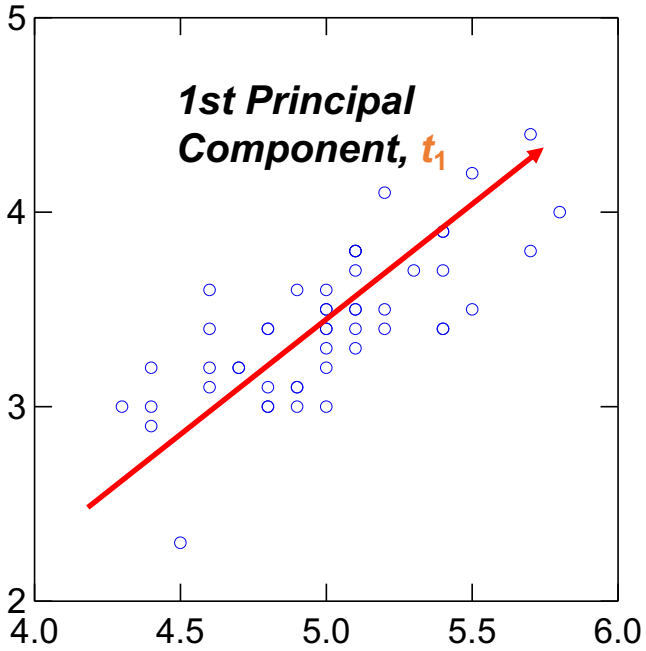
Dissimilarity between two clusters = Distance between two cluster means.

Principal Component Analysis (PCA)

- Project high-dimensional data into lower dimensions that capture the most variance of the data
- Main directions of variance

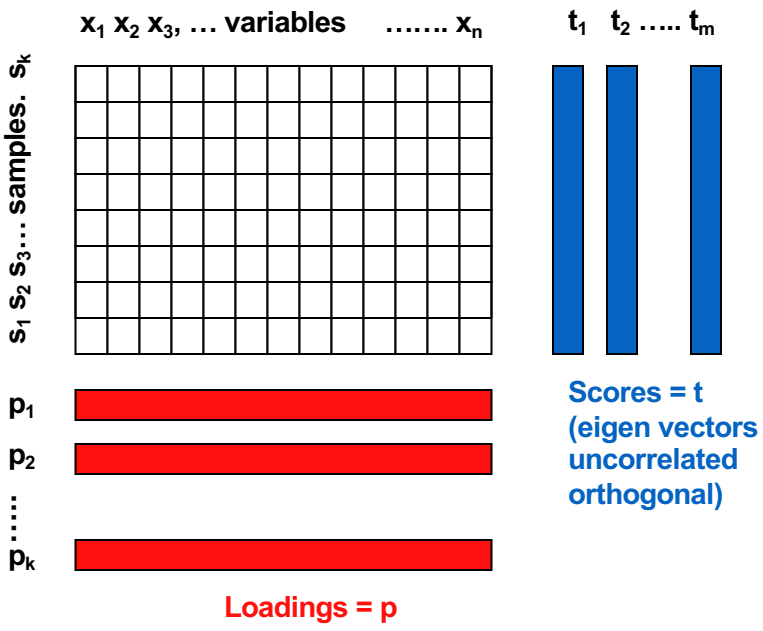
  ≈ major data characteristics





PCA on omics data

- PCA transforms data to a new coordinate system so that the greatest variance of the data comes to lie on the first coordinate (1st PC), the second greatest variance on the 2nd PC etc.
- PCA involves the calculation of the eigenvalue (singular value) decomposition of a data covariance matrix



PCA details

From k original variables: $x_1, x_2, \ldots, x_k$:

Produce k new variables: $t_1, t_2, \ldots, t_k$

$t_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$

$t_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2k}x_k$

...

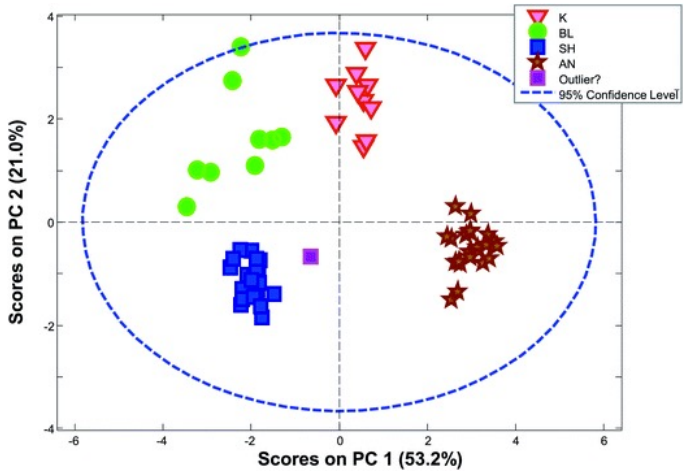
$t_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{kk}x_k$

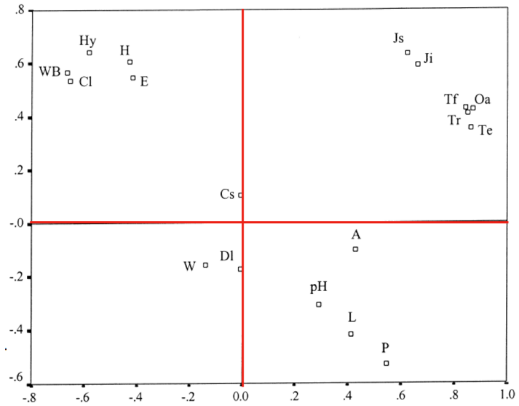
Such that:

- t_k's are uncorrelated (orthogonal)
- t_1 explains as much as possible of original variance in data set
- t_2 explains as much as possible of remaining variance, etc.

Scores and Loadings

- *Scores:* samples in the low-dimensional space defined by the components
- *Loadings:* feature coefficients





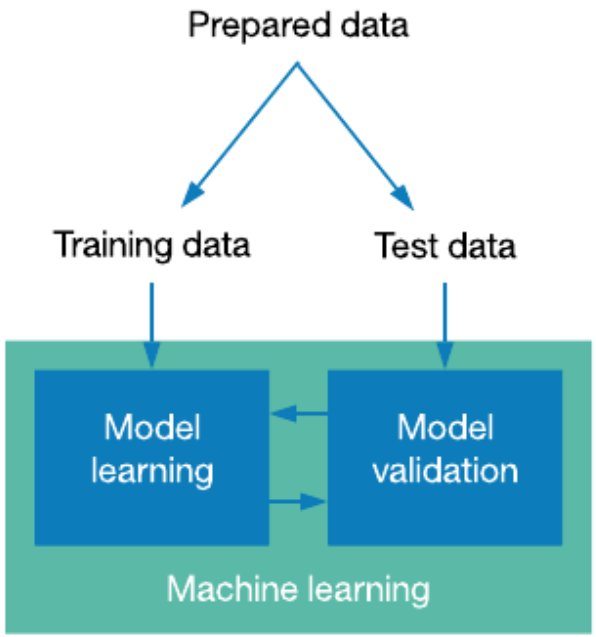
PCA Summary

- Project high-dimensional omics data into a new coordinate system in which the top components capture most data variance
- Common usages:
  - Data overview
  - Outlier detection
  - Look at relationships between samples
  - Look at relationships between features
  - Check which features contribute most to the separation pattern

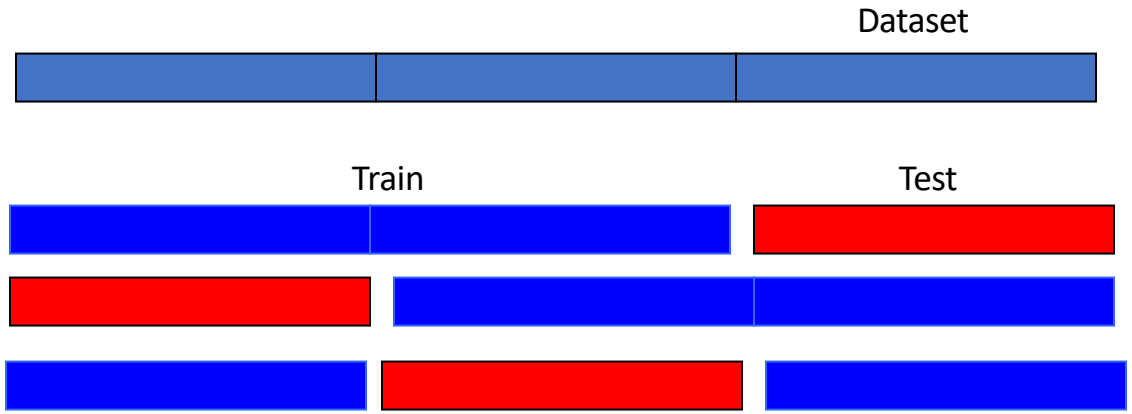
Performance Measures

Supervised Learning

1. Data preparation
2. Algorithm/model selection
   - Support vector machine
   - Random forests
   - Logistic regression
3. Parameter tuning
   - Training
   - Testing



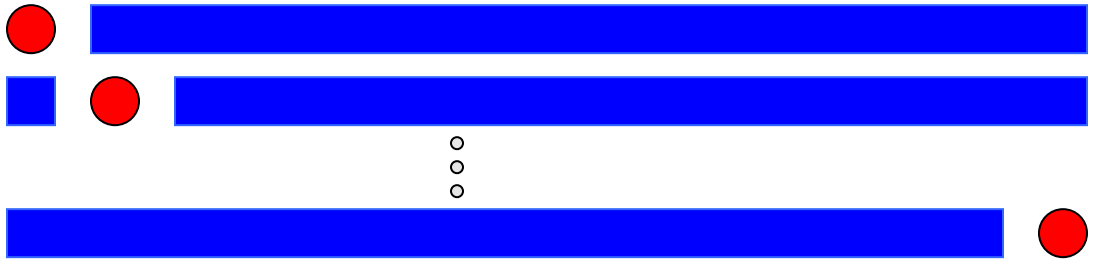
Cross validation (CV)



Three-fold cross-validation

- Repeat three times on each fold and report average performance
- May have variations due to random splitting (outliers) for small sample size

Leave One Out Cross Validation (LOOCV)



- More stable results
- Computationally intensive for large data

Evaluating Performances (I)

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

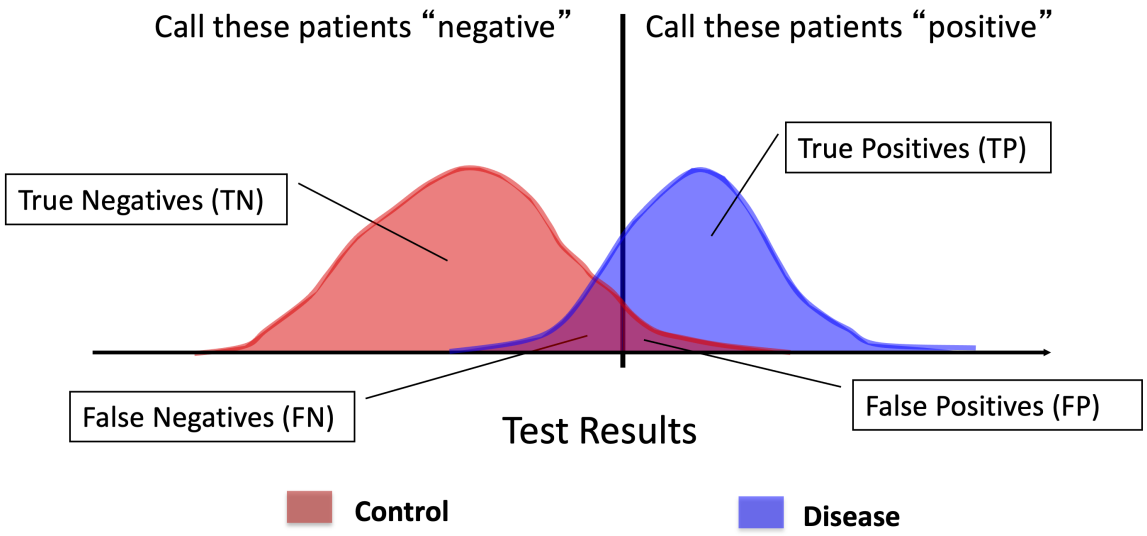
Evaluating Performances (II)

- 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

Sensitivity & specificity

- How good in predicting positive cases
  - Sensitivity (Sn)
    - Sn = TP / (TP + FN)
    - True positive rate
- How good in predicting negative cases
  - Specificity (Sp)
    - Sp = TN/(TN + FP)
    - True negative rate



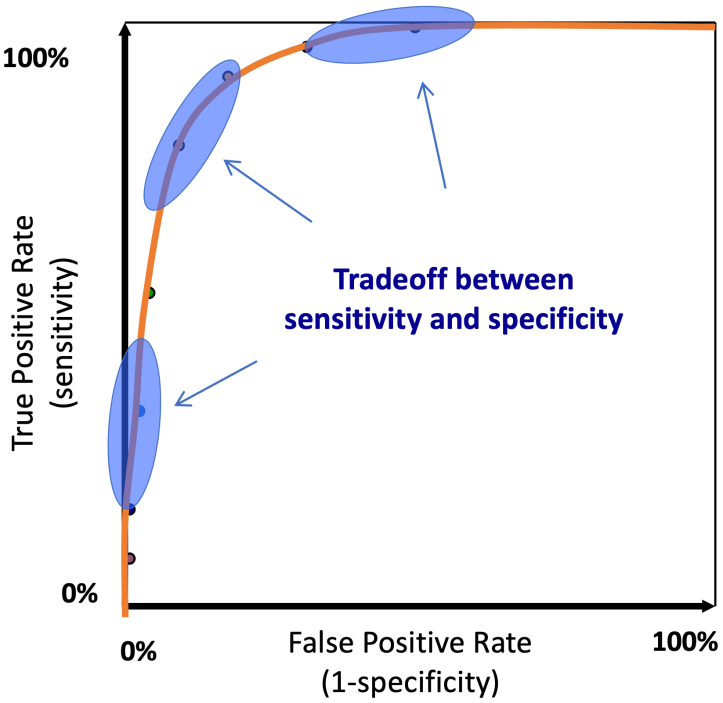
Introduction to ROC curves

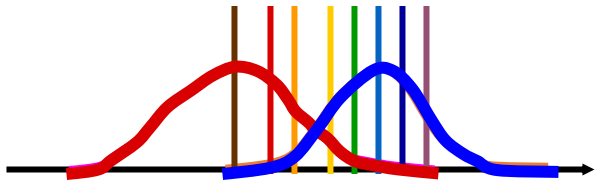
ROC = Receiver Operating Characteristic

- A historic name from radar studies
- Very popular in biomedical applications
  - To assess biomarker performance.
  - To compare different biomarker models

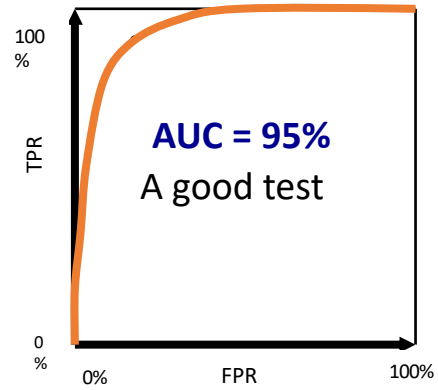
A graphical plot of the true positive rate (TPR) vs. false positive rate (FPR), for a binary classifier (i.e. positive/negative) as its cutoff point is varied

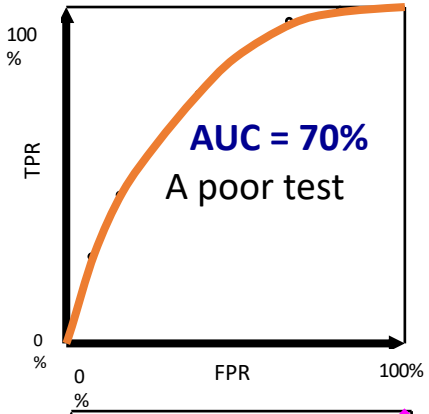
ROC Curve

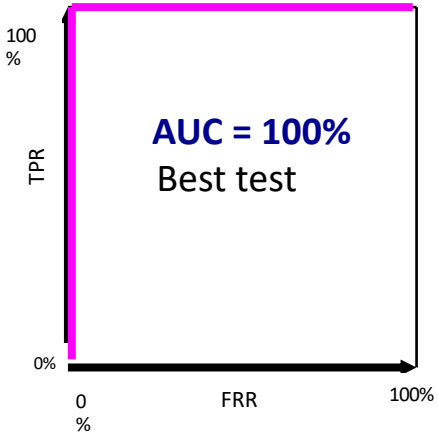

100%
0%
True Positive Rate
(sensitivity)
Tradeoff between
sensitivity and specificity
0%
False Positive Rate
(1-specificity)
100%

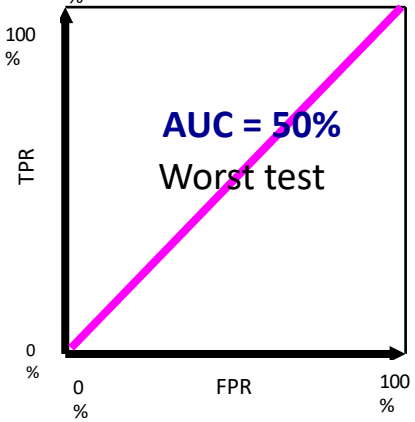


Area under ROC curve (AUC)

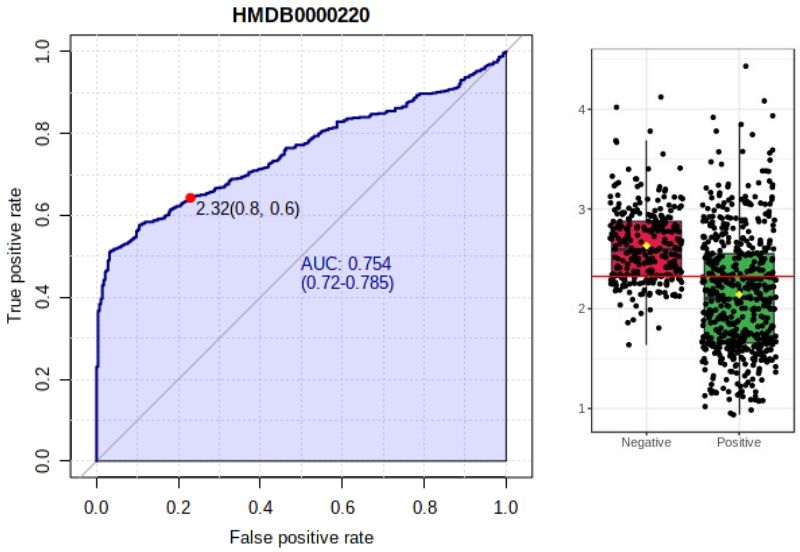




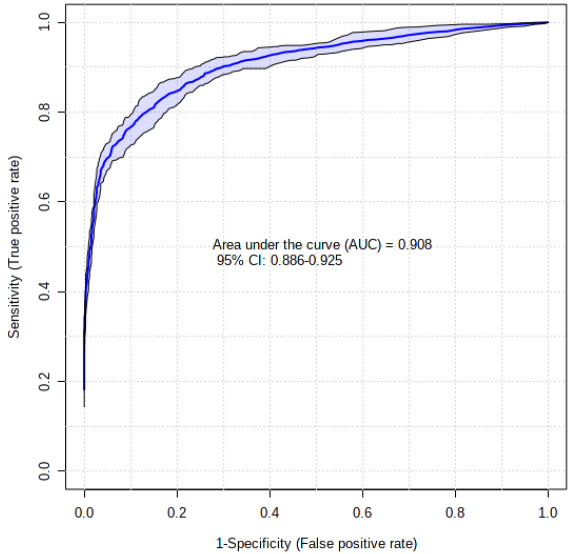




ROC for single and multiple biomarkers



Single biomarker



Multiple features
(Random forests)

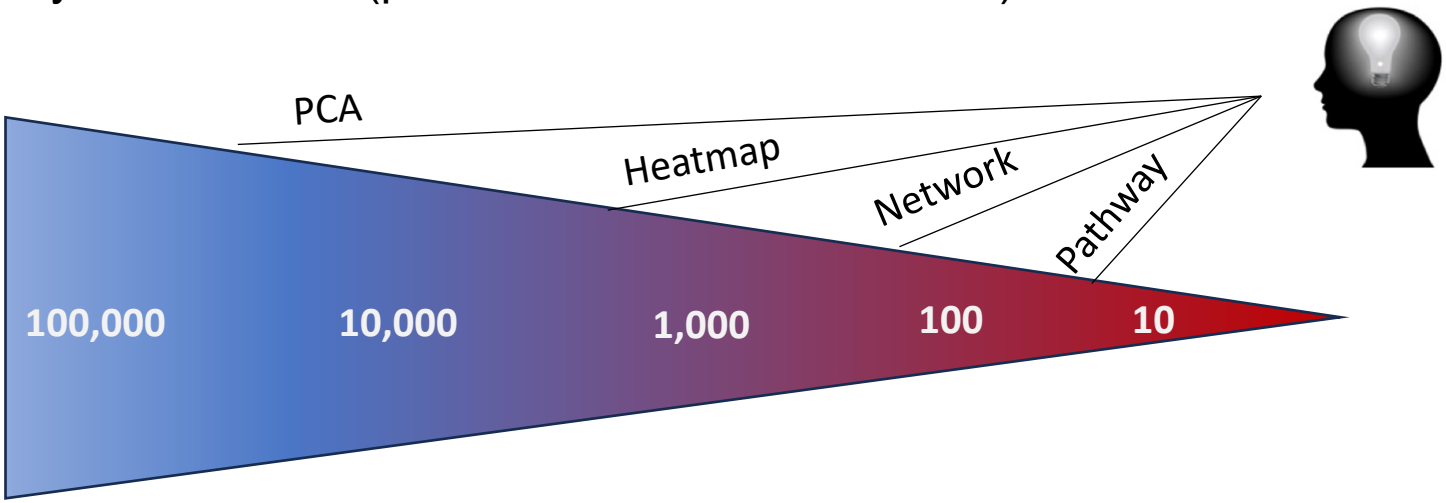
Schedule

Time	Topics
9:00 – 9:20	Introduction & logistics
9:25 – 10:00	Omics data journey
10:05 – 10:50	Pitfalls & performance measures
10:55 – 11:20	Visual analytics on omics data
Comments & Discussions	

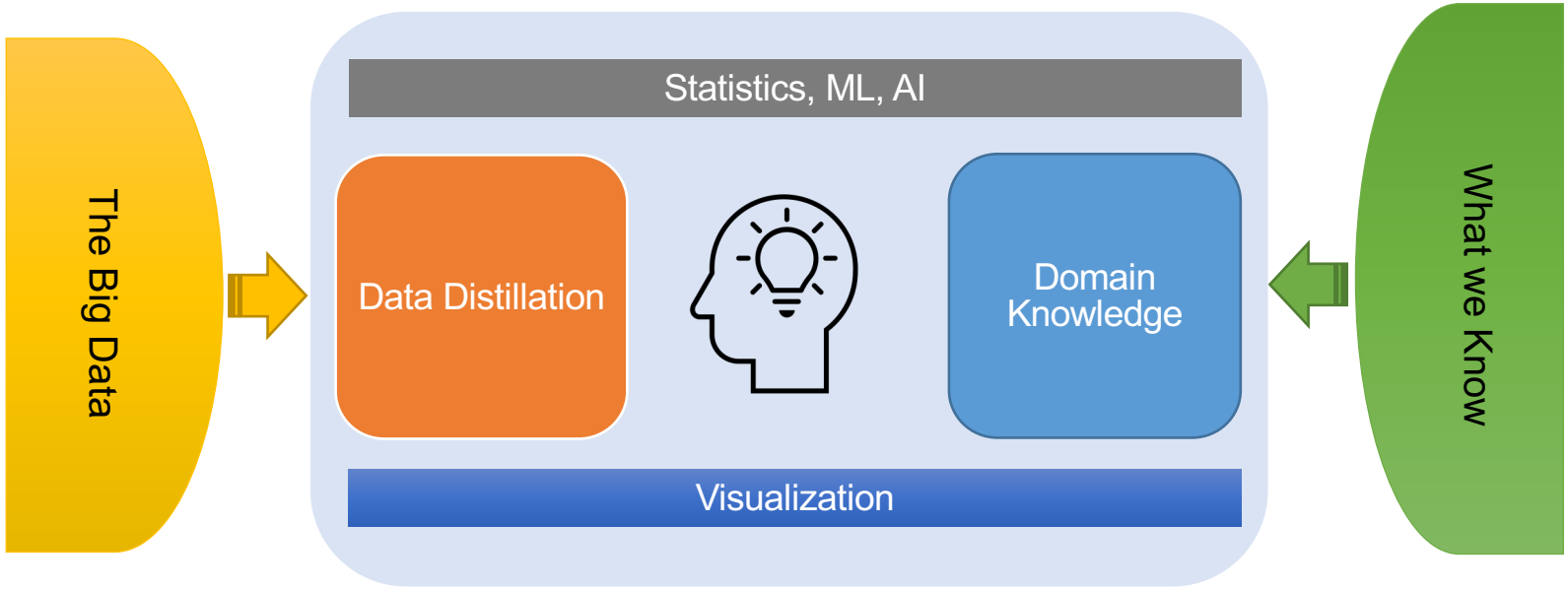
==== 5 minutes break =====

Visualization is critical in omics data analysis

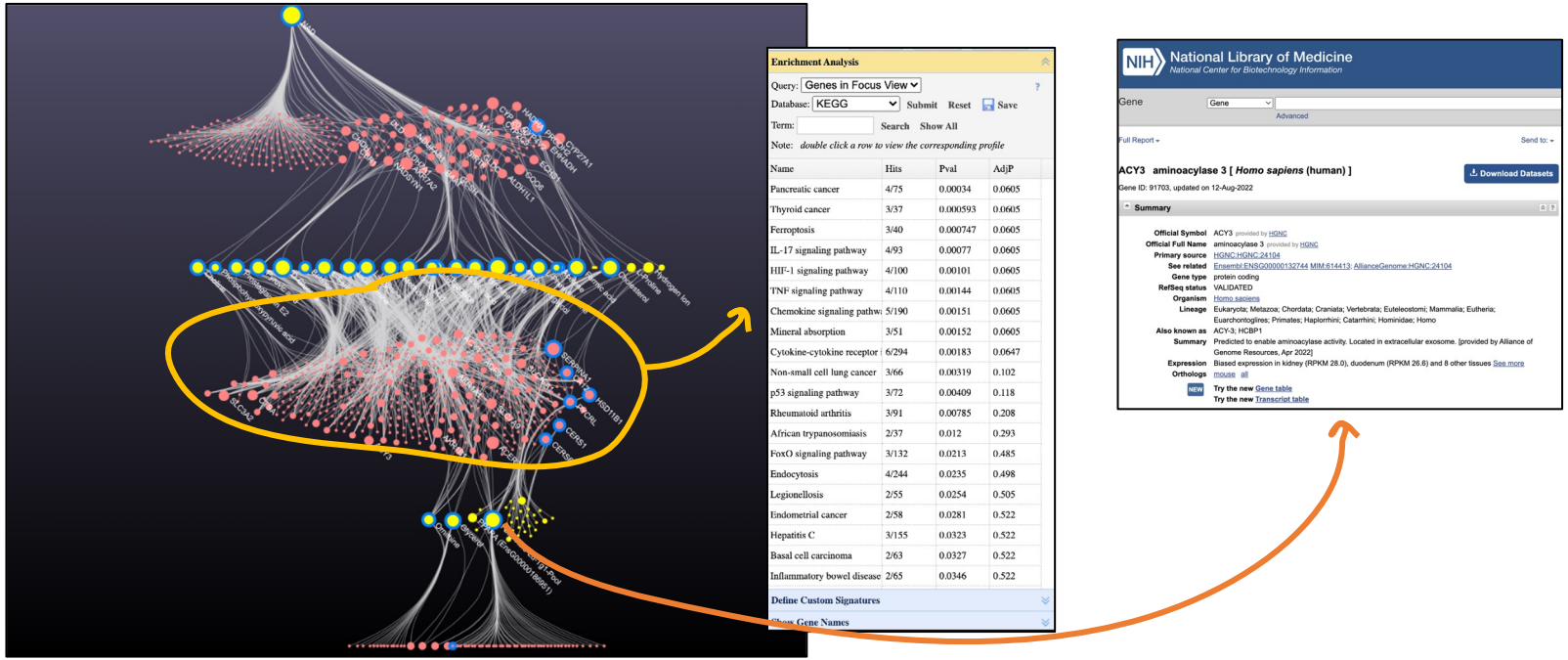
- **"Seeing is believing"**
  - Encourage active learning (user driven)
  - Beyond statistics (p-value alone is not sufficient)



How do we understand data?

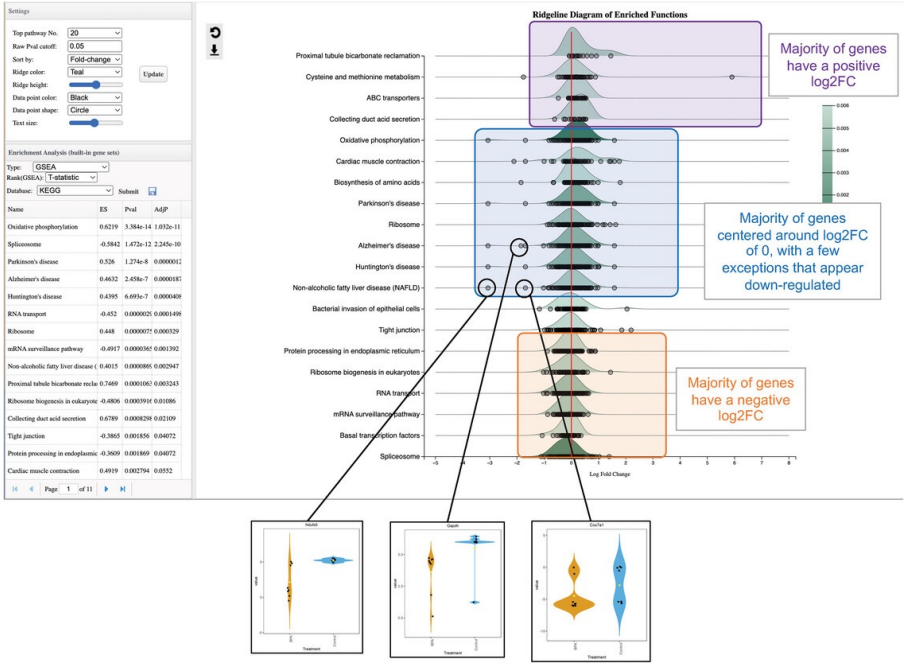


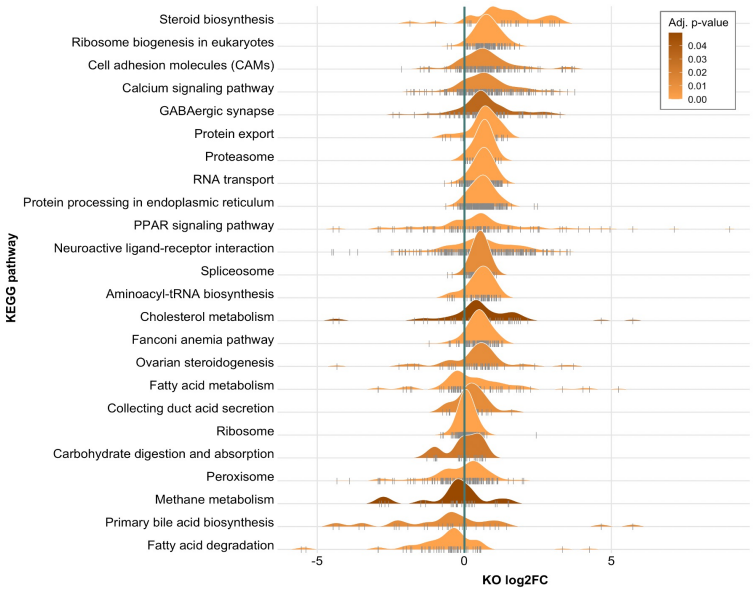
Reduce complexity with visual analytics (NetworkAnalyst)



Visualization + Statistics + Knowledgebase

Revealing functional shift (ExpressAnalyst)

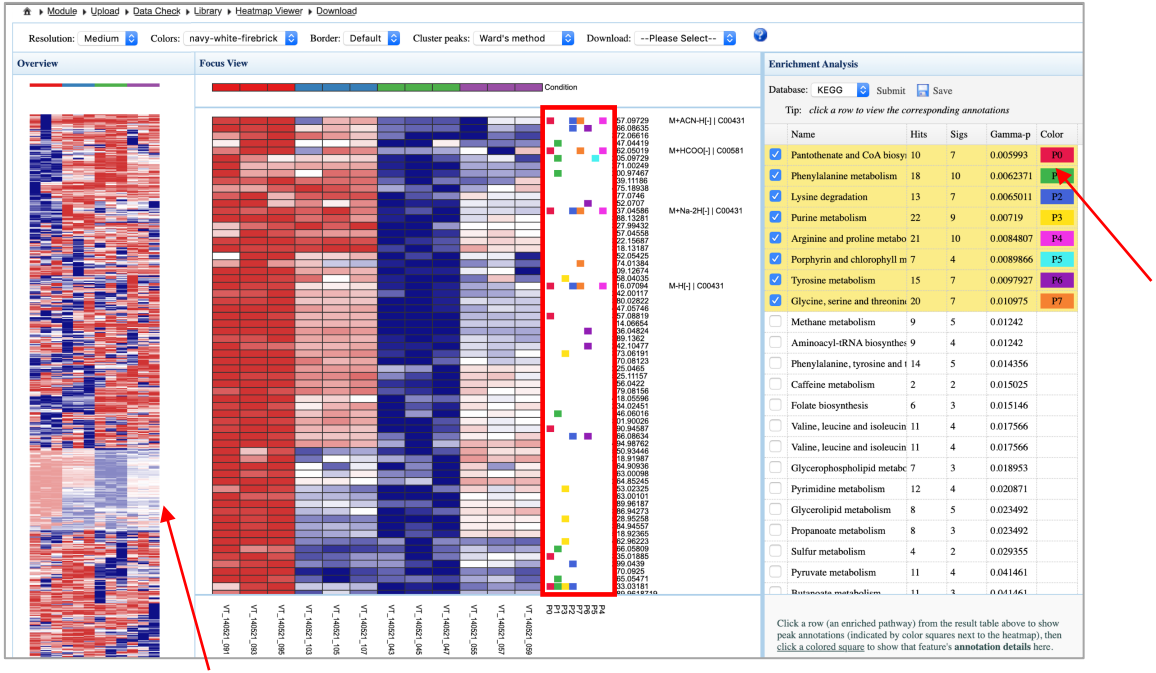




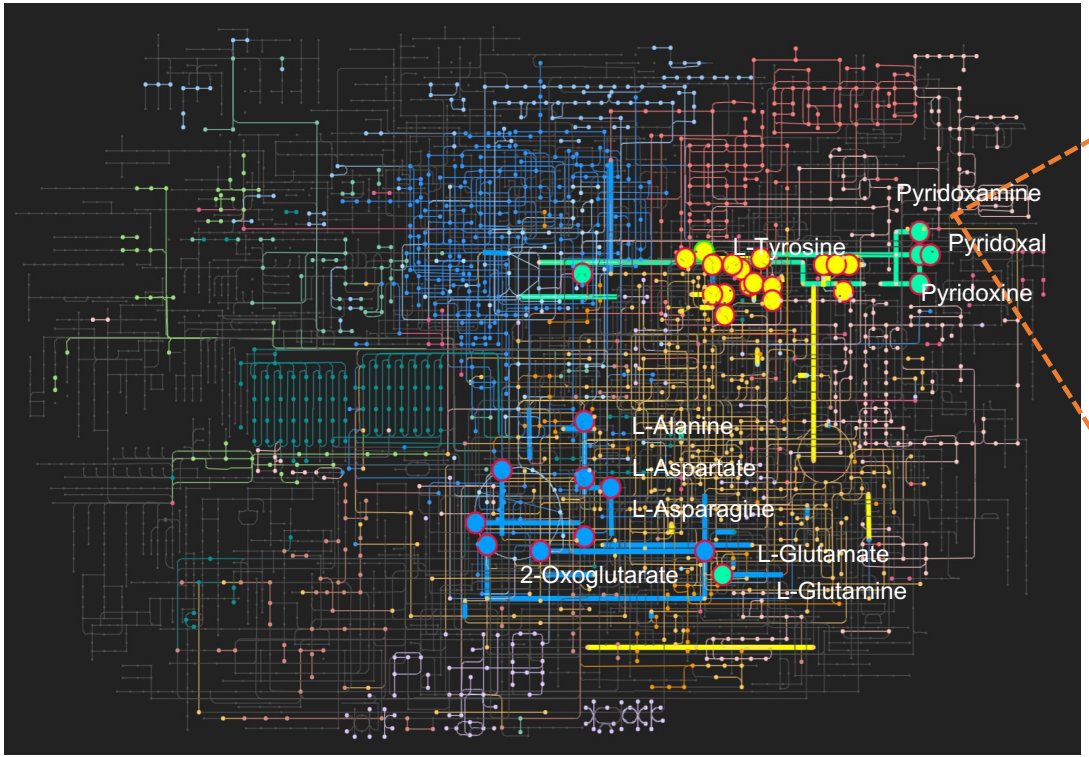
Non-model species

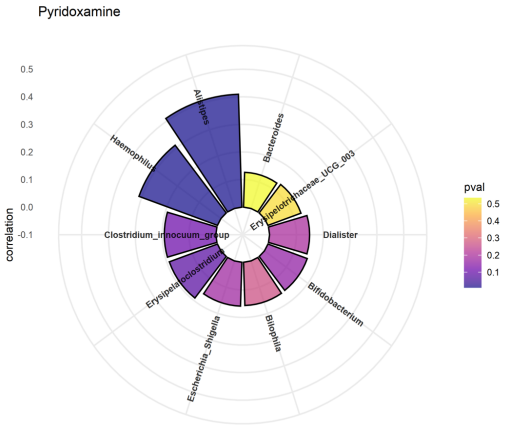
Intuitive & user-driven (MetaboAnalyst)

- Statistics alone is often insufficient to capture the data characteristics of complex omics data
  - Require repeated measures of random, identical exp.
- Disruptive discoveries are often driven by the <u>hunches</u> and <u>leaps of faith</u> of the researcher by interacting with the data
  - Avoid “black-box” method



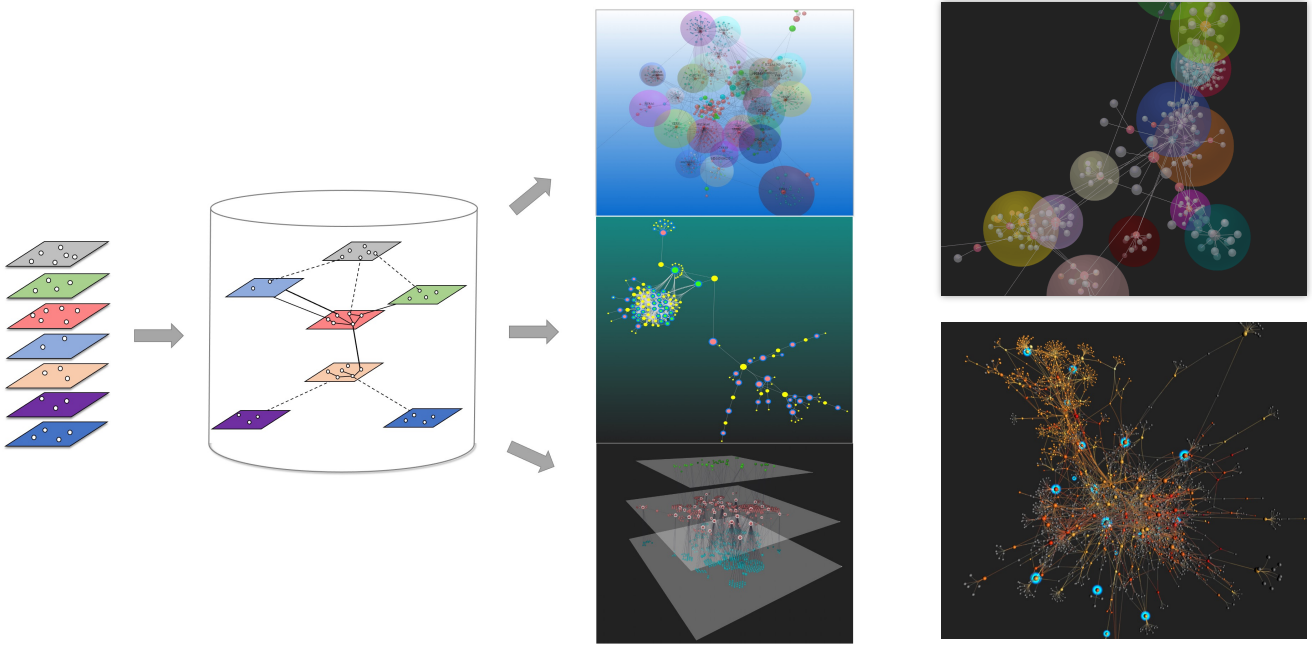
Contextualize data analysis (MicrobiomeAnalyst)





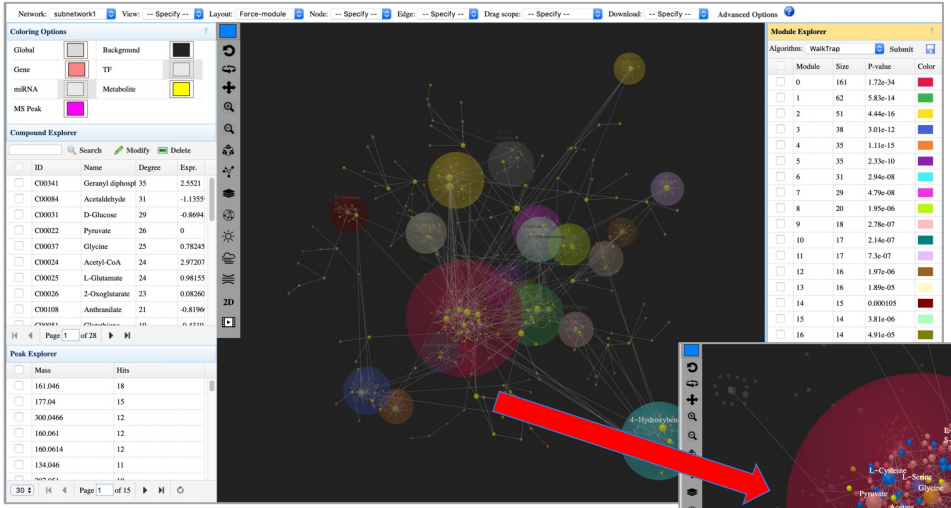
Microbiome Feature	Metabolomics Feature	Correlation	P-value	FDR
Alistipes	C00534	0.4247146	0.00990099	0.04950495
Haemophilus	C00534	0.3124323	0.00990099	0.04950495
Erysipelatoclostridium	C00534	0.1979814	0.04950495	0.1485149
Clostridium_innocuum_group	C00534	0.1926674	0.05940594	0.1485149
Escherichia_Shigella	C00534	0.1593542	0.1188119	0.2376238
Bilophila	C00534	0.1575959	0.1881188	0.3094059
Bifidobacterium	C00534	0.157558	0.2277228	0.3094059
Dialister	C00534	0.1529811	0.2475248	0.3094059
Erysipelotrichaceae_UCG_00:	C00534	0.1315825	0.5544554	0.5544554
Bacteroides	C00534	0.1291311	0.4950495	0.550055

Moving to higher dimensions (OmicsNet)

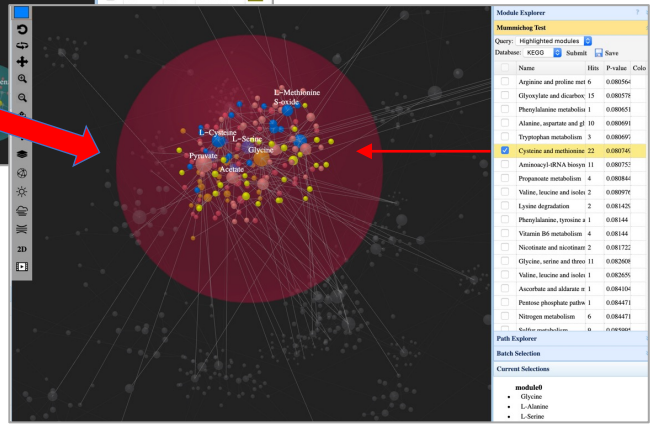


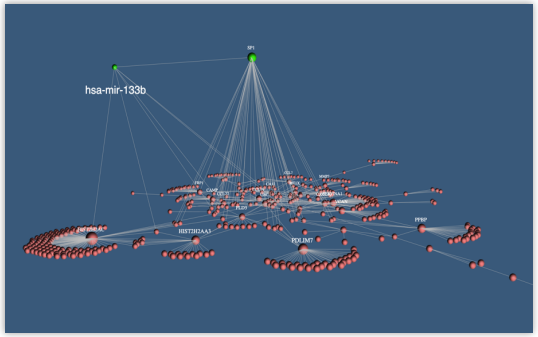
Network ➔ modules ➔ insights

1. Module detection



2. Select a module of interest

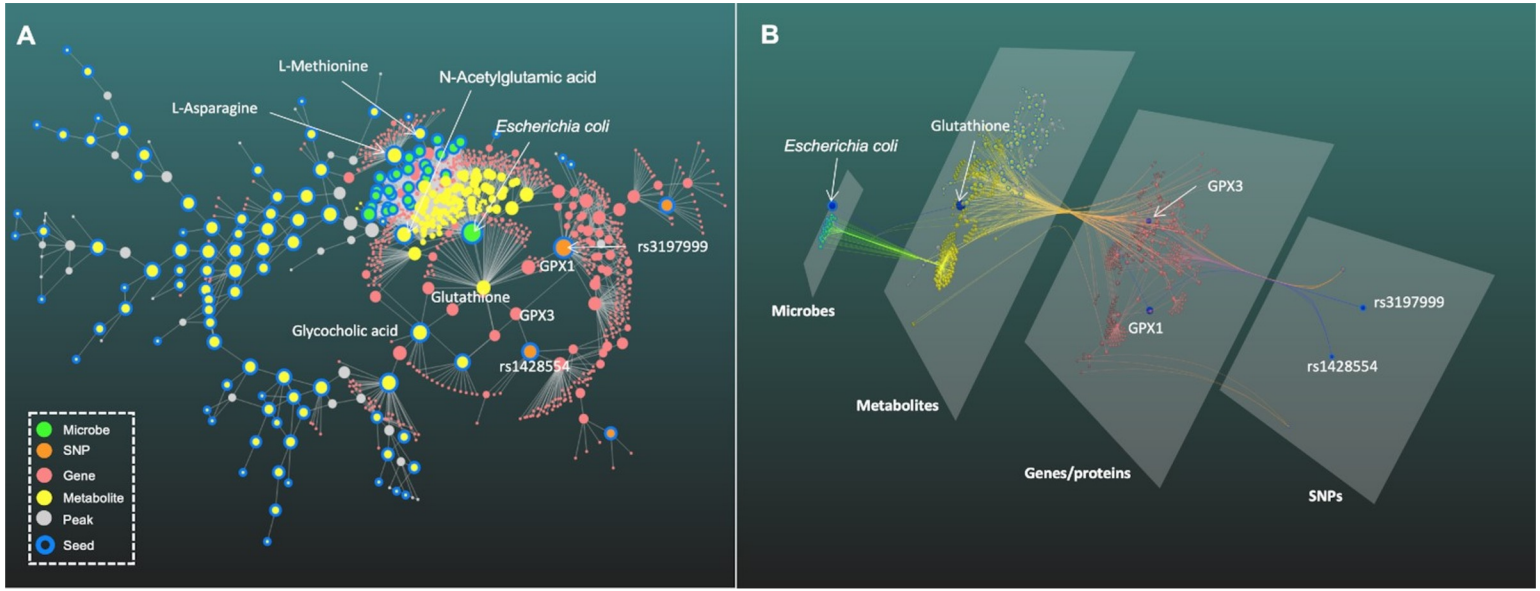




4. Refine layout

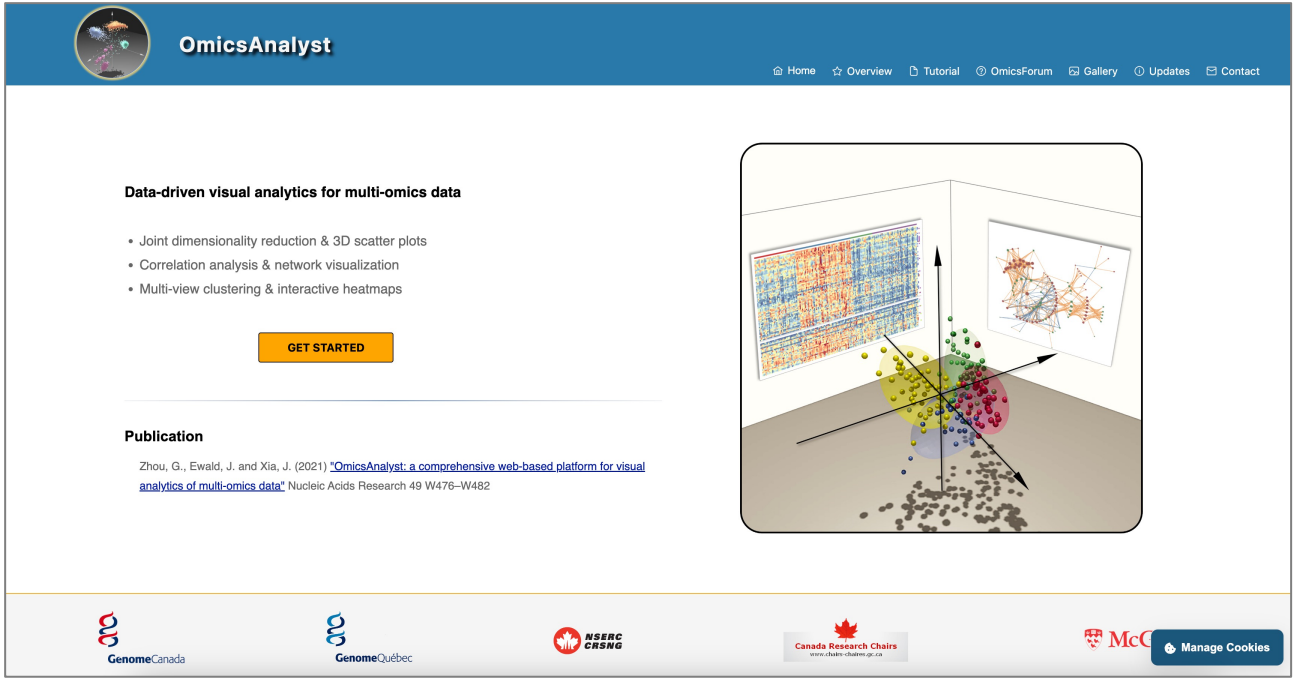
3. Enrichment analysis

Knowledge-driven multi-omics integration

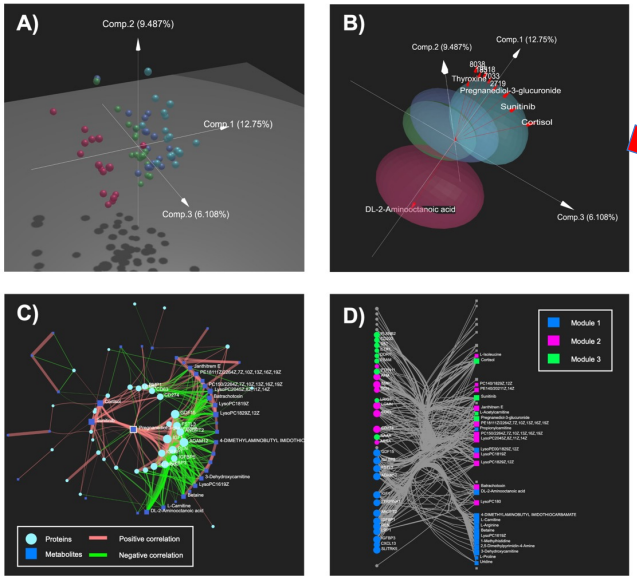


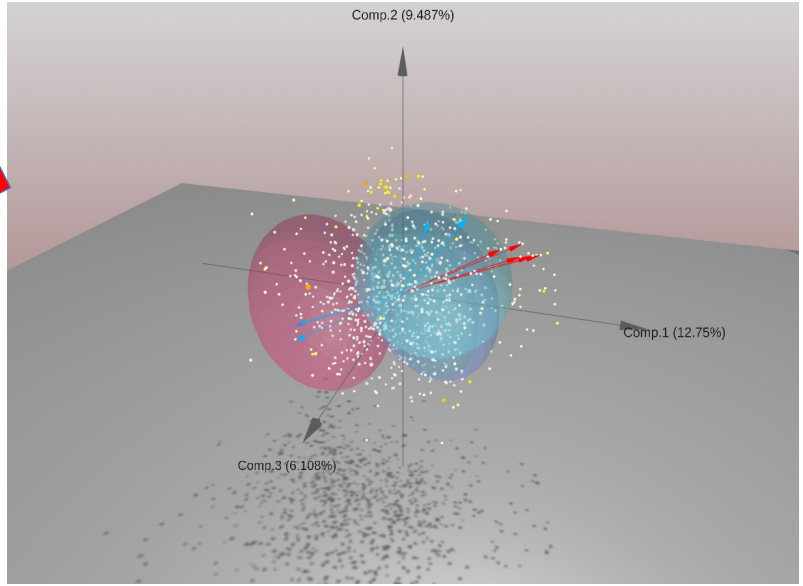
Integrating SNPs, taxa, LC-MS peaks to our knowledge framework

Data-driven multi-omics: OmicsAnalyst



Reduce complexity in multi-omics integration





Towards AI-assisted Self-service Conversational Analytics

Cloud & Edge Computing → Visual Analytics →

← AI Chatbot ← Community Forum & Knowledge base

Next lecture: from raw data to functional insights

1. Raw data processing
   - NGS omics - FASTQ files to count tables
   - MS omics - LC-MS spectra to peak intensity tables
   - Cloud & Docker
2. Functional analysis
   - Over-representation analysis
   - GSEA (transcriptomics)
   - Mummichog (metabolomics)

We would like to hear your comment & feedback

contact@xialab.ca

See you next week!